Learning Progressions in Science

Current Challenges and Future Directions

Alicia C. Alonzo and
Amelia Wenk Gotwals (Eds.)

LeaPS
Learning Progressions in Science

SensePublishers

LEARNING PROGRESSIONS IN SCIENCE

LEARNING PROGRESSIONS IN SCIENCE

Current Challenges and Future Directions

Edited by

Alicia C. Alonzo

Amelia Wenk Gotwals

SENSE PUBLISHERS
ROTTERDAM/BOSTON/TAIPEI

A C.I.P. record for this book is available from the Library of Congress.

ISBN: 978-94-6091-822-3 (paperback)
ISBN: 978-94-6091-823-0 (hardback)
ISBN: 978-94-6091-824-7 (e-book)

Published by: Sense Publishers,
P.O. Box 21858,
3001 AW Rotterdam,
The Netherlands
www.sensepublishers.com

Printed on acid-free paper

CONTENTS

AUTHOR BIOGRAPHIES

Andrés Acher is Assistant Research Professor at the School of Education and Social Policy at Northwestern University. His research examines how to capitalize on ideas students bring to the classroom in order to support them in developing more sophisticated understandings of science knowledge and practices over time. Examples of this research include his participation in (1) MoDeLS, a National Science Foundation (NSF) project that is developing a learning progression for modeling practices and metamodeling knowledge and (2) Modeling Materials at Early School Years, a European Union project that is developing a pathway of articulated ideas to support students in pre-K to fourth grade elementary classrooms to develop more sophisticated understandings of materials while they engage in modeling the properties of and changes in these materials. Previous to his position at Northwestern, he conducted research in four European institutions: King's College London; University of Leeds, UK; University of La Sapienza, Rome; and Autonomic University of Barcelona, where he obtained his Ph.D. in science education.

Alicia Alonzo is an assistant professor of science education in the Department of Teacher Education at Michigan State University. Her research explores knowledge and tools (including learning progressions) that support teachers' formative assessment practices. She has been involved with the development of learning progressions and associated assessment items since 2001, resulting in several co-authored papers. She is a member of the Steering Committee for US Involvement in the Program for International Student Assessment (PISA) and a member of the National Assessment of Educational Progress (NAEP) Science Standing Committee. She served on working groups charged with exploring the inclusion of learning progression items and interactive computer tasks in the 2009 NAEP science assessment. Currently, she serves as an editorial board member for the *Journal of Research in Science Teaching* and *Educational Assessment*. She is a research fellow of the Knowles Science Teaching Foundation and a Lilly Teaching Fellow.

Charles W. (Andy) Anderson is Professor in the Department of Teacher Education, Michigan State University, where he has been since receiving his Ph.D. in science education from The University of Texas at Austin in 1979. Dr. Anderson's current research focuses on the development of learning progressions leading to environmental science literacy for K-12 and college students. He has used conceptual change and sociocultural research on student learning to improve classroom science teaching and science teacher education, science curriculum, and science assessment. Dr. Anderson is past president of the National Association for Research in Science Teaching (NARST). He has been co-editor of the *Journal of Research in Science Teaching* and associate editor of *Cognition and Instruction*. He recently served as a consultant to the National Research Council's (NRC)

Committee on Test Design for K-12 Science Achievement and as a member of the NRC's Committee on Science Learning, K-8. He served as a member of the NAEP Science Framework Planning Committee and the NAEP Science Standing Committee, and he is currently a member of the NRC's Climate Change Education Roundtable.

John Behrens is Director of Networking Academy Learning Systems Development at Cisco and Adjunct Assistant Research Professor in the Department of Psychology at the University of Notre Dame, South Bend, IN. Dr. Behrens is responsible for the design and development of the Networking Academy's global on-line instructional offerings that integrate interactive curricular and assessment technologies with hands-on activities. These curriculum, assessment, and gaming tools serve more than 700,000 students per year in 165 countries across more than 10 languages. Dr. Behrens received his Master's degree in special education and Ph.D. in educational psychology from Arizona State University. He received his B.A. in psychology and philosophy from the University of Notre Dame. Prior to joining Cisco, he was a tenured associate professor of psychology in education at Arizona State University.

Melissa Braaten is an assistant professor in the School of Education's Department of Curriculum and Instruction at the University of Wisconsin-Madison. Her research interests include teacher learning in collaborative inquiry groups, ambitious science instruction that is responsive to students, the construction of explanatory models in school science, and the use of classroom discourse to create generative and equitable learning environments. Earlier in her career she was a middle school and high school science teacher in Tennessee, Texas, and Washington. She received her Ph.D. from the University of Washington.

Derek Briggs is Chair of the Research and Evaluation Methodology Program in the School of Education at the University of Colorado at Boulder, where he also serves as an associate professor of quantitative methods and policy analysis. His research agenda focuses on building sound methodological approaches for the valid measurement and evaluation of growth in student achievement. Examples of his research interests in the area of educational measurement include (1) characterizing the gap between validity theory and practice in the context of high-stakes standardized testing and (2) developing and applying psychometric models to assess learning progressions.

Lea Bullard is a doctoral candidate in science education at the University of Michigan and a research assistant at the University of North Carolina at Wilmington. Her professional work is focused on assessing student learning outcomes in higher education. Her personal research interests center on the nature and evaluation of visitor learning in science museums.

Kristina Chapple, Ph.D., provides research support for the creation and maintenance of assessments throughout the Cisco Networking Academy ecosystem. She works closely with subject matter experts, providing them with the statistical and measurement tools to create tasks, combine tasks on exams, and evaluate the tasks and exams after implementation. She is interested in item response theory, form linking and balancing, and other issues related to the implementation of large-scale assessments.

Younyoung Choi is a Ph.D. candidate in the Department of Measurement, Statistics, and Evaluation at the University of Maryland at College Park. She received her Master's degree in the Department of Measurement, Statistics, and Evaluation at the same university and a Bachelor of Arts from the Department of Psychology at Sungkyunkwan University in South Korea. She is interested in developing simulation/game-based assessments using the Evidence Centered Design framework and investigating statistical methodologies to assess learners' progressions using Bayesian Networks.

Beth A. Covitt is a research assistant professor in the Environmental Studies Program at the University of Montana. She conducts learning progression research addressing students' understanding of water in environmental systems and students' use of science in socio-ecological decision-making. Dr. Covitt also conducts science teacher professional development with elementary and middle school teachers in Montana.

Aaron Crawford is a graduate student in the Division of Psychology in Education at Arizona State University. His current research areas include methodological studies, item response theory, Bayesian networks, and structural equation modeling. He received M.A. and B.S. degrees in psychology from the University of North Florida.

Kristen DiCerbo is an evaluation specialist at Cisco Learning Institute. Her research interests include understanding how students learn technical skills, researching instructional practice around the world, and using new and different techniques to explore teaching and learning. This research occurs primarily in the context of the Cisco Networking Academy Program. Dr. DiCerbo received her M.Ed. and Ph.D. in educational psychology from the School Psychology Training Program at Arizona State University. She continues to teach a course on individual intellectual assessment at Arizona State. She received her B.A. in psychology and sociology from Hamilton College. Prior to joining Cisco Learning Institute, she worked as a school psychologist.

Susan J. Doubler co-directs the Center for Science Teaching and Learning at TERC and is Associate Professor of Education at Lesley University. Her work focuses on the interface between science education and the use of technology to further inquiry-based learning. She is currently leading two projects funded by the

NSF: *The Inquiry Project*, an effort to develop and study a learning progression about matter for grades 3–5, and *Talk Science*, an effort to develop web-based professional development resources aligned with the Inquiry curriculum and focused on productive classroom talk. She recently co-led the development of a web-based science leadership program for K-8 teachers and the development of a fully online master's program in science education. Before coming to TERC and Lesley University, she was an instructional specialist and teacher in the Winchester (MA) Public Schools.

Jacob Foster is Director of Science and Technology/Engineering at the Massachusetts Department of Elementary and Secondary Education. In this role he oversees the state's science and technology/engineering standards and curriculum framework, state-funded professional development opportunities, and support for districts. Dr. Foster has been a member of the writing team for the National Science Teacher Association's *Anchors* project, has served on a design team for the NRC's *Conceptual Framework for New Science Standards*, and is on the writing team for the *Next Generation Science Standards*. Previously Jacob worked with the Coalition of Essential Schools on school reform and conducted school reviews as part of Massachusett's accountability system. He has taught high school physical and earth sciences. In addition, he has served as a middle school science coach and science teacher educator. Dr. Foster earned a B.A. in earth science from Hampshire College and an M.A. and Ph.D. in science education from the University of Michigan.

David Fortus is a senior scientist at the Weizmann Institute of Science in Israel. He is interested in ways to improve middle school science education and to motivate middle-schoolers to continue studying science in high school. At present he is involved in the publication of a coordinated and comprehensive middle school science curriculum and a study of the reasons why students' motivation to learn science declines from the end of elementary school and throughout middle school.

Erin Marie Furtak is Assistant Professor of Education specializing in science education at the University of Colorado at Boulder. As a former public school teacher, Dr. Furtak's research focuses on the development of teachers' knowledge and practices to support reform-oriented science teaching. Currently, she is (1) exploring how learning progressions about natural selection can support teacher learning communities as they develop, enact, and revise common formative assessments and (2) relating teacher enactment of formative assessment to student learning. Her projects have been supported by the Alexander von Humboldt Foundation, the Knowles Science Teaching Foundation, the Spencer Foundation, and a CAREER grant from the NSF.

Amelia Wenk Gotwals is an assistant professor of science education at Michigan State University. Her research interests include examining the ways that students and teachers develop more sophisticated understandings and abilities. Specifically, she is interested in the development, evaluation, and assessment of learning progressions in science and ways to assess complex reasoning in authentic scientific situations. In addition, she is interested in investigating how teachers develop assessment literacy, specifically, how teacher candidates can learn to gather evidence of student understanding and use this evidence to modify their teaching practices. Her work has been funded by the NSF and the Spencer Foundation, through a Spencer Dissertation Research Fellowship.

Kristin L. Gunckel is an assistant professor of science education at the University of Arizona. She received her Ph.D. from Michigan State University. She is co-leader of the water cycle learning progression strand in the Environmental Literacy Project and the Reasoning Tools for Understanding Water Systems Project. In addition to learning progressions, she is also interested in elementary science teacher education.

Hui Jin is an assistant professor at The Ohio State University. Her research interests include learning progressions, students' informal explanations and causal reasoning, conceptual change theories, and qualitative research methodology. She received her Ph.D. from Michigan State University.

Lisa Kenyon is an associate professor in the Department of Biological Sciences in the College of Science and Mathematics and the Department of Teacher Education in the College of Education and Human Services at Wright State University. Her research focuses on engaging students and teachers in scientific practices such as explanation, argumentation, and scientific modeling—specifically, examining how students use their epistemologies of science to support these practices. Her current focus on scientific modeling includes two main areas: (1) developing a learning progression for upper elementary and middle school students and (2) supporting preservice teachers' pedagogical content knowledge for scientific modeling. Other research interests include curriculum design, project-based inquiry, and teacher professional development.

Joseph Krajcik, Professor in the College of Education at Michigan State University (MSU) and Director of the CREATE for STEM Institute, focuses his research on exploring the impact of designing innovative classroom environments in which students find solutions to important intellectual questions that subsume essential learning goals. He has authored and co-authored over 100 manuscripts and makes frequent presentations focused on his research, as well as presentations that translate research findings into classroom practice, at international, national and regional conferences. He is a fellow of the American Association for the Advancement of Science (AAAS) and the American Educational Research Association (AERA), served as president of NARST, and received guest professorships from Beijing Normal University and the Weizmann Institute of Science. In 2009, Ewha University in Korea named him a distinguished fellow. Prior to coming to MSU, Joe spent 21 years at the University of Michigan.

Amy Kurpius, M.S., has been working in the field of higher education for more than 10 years. Her work experience includes project management, research evaluation and analyses, grants management, and financial administration. Her education work has focused on professional and executive education, learning progressions and formative assessments in K-8 science and mathematics education, performance assessment in pre-college and higher education, and admissions assessments of achievement and ability. She has consulted with organizations such as the Council for Aid to Education, the College Board, Indiana University, Organization for Economic Co-Development, Stanford University, Teachers College, and the University of Pennsylvania.

Roy Levy is an assistant professor in the School of Social and Family Dynamics at Arizona State University. His work focuses primarily on the use of statistical and psychometric models in assessment, education, and the social sciences more broadly. His interests include methodological investigations and applications in psychometrics and statistical modeling, with an emphasis on models and procedures traditionally associated with item response theory, structural equation modeling, and Bayesian networks.

Junhui Liu is a doctoral student in the Department of Measurement, Statistics and Evaluation at the University of Maryland, College Park. She has a M.Ed. in Curriculum and Instruction from Texas A&M University. Her research interests include growth modeling, multidimensional item response theory, and Bayesian estimation.

Robert Mislevy is the Frederic M. Lord Chair in Measurement and Statistics at Educational Testing Service (ETS). He was previously Professor of Measurement and Statistics at the University of Maryland at College Park, and before that, Distinguished Research Scientist at ETS. His research applies developments in statistics, technology, and psychology to practical problems in assessment. His work includes the multiple-imputation analysis for the NAEP, the BILOG computer program for estimating item parameters, an evidence-centered assessment design framework, and simulation-based assessment of computer networking proficiency with Cisco Systems. He has served as president of the Psychometric Society, has received awards for career contributions to educational measurement from the National Council on Measurement in Education and Division D of AERA, and is a member of the National Academy of Education.

Lindsey Mohan is an adjunct research scientist at Michigan State University. She received her Ph.D. in educational psychology from Michigan State University. Through her work with National Geographic Society, Lindsey developed video-based professional development on the Environmental Literacy Teacher Guides series and is currently working on carbon-cycling learning progression research with the Environmental Literacy Project. She has also conducted research documenting exemplary teaching and exceptional classroom dialogue in science.

Teresa Neidorf is a principal research scientist at the NAEP Education Statistics Services Institute (NAEP-ESSI) of the American Institutes for Research. Since 2003, she has worked on projects related to cognitive item development and scoring for the NAEP, with a recent focus on the NAEP 2009 Science Assessment. This work includes participation in an expert panel convened by the National Center for Education Statistics (NCES) to explore the potential of learning progressions in the NAEP 2009 Science Framework. Prior to beginning her work on NAEP, Dr. Neidorf was a senior research associate at the International Study Center at Boston College, where she served as the science coordinator for the Trends in International Mathematics and Science Study (TIMSS) 1999 and 2003 assessments. Since 2003, she has conducted studies for NCES comparing NAEP with international assessments. Dr. Neidorf's professional training was originally in the physical sciences. She has a Ph.D. in chemistry from Stanford University and worked for several years as a research scientist at Eastman Kodak Company before obtaining a M.Ed. in Educational Research, Measurement, and Evaluation and changing the focus of her career to science assessment.

Julia Plummer is an assistant professor at Arcadia University where she coordinates the science education program. Prior to this position, Dr. Plummer spent more than a decade teaching children and adults in planetariums. Her research interests include investigating the design of learning environments for developing students' understanding of astronomy and the practices of science as applied to astronomy, in both classroom and informal environments. She is also interested in the role of curriculum and professional development in educators' beliefs about astronomy education. Dr. Plummer has conducted a series of studies that have led to the development of an initial framework for a K-8 observational astronomy learning progression. She has co-authored a middle school astronomy curriculum published through *It's About Time.*

Brian J. Reiser is Professor of Learning Sciences in the School of Education and Social Policy at Northwestern University. Dr. Reiser's research examines how to make scientific practices such as argumentation, explanation, and modeling meaningful and effective for classroom teachers and students. This design research investigates (1) the cognitive and social interaction elements of learning environments supporting scientific practices and (2) design principles for technology-infused curricula that embed science learning in investigations of contextualized data-rich problems. Dr. Reiser leads the MoDeLS (Modeling Designs for Learning Science) project, developing an empirically-based learning progression for the practice of scientific modeling, and BGuILE (Biology Guided Inquiry Learning Environments), developing software tools for supporting students in analyzing biological data and constructing explanations. Dr. Reiser is also on the leadership team for IQWST (Investigating and Questioning our World through Science and Technology), a collaboration with the University of Michigan developing a middle school project-based science curriculum. Dr. Reiser was a founding member of the first graduate program in Learning Sciences, created at

Northwestern, and chaired the program from 1993, shortly after its inception, until 2001. He was co-principal investigator in the NSF Center for Curriculum Materials in Science, exploring the design and enactment of science curriculum materials, and served on the NRC panels authoring the reports *Taking Science to School* (2007) and *Conceptual Framework for New Science Education Standards* (2011).

Daisy Wise Rutstein is a graduate student in the Department of Measurement, Statistics and Evaluation at the University of Maryland, College Park. Her interests include Bayesian networks and cognitive diagnosis modeling.

Christina Schwarz is an associate professor in the Department of Teacher Education at Michigan State University. She was Co-PI of the MoDeLS project, developing a learning progression for scientific modeling and is currently Co-PI of the Scientific Practices project, integrating scientific modeling, explanation and argumentation practices. Her work focuses on determining how to enable students and teachers to understand and engage in the practices of science—particularly model-based scientific inquiry. Her research interests also include student and teacher learning progressions, educational technology in science teaching and learning, and the role of curriculum materials in science teaching and learning.

Richard J. Shavelson is the Emeritus Margaret Jacks Professor of Education and Professor of Psychology (by courtesy) at Stanford University, and former I. James Quillen Dean of the School of Education at Stanford and Senior Fellow in Stanford's Woods Institute for the Environment. He is currently director of research and development at SK Partners, LLC. He served as president of AERA; is a fellow of AAAS, AERA, the American Psychological Association, and the American Psychological Society; and is a Humboldt Fellow (Germany). His current work includes assessment of undergraduates' learning using the Collegiate Learning Assessment, accountability in higher education, assessment of science achievement, the validity of learning progressions, the enhancement of women's and minorities' performance in organic chemistry, and the role of mental models of climate change on sustainability decisions and behavior. Other work includes studies of computer cognitive training on working memory, fluid intelligence and science achievement; the scientific basis of education research; and new standards for measuring students' science achievement in the NAEP (the nation's "report card"). His publications include *Statistical Reasoning for the Behavioral Sciences,Generalizability Theory: A Primer* (with Noreen Webb); *Scientific Research in Education* (edited with Lisa Towne); and *Assessing College Learning Responsibly: Accountability in a New Era* (2010).

Carol L. Smith is Professor in the Department of Psychology, University of Massachusetts at Boston, where she has been since receiving her Ph.D. in developmental studies from Harvard University in 1976 and completing postdoctoral research at the Massachusetts Institute of Technology (MIT) in 1978. She is a cognitive developmental psychologist whose work on conceptual change in science education exemplifies part of the research basis for learning progressions. Over the past 30 years she has studied the conceptual changes that occur as children develop their ideas about matter as well as their ideas about scientific models and knowledge construction in science. She has also collaborated with teachers, scientists, and science educators to create innovative teaching units for elementary and middle school students and to study their effectiveness in facilitating conceptual restructuring compared with more traditional teaching approaches. Most recently, she worked on a team that synthesized current research in order to propose a long-term learning progression for matter and served on the NRC's Committee on Science Learning, K-8, which authored *Taking Science to School*. She is currently collaborating on two longitudinal studies—one with elementary school students (on matter) and the other with college students (concerning their conceptions of science).

Nancy Butler Songer is Professor of Science Education and Learning Technologies in the School of Education at the University of Michigan and the Director of the Center for Essential Science (www.essentialscience.umich.edu). Her research is focused in two areas: (1) addressing 4–10th grade urban students' underperformance in science through the design and evaluation of curricular units and emerging technologies focused on complex reasoning in science and (2) designing and evaluating assessments to measure complex learning in science. She has a Ph.D. in science education from the University of California, Berkeley, a M.S. in Molecular/ Developmental Biology from Tufts University, and a B.S. in Biological Sciences from the University of California, Davis. She was awarded a Presidential Faculty Fellowship from President William J. Clinton in 1996 and is a Fellow of the AAAS.

Jessica Thompson is a research assistant professor in the College of Education at the University of Washington. Her research focuses on engaging underserved students in science and on supporting science teachers in working toward ambitious and equitable pedagogy. Dr. Thompson was awarded an American Association of University Women Dissertation Fellowship, the 2007 Selma Greenberg Dissertation Award, and a 2010 Knowles Science Teaching Foundation (KSTF) Fellowship. The KSTF is supporting a project that examines co-learning of high-leverage practices between teacher candidates and cooperating teachers. With funding from the Carnegie Foundation and the NSF, she has spent the last five years working on the development of tool systems designed to advance early career science teachers' pedagogical reasoning and practice (see http://tools4teachingscience.org/). For the last three years she has also run a video club, in partnership with local school districts, that supports in-service science teachers in examining links between practice and student learning.

Patti West is an employee of Cisco Systems. She is an educational specialist on the Cisco Networking Academy Learning Systems Development team. Her interests include the application of latent variable psychometric and Bayesian inference models to inform the design and development of curriculum, assessment and learning games.

Mark Wilson is Professor in the Graduate School of Education at the University of California, Berkeley. He is a Fellow of the American Psychological Association and Founding Editor of the journal *Measurement: Interdisciplinary Research and Perspectives.* He is currently the President of the Psychometric Society. His interests in measurement range from reforming the approach to measurement in education and, more broadly, across the social sciences, to innovations in mathematical and statistical modeling for measurement to the policy and practical issues involved in educational and psychological assessment.

Mark Windschitl is a professor of science teaching and learning at the University of Washington. His research interests deal with the early career development of science teachers—in particular their trajectories toward ambitious and equitable pedagogy. His research group has developed a set of high-leverage practices for K-12 science instruction and a set of discourse tools that allow beginners an entry point into expert-like dialogic interactions with young learners. This work is supported by a 5-year grant from the NSF. Work from this and related projects has appeared in the *American Educational Research Journal*, *Teachers College Record*, *Cognition and Instruction*, *Phi Delta Kappan*, *Science Education*, and in white papers commissioned by the NRC and the National Academy of Science. Dr. Windschitl is the PI on a Noyce Teaching Scholars grant and has supported approximately 30 teachers in that program in their transitions to urban schools. He also co-administers the Annenberg Fellowship program—also known as the Rhodes Scholarships of Teaching—for teachers at the UW. He is the recipient of the 2002 AERA Presidential Award for Best Review of Research, the co-author of the chapter on science teaching in the new AERA *Handbook of Research on Teaching*, and the 2011 recipient of the Iowa State University Outstanding Alumni Achievement Award.

Marianne Wiser is Associate Professor and Chair of the Psychology Department at Clark University. She has degrees in physics and engineering and in physical oceanology from the University of Liege, Belgium, and a Ph.D. in cognitive and brain science from MIT. Her areas of research are science learning and teaching in elementary and middle school; the development of physical concepts, numeracy, and notations in young children; and the role of language in science learning. Her work in science education focuses on learning progressions of matter and energy for the elementary grades. She is working with a team of researchers, curriculum developers, and teachers from Tufts University, University of Massachusetts Boston, and TERC. She is also collaborating with colleagues at Clark University on an oral language curriculum for early elementary school based on science learning progressions and language development research. She has been a consultant for the College Board as part of a team that redesigned the AP Physics assessment using an evidence-centered model; for NRC on how to organize science standards around core ideas; and for the Massachusetts Department of Education on how to revise science standards around learning progressions.

FRAMING SECTION

AMELIA WENK GOTWALS & ALICIA C. ALONZO

INTRODUCTION

Leaping Into Learning Progressions in Science

Learning progressions—descriptions of increasingly sophisticated ways of thinking about or understanding a topic (National Research Council [NRC], 2007)—offer a promising framework for bringing coherence to multiple facets of the educational system. Learning progressions, which articulate cognitive models of the development of student understanding, have the potential to inform the design of standards, large-scale and classroom assessments, curricula, and teacher professional development. As such, the science education community has taken considerable interest in learning progressions. However, as with any new research agenda, there are challenges. If these challenges are not addressed, they may thwart the promise that learning progressions hold.

Many journal articles and conference presentations gloss over the challenges authors and presenters have encountered in their work. In contrast, this book emphasizes that such challenges in learning progression work need to be a central part of the conversation. The chapters in this book recognize that, in order for learning progressions to fulfill their promise, the community must undertake a critical examination of the challenges in defining, assessing, modeling, and using learning progressions. Therefore this book explores learning progression work through an examination of some of its most important challenges.

This chapter introduces the book. First, we discuss why learning progressions have generated so much interest in the science education community. Next, we describe the current challenges in learning progression work and our impressions of the field. We conclude by outlining the book's structure and discussing our expectations of how the book may advance learning progression work.

THE PROMISE OF LEARNING PROGRESSIONS

Students in the United States (US) consistently perform worse on international standardized tests of science achievement than their peers in other countries. US students perform relatively well in fourth grade. However, by the eighth grade their level of performance drops considerably and remains low throughout compulsory schooling (Gonzales et al., 2004; Schmidt, McKnight, & Raizen, 1997). In addition, there is increasing concern that schools in the US do not adequately teach students the scientific knowledge and skills needed for success in the workforce (National Academy of Sciences, National Academy of Engineering & Institute of Medicine, 2007).

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 3–12.

One possible explanation of these disconcerting findings is that K-12 science education in the US is unfocused, with many topics presented in a disconnected and shallow fashion (Valverde & Schmidt, 1997). US science textbooks address many more topics than textbooks in other countries (Linn, Tsuchida, Lewis, & Songer, 2000) and do so in a superficial and unsystematic manner, such that content is not presented in a logical order and does not build from year to year (Roseman, Kesidou, Stern, & Caldwell, 1999). Thus US students lack the opportunity to develop the coherent understandings and skills needed to develop scientific literacy.

The idea of preparing students to be scientifically literate is not new. In its 1990 report, *Science for All American,* the American Association for the Advancement of Science [AAAS] argued that students should acquire deep understandings of big ideas in science in order to understand socio-scientific issues such as global warming, population growth, nuclear energy, and public health. Although this argument was widely accepted by the science education community, two decades later we still do not have a clear understanding of how to achieve science literacy through science education.

Learning progressions have the potential to organize standards, assessments, and instruction in a way that promotes scientific literacy. Current standards and curricula prioritize the structure of the scientific disciplines, using a top-down approach that creates logical (from scientists' perspective) sequences of ideas. Learning progressions, which use both top-down and bottom-up design approaches, can combine ideas about scientific domains with understandings of how students learn. Thus learning progressions provide a significantly different perspective from that of other currently available frameworks for organizing standards, assessments, and instruction.

Learning progressions prioritize big ideas that are generative and merit extended periods of study. As part of the top-down design approach to learning progressions, scientists and science educators select these big ideas from the core knowledge needed for understanding socio-scientific issues and achieving scientific literacy. However, this logical decomposition of big ideas may not necessarily reveal the paths students take as they learn scientific content. Therefore, the bottom-up design approach to learning progressions promotes the organization of content based on students' thinking as they develop more sophisticated understandings. Students' progression from naïve to more sophisticated understandings may not be linear or easily described. An investigation of the "messy middle" (Gotwals & Songer, 2010, p. 277) of students' learning may thus provide powerful information for formative assessments (e.g., Alonzo, 2011), curriculum development (e.g., Wiser, Smith, & Doubler, this volume), and standards (e.g., Foster & Wiser, this volume).

The top-down and bottom-up processes of developing learning progressions require varied expertise. Learning progressions draw on existing work that has not before been brought together in a coherent and systematic manner. In addition, learning progressions require collaborations to generate new knowledge needed to advance the field even further. In the past, scientists and science educators have articulated core ideas in science that are generative and allow students to integrate knowledge that produces powerful explanations of socio-scientific phenomena (e.g., AAAS, 1990; NRC, 1996, 2007). However, while they have identified goals for scientifically literate citizens, they have not taken the bottom-up design approach described above; thus they have failed to identify and/or prioritize the ways students achieve these goals.

Cognitive and learning scientists have conducted research on how children learn in specific domains and have studied the ideas students bring to school. However, much of this research has been conducted outside the classroom, with limited success in transferring the knowledge acquired to learning environments. In addition, assessment experts have researched ways of ascertaining what students know, and psychometricians have developed sophisticated models of students' responses to assessments. Yet, since there has been little communication between science educators and these measurement experts, new techniques have not been systematically applied to science education (NRC, 2001). Hence the research on learning progressions represents a systematic effort to synthesize the ideas from multiple strands of research into frameworks for scaffolding students in the deep understandings required for scientific literacy.

Learning progressions hold great promise for the science education community. They can harmonize and coalesce multiple aspects of the educational system by their focus on a common framework that is informed by core socio-scientific ideas and by knowledge of how students learn. Standards and large-scale assessments have identified which science topics to teach and curricula have outlined how to teach these topics. However, while students' misconceptions have informed curricula and standards documents (e.g., AAAS, 1990, Davis & Krajcik, 2005), learning progressions go further in that they focus on how students learn these topics (Alonzo, 2011). In the study of learning progressions, students' ideas become, for the first time, an essential component of a framework that guides these multiple aspects of science education.

CONSIDERATION OF CHALLENGES IN LEARNING PROGRESSION WORK

Despite the exciting potential that learning progressions offer, we realized that many researchers lacked a shared definition of learning progressions and a shared vision of their critical features. Therefore we organized an interactive poster session at the 2008 annual meeting of the American Educational Research Association (AERA) entitled *Diverse Perspectives on the Development, Assessment, and Validation of Learning Progressions in Science*. In this session, our goal was to highlight both the promise that learning progressions held and the lack of a shared definition of this concept. While there were similarities in the work on learning progressions presented at this session, there were also significant differences in how researchers conceptualized and used learning progressions. We observed that researchers who took diverse approaches to the development of learning progressions and associated curricula and assessments were using the same language to describe dissimilar studies of and experiences with learning progressions. In addition, these researchers were conducing learning progression work using very different methods. We also realized there was little communication among learning progression researchers about their efforts and results.

The interactive poster session allowed us to make comparisons among multiple projects in terms of how learning progressions are conceptualized. However, AERA poster sessions are not conducive to the deep discussions needed to move the field forward. And, as is typical of most conference presentations, the posters presented at the AERA session tended to highlight the successes—rather than the challenges—of learning progression work.

Following the AERA symposium, we organized the Learning Progressions in Science (LeaPS) conference (funded by the National Science Foundation [NSF]) that was held June 24–26, 2009, in Iowa City, IA (http://www.education.msu.edu/projects/leaps/). Eighty-two science educators, scientists, curriculum developers, assessment specialists, psychometricians, policy makers, and teachers attended the conference, where presentations were made and discussions were held about learning progression work with a specific focus on its challenges.

We structured the conference around four strands of learning progression work:

1. *Defining learning progressions* (the construct of learning progressions and the conceptualization of student progress);
2. *Developing assessments to elicit student responses relative to a learning progression* (the multiple ways to elicit evidence of students' knowledge and practices);
3. *Modeling and interpreting student performance relative to a learning progression* (the inferences made about students' learning progression levels based on their responses to assessment tasks); and
4. *Using learning progressions* (the many ways learning progressions may influence science education, including the design of standards, curricula, and teacher education).

These strands of learning progression work usually overlap and therefore cannot be pursued independently. However, separation of the strands reduces the complexity of the issues involved and allows for a more organized conversation about the challenges of learning progression work.

The LeaPS conference provided a structured forum for discussing the challenges associated with these four strands of learning progression work in both plenary and strand-specific sessions. The plenary sessions offered participants the opportunity to learn about the ongoing work of addressing challenges in the four strands. Three plenary sessions highlighted work in the defining, assessing/ modeling, and using strands. Richard J. Shavelson gave a keynote address in a fourth plenary session. In addition, a graduate student poster session showcased the work being undertaken by early-career scholars across the four strands. Each LeaPS conference participant selected one strand and attended strand-specific work sessions. Strand leaders facilitated working sessions in each strand. In these sessions, participants shared their work, discussed strand-specific issues, and suggested ways of addressing strand challenges. The strand leaders shared key ideas from the strand-specific work sessions at a final plenary session that allowed all participants to hear the ideas from the four strands.

It is important to note that the ideas generated at the LeaPS conference do not represent a consensus of all participants. While many ideas had wide agreement, our intent as conference organizers was not to push for consensus. We agree with the report from the Consortium for Policy Research in Education (CPRE; Corcoran, Mosher, and Rogat, 2009) that states that, at this stage, learning progressions are "… potentially important, but as yet unproven tools for improving teaching and learning… developing and utilizing this potential poses some challenges" (p. 5). However, in contrast to the meetings convened by CPRE, which resulted in this widely cited report, the main purpose of the LeaPS conference was not to achieve consensus. Rather, since learning progression work is still in its early stages, we thought it was important to explore a diversity of approaches. Forcing consensus too early may limit the successes that could come from learning progression research. Thus the main contributions of the conference, and, we hope, this book, are the descriptions of challenges researchers face in learning progression work and the approaches they develop to work with and around these challenges.

THE CURRENT STATE OF LEARNING PROGRESSIONS

Since the LeaPS conference, learning progressions have grown in popularity. Articles about learning progressions have appeared more frequently in journals. For example, the *Journal of Research in Science Teaching* devoted a special issue to learning progression research (Hmelo-Silver & Duncan, 2009). Besides funding much of the learning progression research, the National Science Foundation recently sponsored a "footprint" conference[1] to assess learning progression work in science and learning trajectories in mathematics and to make recommendations for future work. In addition, learning progressions have begun to influence national policies. The NRC (2010) included "prototype" learning progressions in its draft framework for the development of science education standards. Although the final version of its *Framework for K-12 Science Education* (NRC, 2011) did not include these learning progressions, research on learning progressions informed much of the new design of the *Framework*, and this document calls for increased learning progression research that may inform the development of future standards.

Learning progression research has advanced considerably in the last five years. Many researchers, educators, and teachers now recognize the potential of learning progressions throughout the educational system. However, challenges remain that require resolution before learning progressions can be effectively incorporated into a comprehensive framework for science education. Shavelson and Kurpius (chapter 2) discourage pushing learning progressions in science education prematurely into "prime time"; they caution that additional learning progression research is still required. We view this book as a significant contribution towards addressing these challenges such that learning progressions can fulfill their promise for widespread impact on science education.

CREATION AND STRUCTURE OF THE BOOK

Selecting and Reviewing the Exemplar Chapters

This book evolved from presentations and discussions at the 2009 LeaPS conference. For formal presentation at the conference (including both plenary sessions and strand-specific sessions), we selected 23 proposals from the 38 proposals submitted. Of the 23 conference presentations, we selected 12 presentations for inclusion in this book. We asked two author groups to combine the ideas from their presentations into joint chapters. Therefore, conference presentations resulted in 10 of the 12 "exemplar" chapters that highlight research in the four strands of learning progression work. Thus, the overall acceptance rate from conference proposal to book chapter was 32%. We solicited two more exemplar chapters in order to describe the use of learning progressions in standards development and in large-scale assessment. Consistent with instructions for the conference proposals, we asked the authors to feature challenges from their work on learning progressions as an integral and focal part of their chapters.

After at least one exchange of editorial comment and feedback with each chapter author, we sent all exemplar chapters to three or four external reviewers. We provided these reviewers with a vision statement for the book that highlighted the theme of challenges in work on learning progressions. We asked the reviewers to evaluate how well the chapters identified challenges in learning progression work and how critically the chapters considered the work presented. In addition, we asked the reviewers to rate the clarity and coherence of the writing, to give the chapters an overall rating, and to make comments pertinent for revisions. We asked the authors to respond to the reviewers' comments and to make revisions as needed. We reviewed the revised chapters before accepting them for inclusion in the book.

Reviewing the Other Chapters

In addition to the 12 exemplar chapters, the book contains this introductory chapter and two framing chapters. There is also a synthesis chapter for each strand and a conclusion chapter. Strand leaders[2] wrote the synthesis chapters that were revised in a feedback process with the book's editors.

Format of the Book

Section I: Framing section. After this introductory chapter (chapter 1), two framing chapters (chapters 2 and 3) set the tone for the book. Shavelson and Kurpius (chapter 2) recommend that a cautious view should be taken of the recent excitement generated by learning progressions in science education. They argue that researchers in science education should critically examine learning progression work to ensure that the resulting products (learning progressions and associated tools) live up to their promise. Krajcik (chapter 3) responds to this recommendation and describes learning progression research that can advance science education.

Following the Framing Section, there are four sections based on the four strands of learning progression work (Defining, Assessing, Modeling, and Using) used at the LeaPS conference. Each section has four chapters: three exemplar chapters on strand-specific challenges in learning progression work and one chapter that synthesizes ideas from the exemplar chapters and the discussions at the LeaPS conference.

Section II: Defining learning progressions. Chapters 4–7 describe the challenges associated with defining learning progressions. Defining learning progressions involves identifying a big idea or core concept and being explicit about what progresses as students develop more sophisticated knowledge and/or practice. There is significant variation in how different projects define learning progressions. Learning progressions have been developed for both content and scientific practices; thus what constitutes "progression" differs by project. The chapters in this section describe how students learn to provide scientific accounts of water and carbon in socio-ecological systems (Gunckel, Mohan, Covitt, & Anderson); how students learn to coordinate observations and explanations of celestial motion (Plummer); and how students learn to engage in scientific modeling practices (Schwarz, Reiser, Acher, Kenyon, & Fortus). Chapter 7 synthesizes these challenges and highlights the often implicit decisions made in defining learning progressions.

Section III: Assessing learning progressions. Chapters 8–11 describe the challenges associated with developing assessments that elicit student responses relative to learning progressions. Learning progression assessments are created and used for different purposes—for example, to validate the learning progressions or to evaluate student learning using formative and summative classroom assessments as well as large-scale assessments. In all learning progression assessments, the goal is to develop tasks that may be used to validly and reliably place students at a given level of a learning progression. The chapters in this section describe challenges in the following contexts: gathering evidence about students over a wide age range and across cultures and languages (Jin & Anderson); designing assessment tasks based on a learning progression that contains both content and practices (Gotwals, Songer, & Bullard); and working within (or possibly changing) existing large-scale assessment systems whose purposes and existing structures are not necessarily aligned with the purposes and design of learning progression assessments (Alonzo, Neidorf, & Anderson). Chapter 11 synthesizes these challenges in learning progression assessments and examines other issues in the design of such assessments.

Section IV: Modeling learning progressions. Chapters 12–15 describe the challenges psychometricians face when modeling (and interpreting) student performance relative to a learning progression. Because learning progression researchers have a more complex view of student thinking than the "gets it"/ "doesn't get it" perspective, new measurement approaches may be required to interpret student responses to assessment items with respect to the underlying learning progression. The learning progression chapters in this section explore measurement models based on Bayesian Networks (West et al.), Attribute Hierarchy Modeling (Briggs & Alonzo), and Item Response Theory (Wilson).

Chapter 15 synthesizes these modeling efforts and identifies common themes. This chapter also examines the role of grain size and misfit in modeling student responses with respect to learning progressions and explains the importance of including considerations of modeling in all aspects of learning progression work.

Section V: Using learning progressions. Chapters 16–19 describe challenges in the use of learning progressions for various purposes. Researchers have proposed that learning progressions can be used as tools in the development of standards and curricula and in teacher preparation and professional development. However, guidelines for translating learning progressions into tools for specific purposes and audiences are still being developed. The chapters in this section explore the requirements of and challenges inherent in this work: designing curricular resources to support student progress with respect to a learning progression (Wiser et al.); designing learning-progression-based tools useful for supporting development of ambitious teaching practices (Furtak, Thompson, Braaten, & Windschitl); and using learning progressions in the design of state standards (Foster & Wiser). Chapter 19 synthesizes these challenges in using learning progressions and describes themes and issues related to learning-progression-based products. This chapter also argues for the development of learning-progression-based tools, discusses the role of misconceptions in learning progression levels, and addresses the decision of when learning progressions are ready to use.

Section VI: Concluding section. In Chapter 20, the book's editors summarize the four strands of learning progression work. The chapter also identifies major cross-strand themes in order to make recommendations for how learning progression work might advance through more collaborative research, contributions to policy conversations, and interactions with other stakeholders. The editors emphasize that while the book focuses on the challenges of learning progression work, the science education community should not lose sight of the promises that learning progressions hold. It is only by addressing these challenges that learning progressions can have a significant impact on science education.

GOAL OF THE BOOK

As work on learning progressions is still in its early stages, it is doubtful that all the ways of defining, assessing, modeling, or using learning progressions have been identified. In fact, it is important to explore multiple options for addressing learning progression challenges. Therefore, our hope is that this book, with its focus on identifying and addressing such challenges, will stimulate further interest in learning progression research and model ways of addressing challenges in this complex work.

ACKNOWLEDGEMENTS

The LeaPS conference and this book were supported by the National Science Foundation under grant #DRL 0824432. We offer special thanks to our National

Science Foundation program officer, Janice Earle. The Iowa Math and Science Education Partnership and the Iowa Measurement Research Foundation provided additional funding for the LeaPS conference. Suzanne Wilson, Department Chair of Teacher Education at Michigan State University, arranged funding support for the language editing of the book. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation or other funding sources.

Our advisory board members and our strand leaders played critical roles in structuring this book through their conference assistance and their chapter contributions. Each advisory board member provided guidance about a particular learning progression strand: Carol L. Smith (Defining), Charles W. (Andy) Anderson (Assessing), Mark Wilson (Modeling), and Joseph Krajcik (Using). The strand leaders guided discussions at the LeaPS conference and wrote synthesis chapters. Alicia Alonzo and Amelia Wenk Gotwals were strand leaders for the Assessing and Using strands respectively. Lindsay Mohan and Derek Briggs were strand leaders for the Defining and Modeling strands, respectively.

In addition, LeaPS conference participants—see Appendix A—contributed to the discussion of challenges in learning progression work. Their ideas influenced the chapter authors and helped shape the synthesis chapters. Comments by chapter reviewers greatly improved the book. These reviewers are listed in Appendix B. In addition, we thank Marcia Halvorsen, whose careful language editing helped us achieve consistency in the chapters and comprehensibility of complex ideas. We also thank David Zinn, who created the LeaPS logo.

NOTES

[1] Charles Anderson, Principal Investigator; DUE-1132562.
[2] Julia Plummer also contributed to the Defining Learning Progressions synthesis chapter.

REFERENCES

Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research and Perspectives, 9,* 124–129.

American Association for the Advancement of Science. (1990). *Project 2061: Science for all Americans*. New York, NY: Oxford University Press.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher, 34*(3), 3–14.

Gonzales, R., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2004* (NCES 2005–005). Washington, DC: U.S. Government Printing Office.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*, 259–281.

Hmelo-Silver, C., & Duncan, R. G. (Eds.). (2009). Learning progressions [Special issue]. *Journal of Research in Science Teaching, 46*(6).

Linn, M., Tsuchida, I., Lewis, C., & Songer, N. B. (2000). Beyond fourth grade science: Why do US and Japanese students diverge? *Educational Researcher, 29*(3), 4–14.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Retrieved from the National Academies Press website: http://www.nap.edu/catalog/11463.html

National Research Council. (1996). *National science education standards.* Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

National Research Council. (2010, July). *A framework for science education: Preliminary public draft*. Retrieved from the National Academies website: http://www7.nationalacademies.org/bose/Standards_Framework_Homepage.html

National Research Council. (2011). A *framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Roseman, J. E., Kesidou, S., Stern, L., & Caldwell, A. (1999). Heavy books light on learning: AAAS Project 2061 evaluates middle grades science textbooks. *Science Books & Films, 35*(6), 243–247. Retrieved from Project 2061 website: http://www.project2061.org/publications/textbook/articles/heavy.htm

Schmidt, W. H., McKnight, D. C., & Raizen, S. A. (1997). *Splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer Acadmic Publishers.

Valverde, G.A., & Schmidt, W. H. (1997). Refocusing U.S. math and science education. *Issues in Science and Technology, 14*(2), 60–66.

RICHARD J. SHAVELSON AND AMY KURPIUS

REFLECTIONS ON LEARNING PROGRESSIONS

> [The Center on Continuous Instructional Improvement] views learning progressions as potentially important, but as yet unproven tools for improving teaching and learning, and recognizes that developing and utilizing this potential poses some challenges.
>
> Corcoran, Mosher, and Rogat (2009, p. 5)

Learning progressions have captured the imagination and the rhetoric of school reformers and education researchers as one possible elixir for getting K-12 education "on track" (Corcoran et al.'s metaphor, 2009, p. 8). Indeed, the train has left the station and is rapidly gathering speed in the education reform and research communities. As we are concerned about this enthusiasm—and the potential for internecine warfare in a competitive market for ideas— we share the Center on Continuous Instructional Improvement's view of the state-of-learning-progressions as quoted above. Even more, we fear that learning progressions will be adapted to fit various Procrustean beds made by researchers and reformers who seek to fix educational problems. We believe that learning progressions and associated research have the potential to improve teaching and learning; however, we need to be cautious—learning progressions are especially vulnerable to data fitting in the manner depicted in the *Non Sequitur* cartoon (Figure 1). As with any innovation, there are both promises and pitfalls associated with a learning progression reform agenda. Moreover, we fear that the enthusiasm gathering around learning progressions may lead to preferential treatment of one solution when experience shows single solutions to education reform come and go, often without leaving a trace. The best of intentions can go awry.

With this preamble, it is understandable that the LeaPS conference[1] organizers—Alicia Alonzo and Amelia Gotwals—would invite this chapter's first author to keynote the conference as a *friendly curmudgeon* who would raise issues and concerns about the ability of learning progressions to keep the train "on track." As veterans of formative assessment, learning progressions, and cognitive research on learning and memory, we have learned firsthand how tricky it is to attempt to model cognition and the multitude of differences among individuals. For example, in his doctoral dissertation, Jeffrey Steedle (2008; see also Steedle and Shavelson, 2009) revealed how fragmented students' knowledge structures are in explanations of force and motion. Knowledge comes in pieces that seem to be cobbled together in a particular context that calls for a particular explanation; this cobbled-together

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 13–26.

Figure 1. Procrustean fitting of data to (learning progression) theory. NON SEQUITUR © 2009 Wiley Miller. Dist.By UNIVERSAL UCLICK. Reprinted with permission. All rights reserved.

explanation may or may not align neatly with a learning trajectory (e.g., diSessa, 1988). More problematic, imposing a particular learning trajectory on data leads to misinterpretations and mis-prescriptions for teaching. As we discuss in this chapter, there is likely no single linear path within and across students' knowledge structures that has the potential to provide a tidy learning progression and prescriptions for teaching.

We have learned from experience how appealing and (superficially) compelling innovative teaching practices can be, especially when implemented by teachers who have their own conceptions of teaching and learning. Research in the Stanford Education Assessment Laboratory (SEAL) on formative assessment, which incorporated a learning progression for students' learning about sinking and floating, led to the following conclusion that appeared in a special issue of the journal *Applied Measurement in Education*:

> After five years of work, our euphoria devolved into a reality that formative assessment, like so many other education reforms, has a long way to go

> before it can be wielded masterfully by a majority of teachers to positive ends. This is not to discourage the formative assessment practice and research agenda. We do provide evidence that when used as intended, formative assessment might very well be a productive instructional tool. Rather, the special issue is intended to be a sobering call to the task ahead. (Shavelson, 2008, p. 294)

We have also discovered how learning progressions can derail the train by reinforcing naïve conceptions and by prematurely imposing constraints on instruction and cognition that ultimately may not be advantageous. For example, with respect to naïve conceptions, the SEAL research on sinking and floating followed a middle school science inquiry unit (Pottenger & Young, 1992) that was sequenced in a manner consistent with scientists' evolving explanations of sinking and floating: from mass to volume to volume and mass to density to relative density. One major, unintended consequence of the curricular learning progression approach was that the unit reinforced the mass explanation of sinking and floating, complicating subsequent conceptual development and conceptual change.

With respect to the premature imposition of constraints on instruction, SEAL research (discussed below) tested competing models of cognitive progression—a learning progression and a knowledge-as-pieces conception of growth (Steedle & Shavelson, 2009). We found that constraining students' ideas to the learning progression led to clumping incommensurate beliefs about force and motion into a single level. Using the learning progression in teaching, then, might work for some students identified at a given level but not other students with a similar level diagnosis. The evidence, rather, supported the knowledge-as-pieces conception in which students cobble together sets of beliefs into a "model" that they use to explain a phenomenon in a particular situation; the cobbling might lead to a different model of the same phenomenon when surface features of the situation change.

In the remainder of the chapter we first present a simplified view of how the field of learning progression conceptualization and research is evolving along two strands: (a) curriculum and instruction and (b) cognition and instruction. Given the possibility of fragmentation, this view may say more about the perceivers than the perceived; we leave that judgment to the reader. We then discuss each strand, drawing lessons learned and proposing approaches for further research. Finally, we try to put the pieces together in a summary.

Before proceeding, it seems appropriate to attend to definitional matters. Along with a number of others who also attended the LeaPS conference, we had the good fortune to serve on the Planning Committee for the Science Framework for the 2009 National Assessment of Educational Progress (NAEP). We described a learning progression as "a sequence of successively more complex ways of reasoning about a set of ideas" and stated that learners move from novice to expert after extensive experience and practice. We added that "learning progressions are not developmentally inevitable but depend on instruction interacting with students' prior knowledge and construction of new knowledge." Moreover, we recognized that there was no one "correct order" of progression. We also noted that learning evolves in a "succession with changes taking place simultaneously in multiple interconnected ways." Finally we warned that learning progressions are "partly hypothetical and inferential since long-term longitudinal accounts of learning by individual students do not exist" (National Assessment Governing Board [NAGB], 2008, p. 90). We believe that this description constituted a reasonably accurate characterization of learning progressions and what was known in 2006 when the framework was being written.

Corcoran et al. (2009), reporting for a committee of researchers engaged in work on learning progressions, provided a more recent yet consistent definition of learning progressions based on a National Research Council (2007) report: "empirically grounded and testable hypotheses about how students' understanding of, and ability to use, core scientific concepts and explanations and related scientific practices grow and become more sophisticated over time, with appropriate instruction" (p. 8). Corcoran et al. (2009) also noted that the hypotheses describe pathways students are likely to follow as learning progresses, with the number and nature of such pathways empirically testable and influenced by instruction. These learning progressions are based on "research… as opposed to selecting sequences of topics and learning experiences based only on logical analysis" (p. 8).

There seems to be considerable overlap in the two definitions. Both characterize learning progressions as the sequence or growth of successively more complex ways of reasoning about a set of ideas. They both recognize the centrality of instruction in the evolution of the progressions. They both recognize that such growth is not simple but may take complex forms as learners move from novice to expert. And both definitions recognize the hypothetical character of learning progressions and the need for a strong research base on which to justify policy recommendations for widespread use of such progressions.

It is the hypothetical and under-researched nature of learning progressions that frightens us. It is premature to move learning progressions into prime time, as seems to be happening; significant empirical research is required to establish these progressions. When we think of each set of core ideas that might be the focus of learning progression research and subsequently incorporated into teaching and learning, the amount of research required is staggering. Moreover, by the time this research is completed, the policy and reform circus will have long ago taken down its tents and headed for another apparently greener pasture. Just what are we embarking on and recommending? Might it be premature? Or might we recognize the hypothetical nature of learning progressions, call for more research, but push ahead with the empirically-based revision of progressions in the meantime? That is a question we pose to our community as we move forward.

TWO ROADS TO LEARNING PROGRESSIONS

Robert Frost's (1916) well-known poem "The Road Not Taken" describes the choice a traveler faces when meeting a fork in a wood:

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth.
Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same.
. . .
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Like the traveler in the poem—although more simply—we face a choice between two roads: interrelated roads traveled by learning progression reformers and researchers. One appears more worn, but like the roads in Frost's poem, both are really worn about the same. It is the choice that makes all the difference.

We call the first road the curriculum and instruction road and the second road the cognition and instruction road. Fortunately, we are more than one traveler and do not have to choose (or should not choose) at a glance. President Obama's stimulus package (e.g., see http://www2.ed.gov/programs/racetothetop-assessment/index.html accessed November 3, 2010) has the potential to allow researchers and reformers the pursuit of both in order to see if, in fact, one of the two roads makes all the difference, whether both do, or whether neither does.

The Curriculum and Instruction Road

The curriculum and instruction road may be characterized by the development of instructional units on, say, living organisms (Lehrer & Schauble, 2000) or sinking and floating (Shavelson, Yin, et al., 2008); K-8 curricular specifications for, say, atomic structure (Smith, Wiser, Anderson, & Krajcik, 2006); or even content specifications spanning K-12 science (e.g., Valverde & Schmidt, 1997).

To be sure, cognition is not omitted from the curriculum and instruction road. Yet we believe that curriculum and instruction progressions are based largely on logical analysis of content structure—perhaps a kind of spiral curriculum as envisaged by Jerome Bruner (1960) in *The Process of Education*. This logical content analysis is combined with what we call "psychologizing" as to how students might develop ideas cognitively.[2] Yet such psychologizing is always limited to the person engaged in this process. When concrete data are brought to bear on the cognitive processes students employ, complication and surprises arise. This is evident as students "think aloud" when they wrestle with solving a problem or explaining why things sink and float.

Perhaps an example of a learning progression that follows the curriculum and instruction road would be helpful. In SEAL research on the use of formative assessment in teaching about sinking and floating (e.g., Shavelson, Young, et al., 2008), we posited a learning progression that followed the series of investigations described in *Foundational Approaches in Science Teaching* (Pottenger & Young, 1992; see Figure 2). The dependency of the learning progression on teaching and learning is evident in the performance of two students, one from a "successful" guided-inquiry teacher (Gail) and another from an "unsuccessful" open-ended discovery teacher (Ken). Gail's student appears to follow the learning progression; Ken's student does not. Rather, Ken's student is mired in the conception that heavy things sink and light things float. That is, Gail's guided-inquiry teaching provided empirical support for the learning progression, but Ken's open-ended discovery teaching did not.

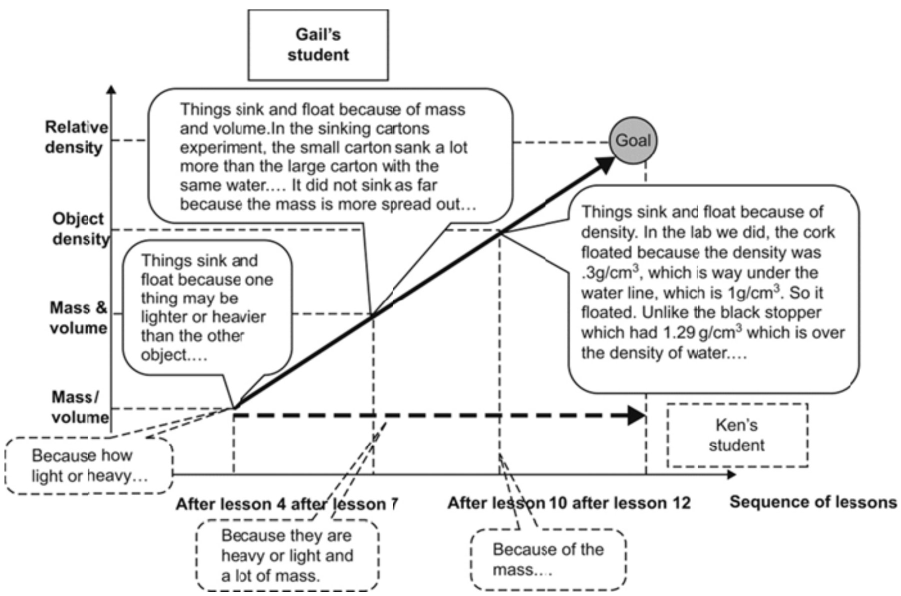


Figure 2. Learning progression for sinking and floating differs for two students: one student taught by a "successful" guided-inquiry science teacher (Gail) and one student taught by an "unsuccessful" open-ended discovery science teacher (Ken). From "On the Role and Impact of Formative Assessment on Science Inquiry Teaching and Learning," by R. J. Shavelson, Y. Yin, E. M. Furtak, M. A. Ruiz-Primo, C. C. Ayala, D. B. Young, M. L. Tomita, P. R. Brandon, and F. Pottenger in Assessing Science Learning: Perspectives from Research and Practice by J. E. Coffey, R. Douglas, and C. Stearns (Eds.) (p. 34), 2008, Arlington, VA: National Science Teachers Association. Copyright 2008 by the National Science Teachers Association. Reproduced with permission of National Science Teachers Association via the Copyright Clearance Center.

With a few exceptions, learning progressions following the curriculum and instruction road have not been empirically validated, at least in the strong sense that each learning progression posited has been researched, replicated, and validated as described by Corcoran et al. (2009). Empirical validation of learning progressions might be obtained through cognitive workshops, short essays, predict-observe-explain probes, teaching experiments, and the like that elicit students' explanations of natural phenomena (e.g., why do things sink and float?). Indeed, SEAL research suggests that context—in this case teacher and teaching method—will greatly influence the validity of a learning progression interpretation of student performance.

Even though learning progressions following the curriculum and instruction road have seldom been adequately validated empirically, we need to follow this road to the development of learning progressions for a number of reasons. Logical analysis and psychologizing can only take us along the road; empirical research can help guide us. But given the immensity of the curriculum, how might we accomplish the kind of self-correcting research needed to fine-tune and validate learning progressions? We don't know, but we have a proposal—one that might be surprising. We believe that teaching experiments and action research with collaborating teacher and researcher teams might amass the evidence and provide the practical wisdom needed to study and refine learning progressions. (Our proposal contrasts with that of Shavelson, Phillips, Towne, and Feuer, 2003, who argue that such experiments are only a beginning and need to be replicated on large scale). We envision such teams working on particular progressions, learning what does and does not work, fine-tuning the progressions, and making their findings available to others working on the same progression. In this way, we might expand both our knowledge of developing and validating learning progressions and our practice in using them. If we assemble a critical mass of teams working on important learning progressions, we might jump start the research and development agenda and create enough replications to evaluate the validity and utility of the proposed progressions. We would then be in a position to know if this is a road worth taking as we logically analyze and psychologize learning progressions. All of this might then lead to new and improved methods for studying learning progressions, which seem a practical necessity.

The Cognition and Instruction Road

While the curriculum and instruction road begins with a logical analysis of content, the cognition and instruction road begins with a psychological analysis of cognition underlying content—what does it mean to understand core ideas in science? How can we use knowledge about cognition to build instruction that improves the chances of all students learning at high levels?

There is a long tradition in the psychological analysis of cognition related to subject-matter learning; studies by David Ausubel (1963), Robert Gagne (1965), Jerome Bruner (1966), and Robert Glaser (1963) are early examples. The goal of this work is to map the growth of cognition as a student learns about particular concepts, such as force and motion. That is, what does the path look like as a student, over time, moves from naïve conceptions of force and motion to expert conceptions consistent with the understanding accepted by the scientific community? Most importantly, what do the paths look like between novice and expert, and how might they inform curriculum, teaching, and assessment?

More recently, Mark Wilson (2009) and Alicia Alonzo and Jeffrey Steedle (2009) have mapped learning progressions from a cognitive perspective. A third example, a learning progression for force and motion—specifically for explaining constant speed—is shown in Table 1. The progression describes what the student knows and can do when confronted by force and motion phenomena, more specifically when a force is and is not present, and when the object is and is not in motion. That is, the learning progression maps a cognitive progression for "understanding" force and motion from naïve (Level 1) to expert (Level 4).

Table 1. Force and Motion Learning Progression.

Level (Facets)	Description and Expected Responses to Item Types
4 (00)	When balanced forces act on an object, the object is either at rest or moving with a constant speed. When unbalanced forces act on an object, the object's speed changes
	Balanced forces: When balanced forces act on an object, it is either at rest or moving with a constant speed
	No force: After a force is removed, an object slows down because of friction, which acts in the direction opposite motion
	Constant motion: An object is moving with constant speed when forces are balanced
	No motion: An object remains at rest when a horizontal force is equal to an opposing friction force. The force of gravity is equal to the upward force for an object at rest on a surface
3 (30, 70)	When balanced forces act on an object, the object is at rest or slowing down. An unbalanced force in the direction of motion is needed to maintain constant speed. Speed is proportional to applied force
	Balanced force: When balanced forces act on an object, it is either at rest or slowing down
	No force: After a force is removed, a force continues to act on an object as it slows down
	Constant motion: A constant net force or unbalanced force or force of motion maintain constant speed
	No motion: *Same as level 4*
2 (90)	No motion implies that no force is acting on an object. Exception: gravity may act on objects at rest. Motion implies that a force is acting on an object
	Balanced force: *Same as level 3*
	No force: *Same as level 3*
	Constant motion:
	No motion: A horizontal force on an object that remains at rest is zero. No force or gravity only acts on an object at rest on a surface
1 (80)	If an object is pushed horizontally and remains at rest, there must be a greater force keeping the object at rest
	Balanced force: When balanced forces act on an object, it is moving at a constant speed
	No force:
	Constant motion:
	No motion: An object remains at rest when a horizontal force is not great enough to overcome a larger friction force, gravity, or the inertia of the object. The force of gravity is not equal to the upward force for an object at rest on a surface

Note. From "Supporting Valid Interpretations of Learning Progression Level Diagnoses," by J. T. Steedle and R. J. Shavelson, 2009, Journal of Research in Science Teaching, 46, p. 707. Copyright 2009 by Wiley Periodicals, Inc. Reproduced with permission of Wiley Periodicals, Inc. via Copyright Clearance Center.

An issue with this kind of learning progression is whether it accurately reflects cognition. Put another way, does students' knowledge actually grow in this linear, progressive way? Put still another way, does the progression provide a valid and practically useful way of portraying the pathway of cognitive development? By valid we mean whether students' knowledge actually grows in this way. By useful we mean that if their knowledge does grow this way, can the progression inform curriculum development, classroom teaching, and assessment?

There is another way to conceive of the pathway from naïve to expert understanding of a core science conception. It builds on two principles in cognitive science. The first principle is that knowing and doing are embedded in a cognitive network. The second principle is that memory is reconstructive. Together these principles lead to the hypothesis that when confronted by a natural phenomenon and posed a problem, students will construct an explanation that is context-dependent, drawing on bits and pieces of knowledge embedded in a memory network to reconstruct their knowledge and, thus, to provide an explanation. Note that if students at different places in the evolution from naïveté to expertise have bits and pieces of knowledge organized in a coherent linear manner, their cobbled-together explanations would most likely follow a linear learning progression, such as the one shown in Table 1.

But suppose students' knowledge is not so orderly. Suppose they have bits and pieces of loosely related knowledge about force and motion in their cognitive networks, garnered from extensive personal experiences and brief classroom encounters. In this case, their explanations will most likely be quite context-specific; if superficial characteristics of the problem change, we suspect students would change their explanations in ways not explicated by the learning progression in Table 1. Progress might not be neat and linear, although our statistical and qualitative modeling might force it, Procrustean style, into something neat and linear. Rather, progress from novice to expert might be better conceived as somewhat hectic and non-linear. If we conceive of memory as a complex network, at various times a student might make progress by building up bits and pieces of knowledge about force and motion into a small subnet, but other bits and pieces might still be unconnected. Of course, students might vary on which subnets they develop and which bits and pieces of knowledge lie scattered in memory. Depending on the context of a force and motion problem, an appropriate subnet might be accessed by one group of students but not by other groups.

If knowledge comes in bits and pieces, then the knowledge appears organized and coherent only when a high level of competence is reached. Anything less than expertise gives rise to multiple "mental models" and explanations for the same underlying phenomenon by the same person under different contexts. And if this is so, prescriptions based on a linear learning progression might not be accurate; if inaccurate, they might be heuristic at best and misleading at worst.

Jeffrey Steedle's (2008) doctoral dissertation provides examples of our concern. He examined the extent to which students' responses to force and motion test items fit a learning progression. He made this examination for three different learning progressions dealing with conceptions in force and motion, including constant speed as shown in Table 1. He used multiple-choice item data where the alternatives included naïve conceptions or "facets" of understanding from Jim Minstrell's Diagnoser (Minstrell, 2000). In a Bayesian latent class analysis of the data, comparing models based on the learning progressions and models based on "knowledge in pieces" in a cognitive network, Steedle and Shavelson (2009) report:

> Students' actual response patterns aligned with the proposed learning progressions for two sorts of students: those whose understanding is (nearly) scientifically accurate and those [naïve students] who believe that velocity is linearly related to force. Learning progression diagnoses for these levels could be interpreted validly (with few caveats), but diagnoses for the other levels could not because students diagnosed at those levels are not expected to consistently express the ideas associated with their learning progression levels … This suggests that it is not feasible to develop learning progressions that can adequately describe all students' understanding of problems dealing with Explaining Constant Speed. Finally, an analysis of relationships between learning progression levels and facet classes indicated that the confirmatory [learning progression] model failed to make important distinctions between latent classes that the exploratory [knowledge in pieces] model made. (p. 713)

Therefore, Steedle and Shavelson (2009) conclude:

> Students cannot always be located at a single level of the learning progressions studied here. Consequently, learning progression level diagnoses resulting from item response patterns cannot always be interpreted validly. It should be noted that the results presented here do not preclude the possibility that some individuals systematically reason with a coherent set of ideas. These results do, however, provide strong evidence that there are few substantial groups of physics-naïve students who appear to reason systematically about the forces acting on objects with constant speed. Further, these results corroborate findings from other physics education research indicating that many physics-naïve students should not be expected to reason systematically across problems with similar contextual features. (p. 713)

There is, then, evidence gathered on the cognition and instruction road that gives us pause as we proceed in the pursuit of learning progressions. This evidence suggests re-thinking how we conceive of learning progressions or even if learning progressions are the "right" way to think about the growth of students' knowledge. Indeed, the evidence supports the not-so-tidy definition of learning progressions used in the NAEP 2009 Science Framework (NAGB, 2008). Progressions are not developmentally inevitable but depend on instruction interacting with students' prior knowledge and new knowledge construction; there is no one "correct order" for the progression. That is, progressions evolve in a succession of changes that take place simultaneously in multiple, interconnected ways. Progressions are, to date, partly hypothetical and inferential since long-term longitudinal accounts do not exist for individual learners.

Perhaps conceiving of knowledge growth as a learning progression, let alone attempting to order levels in a learning progression, needs some re-thinking. Rather, conceiving of knowledge growth as a hectic, opportunistic, constructive process of cobbling together bits and pieces of knowledge, as Steedle's (2008) dissertation suggests, might prove to be beneficial as we attempt to assist teachers in building students' understanding of the natural world. Then we would need to figure out how the bits and pieces evolve into coherent models of the natural world with instruction.

CONCLUDING COMMENTS

The first author was asked to act as a friendly curmudgeon at the LeaPS conference in order to raise issues and ask questions as the learning progression train gathers steam and leaves the station. If we have accomplished anything, it has been to be curmudgeon-like. Our overriding concern is that an inadequately tested idea for improving curriculum, teaching, and assessment is being moved into prime time prematurely. We state this concern with full recognition that the learning progression concept has legs. If the concept is not developed in practice, it will languish in researchers' arcane journals. Nevertheless, we warn that there is the potential that a premature rush to implementation may result in more unintended mischief than intended good at this point.

We must, for example, guard against fitting our data to a preconceived notion of a learning progression. Rather, in a Popperian sense, we should seek disconfirmation; only when we fail should we move the progression into prime time. Even at this point, we need to monitor how well the progression works and agree to modify it as evidence demands.

We also need to make a concerted effort to gather evidence from the field that learning progressions embedded in curricular materials are operating as intended. We posed one possible approach that would move this agenda forward—that of teaching experiments and action research conducted by collaborating teacher-researcher teams. Such teams, on a large scale, might gather the empirical evidence and provide the practical wisdom needed to refine and improve learning progressions. Teams can work on particular progressions, learn what works and what does not, fine-tune the progressions, and make their findings available to others working on the same progression. We trust that those conducting learning progression research will think of other ways to address this area of concern.

A concerted effort should also be made to ensure that cognitive interpretations of learning progressions are accurate, useful, and lead to intended learning with minimal unintended consequences. Learning progressions may not be nice and linear. Teachers need to know this as researchers pursue heuristic representations of progressions to assist in practice, with an expectation of evolution and correction through research and practice over time. It seems that progress from novice to expert may not be linear but may be better conceived as a wandering journey through a complex memory network comprised of bits and pieces of information. Students might be nested in non-linear subnets for particular contextual representations of a problem. Steedle's (2008) research suggests a methodological approach for guarding against imposing theory on data by testing theory—our notion of a particular learning progression—with data. Both substantive psychological theory building and research into learning progressions are needed urgently for the most important science conceptions in the curriculum. A concerted research effort is needed. We again trust that those in the learning progression community will think of other ways to address this area of concern.

We have one final curmudgeonly thought. Whatever we come up with as a learning progression research and development agenda for reform, *it must take into account the capacity of teachers to implement.* The four million teachers in the United States are not, in general, like the teachers who volunteer to work with researchers in developing and testing cutting-edge ideas. It is well known that the former group of teachers lack, on average, the critical content knowledge needed to use learning progressions. They also lack the time needed to acquire that knowledge so that they may address the challenges that emerge when students do not nicely and neatly follow the prescriptions of the progressions and the textbooks. Whatever we do needs to take this reality into account; teacher professional development may not be extensive enough to address this challenge. So, finally, we trust that learning progression researchers will also think of ways to address this area of concern.

In closing, we have discussed two roads taken in the pursuit of learning progressions. In truth, the two roads don't diverge in a yellow wood nearly as much as Frost's roads. Rather, they continually intersect at the point of instruction. So the final challenge is to merge these roads as a major highway of coherent research to support the policy engine that is now steaming down the track… can we even catch up before it derails?

ACKNOWLEDGEMENTS

We would like to thank Alicia Alonzo and Amelia Gotwals for inviting the first author to address the Learning Progressions in Science (LeaPS) conference. We would also like to thank Jeffrey Steedle and Alicia Alonzo for their comments on earlier drafts of this chapter. They made invaluable suggestions. Any errors of omission or commission are, of course, ours.

NOTES

[1] The Learning Progressions in Science (LeaPS) conference took place from June 24–26, 2009, in Iowa City, IA.

[2] Incidentally, Bruner (1966) had a particular version of psychologizing in building curriculum. Curricular materials should move from initially enactive (physical manipulation) to iconic (mental image of physical manipulating) to symbolic (symbol replaces mental image).

REFERENCES

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

Ausubel, D. P. (1963). Cognitive structure and the facilitation of meaningful verbal learning. *Journal of Teacher Education, 14,* 217–221.

Bruner, J. S. (1960). *The process of education*. Cambridge, MA: Harvard University Press.

Bruner, J. S. (1966). *Toward a theory of instruction.* New York, NY: Norton.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Hillsdale, NJ: Lawrence Erlbaum Associates.

Frost, R. (1916). The road not taken. In *Mountain interval* (p. 9). New York, NY: Henry Holt and Company. Retrieved from Gutenberg Project website: http://www.gutenberg.org/files/29345/29345-h/29345-h.htm

Gagne, R. M. (1965). *The conditions of learning.* New York, NY: Holt, Rinehart and Winston.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 510–522.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in Instructional Psychology: Vol. 5. Education design and cognitive science* (pp. 101–169). Mahwah, NJ: Erlbaum.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44–73). Washington, DC: National Academy Press.

National Assessment Governing Board (2008, September). *Science framework for the 2009 National Assessment of Educational Progress*. Retrieved from the National Assessment Governing Board website: http://www.nagb.org/publications/frameworks/science-09.pdf

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.

Pottenger, F. & Young, D. (1992). *The local environment: FAST 1 Foundational Approaches in Science Teaching*. University of Hawaii Manoa: Curriculum Research and Development Group.

Shavelson, R. J. (2008). Guest editor's introduction. *Applied Measurement in Education, 21*, 293–294.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher*, *32*(1), 25–28.

Shavelson, R. J., Yin, Y., Furtak, E. M., Ruiz-Primo, M. A., Ayala, C.C., Young, D. B., . . . Pottenger, F. (2008). On the role and impact of formative assessment on science inquiry teaching and learning. In J. E. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice.* (pp. 21-26). Washington, Arlington, VA: NSTA Press.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P., Furtak, E. M., Ruiz-Primo, M. A., . . . Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*, 295–314.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

Steedle, J. T. (2008). *Latent class analysis of diagnostic science assessment data using Bayesian networks* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3313668)

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*, 699–715.

Valverde, G. A., & Schmidt, W. H. (1997). Refocusing U.S. math and science education. *Issues in Science and Technology, 14*(2), 60–66.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716–730.

JOSEPH S. KRAJCIK

THE IMPORTANCE, CAUTIONS AND FUTURE OF LEARNING PROGRESSION RESEARCH

Some Comments on Richard Shavelson's and Amy Kurpius's "Reflections on Learning Progressions"

Amelia Wenk Gotwals and Alicia Alonzo asked me to write a response to "Reflections on Learning Progressions" by Richard Shavelson and Amy Kurpius (chapter 2). I am pleased to contribute these remarks to further the discussion on learning progressions.

Shavelson and Kurpius, who describe themselves as "friendly curmudgeons," begin their reflections by warning the community not to adopt simple solutions by forcing data to fit learning progressions. This is a wise caution since force fitting of data, before we have learned all that we can from this research paradigm, will prove fatal to learning progression research. Such research will improve the learning and teaching of science only if the community conducts careful, systematic, and unbiased research.

When I first entered the world of education research in the mid-1980's, misconceptions research was in full bloom and, as learning progressions do now, offered the promise of guiding research in science education and of helping students learn science. Unfortunately, although some researchers conducted excellent research and advanced our knowledge, *everything* was interpreted as a misconception. Laundry lists of misconceptions were published and presented at conferences. If students gave incorrect responses to assessment items or in interviews, the conclusion was that they held a misconception about the idea. Any response that did not match canonical science was considered a misconception. Many researchers failed to recognize that students may have simply been unfamiliar with the particular idea being assessed or that the items used to assess understanding were poorly written. I raise this issue because it represents a good illustration of what Shavelson and Kurpius refer to as "Procrustean bed" methodology—when researchers force a pattern onto data to fit the main idea of science education research.

To avoid such force fitting of data to learning progressions, the community needs to critically monitor and evaluate its own work. Without such self-evaluation, like Shavelson and Kurpius, I fear that an abundance of learning progressions will be developed that has not been carefully researched. As a result, learning progression research will fail to produce findings that promote learning of core ideas in science.

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 27–36.

Learning progressions, like formative assessment, appeal to researchers, practitioners, and policy makers. However, this appeal should not allow for superficial development of ideas. Nor should it prevent systematic research and development to occur so that learning progressions can quickly be developed and used in schools. Individuals who conduct learning progression research know the difficulties and challenges of the work, especially since clear procedures for developing and testing learning progressions are underspecified. Of course, as Shavelson and Kurpius state, a challenging situation exists because policy makers, curriculum designers, and teachers wish to use learning progressions to guide assessment and curriculum materials development and instruction.

Creating learning progressions, unfortunately, requires years of systematic research. One major premise of learning progression work is that it requires empirical accounts of how students learn. Although some research findings exist that the community can draw upon in the design of learning progressions, much more is needed. As Shavelson and Kurpius state, while some initial learning progressions are based on research findings, these learning progressions require further development. Until the results from systematic research are available, the gaps in our knowledge need to be filled with our *best guesses.* I agree with this approach; however, in the absence of such research, rationales based on theoretical foundations for decisions need to be provided when articulating levels of learning progressions.

Although the community should heed this warning by Shavelson and Kurpius, it also needs to move forward. Because the support of students' learning of science has never taken a developmental perspective, science learning has suffered. Learning progressions focus the science education community on how ideas gradually become more sophisticated over time based on coherent unfolding of ideas, instruction, and prior experiences. Science education would benefit from a developmental perspective that extends beyond thinking about how science ideas progress based on the logic of the discipline to include how students can reason with core ideas based on the examination of more targeted experiences under optimal teaching and learning conditions.

Let's look at an example to see the value of learning progressions in designing curriculum materials. It is not unusual for curriculum materials to introduce ideas about the water cycle before students develop understandings related to the particle model of matter. Although students can certainly learn about the water cycle in a descriptive sense without understanding the particle model of matter, they cannot provide a causal mechanism that explains the water cycle in the absence of the particle model. Unfortunately, curriculum materials and teachers too quickly introduce the particle model, failing to build an evidence-based model that has explanatory power. What is so unfortunate about this instructional approach is that students perceive and memorize the particle model as fact rather than as a framework that can provide causal accounts of phenomena encountered in their daily lives. However, if instruction could help students develop the beginnings of a particle model, learners could use this model to explain phase changes. Then, when students explore ideas related to the water cycle, the particle model is reinforced and more connections are made. At this point, students develop a more integrated understanding that allows them to offer more sophisticated causal accounts of other everyday phenomena. This is the power behind learning progressions. Those of us in the community should not let this power slip away because of sloppy and rushed research.

As someone whose expertise is in curriculum development and the design of learning environments, I realize that the water cycle could drive the learning of the particle model of matter. I am not criticizing this approach since there are many contexts that could be used for learning core ideas; however, my point is that instruction cannot simply provide students with a model and expect them to see its power to explain phenomena. Instruction needs to help learners develop understanding of the model and see the potential the model has as a causal mechanism to explain and predict phenomena.

Learning progressions move away from the one-paragraph/one page description of content ideas that too often are presented in science education. Rather, learning progressions provide curriculum designers with the tools needed to purposefully build upon students' current understandings in order that they will form richer and more connected ideas over time (Margel, Eylon, & Scherz 2008; Merritt, Krajcik, & Shwartz, 2008). Curriculum materials, as Shavelson and Kurpius state, must not only build from the nature of the discipline but also from what is known about how students learn and reason. Fortunately, such materials are beginning to appear, for example, in research by Hannah Margel and her colleagues (Margel et al., 2008), Andy Anderson and his colleagues (Mohan, Chen, & Anderson, 2009) as well as in my work with various colleagues (Krajcik, Reiser, Sutherland, & Fortus, 2008; Merritt, 2010; Merritt et al., 2008). These materials were carefully researched in classrooms to examine if and how they support student learning. In many respects, these materials were developed as *teaching experiments* (the term used by Shavelson and Kurpius). I do not claim that one curriculum fits all teaching situations or that curriculum materials are an essential aspect of learning progression research. Rather, it is my contention that curriculum developers can take advantage of learning progression research to design curricula with a developmental focus. Such a focus can provide learners with opportunities to build more connected and sophisticated understandings as they examine related ideas in new and more complex situations.

This narrative, which describes how I envision the essence of the learning progression approach, matches the descriptions of learning progressions by Shavelson and Kurpius (chapter 2) and by Corcoran, Mosher, and Rogat (2009). A learning progression provides a sequence of successively more complex ways of reasoning about a set of ideas as learners move from novice to experts as they engage in key learning experiences (National Assessment Governing Board, 2008). Perhaps science education has created a theoretical and research paradigm that can drive the learning of science in the years ahead. However, the research community needs to ensure that good development and research work continue.

Next I highlight some important aspects of the design and research of learning progressions that, although presented in the Shavelson and Kurpius chapter, require further articulation.

First, learning progression work is based on a solid foundation of theoretical ideas. In some respects, researchers are in a unique position as far as science education since the cognitive and learning sciences communities have provided key theoretical underpinnings for learning progressions. *Taking Science to School* (National Research Council, 2007) and *How People Learn* (National Research Council, 1999) articulate what is known in the field about student learning. Never before has the science education community possessed such a solid and robust understanding of how to advance student learning.

The structure of expert knowledge as tightly connected around core ideas that drive thinking, observing, and problem solving is a key that informs learning progression research. This understanding suggests that only a few learning progressions for core/big ideas in the disciplines require careful development. The National Research Council identified a few core ideas that drove their work in the design of the new conceptual framework for K-12 science education.[1] This does not mean that there is only one trajectory for each core idea since learning is influenced by instruction and personal experience. However, to help students develop and refine their ideas, a few hypothetical learning progressions for core ideas should be developed. Thereafter classroom research/teaching experiments should further refine and articulate each learning progression. Unfortunately, very little empirical work is available to help guide the design of learning progressions around these big ideas. For instance, I question if the idea of density should form a learning progression. Certainly, density is a useful idea, but it is not a big idea that helps explain a host of phenomena (Stevens, Sutherland, & Krajcik, 2009). In contrast, a big idea might be:

> The properties of matter can change with scale. In particular, as the size of a material transitions between the bulk material and individual atoms or molecules it often exhibits unexpected properties that lead to new functionality – generally at the nanoscale – a material often exhibits unexpected properties that lead to new functionality. (Stevens, et al., 2009, p. 37)

As this big idea is developed into a learning progression (and this is only a hypothetical case), perhaps the idea of density may become part of the learning progression.[2] If this is the case, then researchers need to learn how to support students in building the idea of density by determining which key phenomena, experiences, and analogies are useful to this learning. For the project Investigating and Questioning our World through Science and Technology [IQWST] Krajcik, Reiser et al. (2008) describe an activity a classroom teacher created that seemed useful for developing middle school students' understanding of density. In this activity, there are two boxes of equal size: One is filled with Styrofoam packing peanuts and the other with books. Although the box with the books has much greater mass than the box with Styrofoam packing peanuts, their volumes are the same. The teacher then invites different students to lift the boxes. This multimodal experience helps students build connections they will remember. The teacher uses this activity to help students realize that two variables, mass and volume, are important in the density of materials. Building this understanding is critical to understanding density. This example points to an important research component in developing, refining, and articulating learning progressions: Once learning progressions are developed, the community needs to conduct research to gather learning data for various segments of the progressions. If density is part of the learning progression on the properties of materials, then it is essential to identify key instructional activities that can develop student understanding.

Second, I agree with Shavelson and Kurpius that learning progressions are not linear, that students need to revisit ideas under new contexts to refine their understanding, and that ideas often link across learning progressions. For instance, ideas related to force and interactions intersect with learners' understanding of ideas related to transformations of matter. Without building certain key ideas in forces and interactions, a learner can only proceed so far in understanding transformations of matter and can only develop a descriptive model of transformation instead of a mechanistic model.

In my view, there are four requirements for learning progressions. First, the big idea should be identified and explained. This step involves unpacking the big idea (see below for additional ideas about unpacking). Second, the learning progression should be clearly described at each level. This description should explain the student reasoning expected at each level, prerequisite understandings that often are part of the previous level, and the links to related ideas. Research-based articulations of the difficulties and challenges learners experience as they move to higher levels are essential. As Anderson (2008) notes, the levels should relate logically from the learner's perspective not necessarily based only on the logic of the discipline. As part of considering the learner's perspective, we should also examine how students learn and how to identify the content they find engaging. Third, each learning progression should include psychometrically validated assessment items that can identify students at a particular level. Fourth, each learning progression should include classroom-tested instructional components—key phenomena, analogies, and tasks—to use in advancing learners to the next level of understanding. These instructional components are not curriculum materials, but they provide teachers and curriculum developers with instructional components that are useful for building curriculum materials.

A learning progression for the particle nature of matter will require that students understand that gases are matter and as such have mass. It has been well documented that many middle school students do not think of gases in this way. An instructional activity for this learning progression could be to ask students to determine the mass of a filled CO_2 cartridge, empty the CO_2 cartridge, and then to determine its mass again. The change in mass of the cartridge is evident. But one example or illustration is never enough to build understanding. This also works with a filled gas tank for a barbeque grill. Anyone who has filled a barbeque gas tank knows that the filled tank is much heavier than the empty tank. However, the difference in weight between the filled and empty tank comes to a surprise to many. Other ideas might work as well.

For instance, I spoke with Dr. Phil Johnson, a science educator and divisional director for postgraduate education at Durham University in England, about his work on learning progressions. He described how students determine that the mass of a closed system does not change when a substance transforms from the liquid stage to the gaseous stage (P. Johnson, personal communication, November 5, 2010). The critical point I am trying to make is that although there could be several key instructional activities to help students move from one level to the next, there is probably a finite set. These key instructional components are linked to the current level and to assessments associated with the learning progression and show how to support students moving to the next level.

These learning progression activities for the particle model of matter contain critical instructional components that support students constructing understanding to move to the next level. If such instructional components are not addressed, in my opinion, learning progression work is vacuous, as a learning progression by itself cannot support teachers or curriculum. I know a difference of opinion exists in the learning progression community about the value of these instructional components and whether they should be considered part of the learning progression or if they are instead part of a teaching progression. However, without them, I do not see how researchers could validate a learning progression; in my work, these instructional components are essential in learning progressions. In testing a learning progression, researchers need to examine how students progress when opportunities to learn arise; otherwise, research measures the result of instruction that is based, not on how students develop ideas across time, but on curriculum materials that may be less than optimal for supporting learning (Roseman, Stern, & Koppal, 2010). In some sense, the conversation about whether these instructional components are/are not aspects of the learning progression is misguided. To validate a learning progression, opportunities to learn are necessary.

The work by Anderson and colleagues on a learning progression for environmental literacy (e.g., Mohan et al., 2009) illustrates the importance of conducting learning progression work in classrooms that target big ideas. Many high school students don't achieve upper levels of environmental literacy because they lack the opportunity to learn the big ideas in the learning progression with the typical high school curriculum. However, high school students who have participated in teaching experiments have experienced coherent instruction take a different path and reach higher levels of performance. Hence, the validation of learning progressions should take place in environments where students have opportunities to learn the ideas in the learning progressions. The challenge is to develop learning environments in which the components work together—the learning progression, assessments and instruction. In the absence of one, you cannot make progress with the others. Other colleagues (e.g., Richard Lehrer and Leona Schauble) argue for the importance of linking professional development to learning progressions. I agree with the important role that professional development plays in helping teachers learn the new practices that learning progressions require. Without professional development, it is unlikely we will see such practices in classrooms.

Like good curmudgeons, Shavelson and Kurpius set up two non-intersecting roads traversed in learning progression work—the curriculum and instruction road and the cognition and instruction road. (My background and work on the design of curriculum materials and learning progressions seems to place me along the curriculum and instruction road). But I fear that Shavelson and Kurpius have over-simplified the situation. A number of researchers have created a third road that merges the first two roads—the cognition, curriculum, and instruction superhighway (see below).

In IQWST (Krajcik, Reiser et al., 2008), my colleagues and I think seriously about what students need to learn (Krajcik, McNeill, & Reiser, 2008). We rigorously address content ideas by identifying and unpacking science standards. Unpacking refers to breaking apart and expanding various concepts to elaborate the intended content. In unpacking a science standard, in addition to examining the importance of the content, we look at how students reason and how they use knowledge. We examine common student difficulties, prior understandings needed to build the target understanding, and aspects of prior conceptions that may pose difficulties. Thus unpacking a science standard helps articulate its content and helps identify the reasoning and use of knowledge appropriate for learners.

Once the key science ideas are identified and unpacked, we develop "learning performances" to articulate the cognitive tasks that students should accomplish with this content. The development of learning performances is a critical step because science standards are typically presented as declarative statements of scientific facts that do not specify the type of reasoning students should engage in with ideas. My work with Namsoo Shin and Shawn Stevens uses a similar process of unpacking, but we refer to learning performances as "claims" and identify the evidence students must show the claim is met (Shin, Stevens & Krajcik 2010).

Several other groups (e.g., Duncan, Rogat, & Yarden, 2009; Songer, Kelcey, & Gotwals, 2009) create learning progressions that blend key science ideas and practices with students' reasoning. Although these researchers may appear to be on the curriculum and instruction road, their work suggests they have split off somewhat from that road. They seem to be taking a new road that intersects both the curriculum and instruction road and the cognition and instruction road. Still other researchers (e.g., Lehrer & Schauble, 2008) appear to be transversing the cognition and instruction road as they unpack what it means to model with various mathematical ideas such as probability. Schauble and Lehrer also follow the same road that Anderson, Duncan, Songer, and I have taken (Krajcik, Reiser et al., 2008). This road is neither the curriculum and instruction road nor the cognition and instruction road but rather a blend—the cognition, curriculum and instruction superhighway. These research efforts represent the third road that Shavelson and Kurpius envision. The works cited clearly show this new road is already under construction and is being used. It is not yet the superhighway the field needs, but researchers are laying its foundation by blending the work of diverse groups who have expertise in the cognitive and the learning sciences, science or mathematics, science/mathematics education, and the teaching and learning of science/ mathematics. Researchers are blending knowledge of learning, classrooms, psychometrics and the sciences in order to explain and advance student learning. The community still has a long way to go to complete this superhighway, but this work is currently underway laying a foundation for improving teaching and learning.

As Shavelson and Kurpius envision, the curriculum and instruction and the cognition and instruction roads have come together, and the learning progression community needs to follow this road to make the type of progress the field needs. However, it will take the work of diverse individuals to build these learning progressions. My work with learning progressions (e.g., Shin & Krajcik, 2008) suggests that the psychometric expertise that individuals such as Mark Wilson and Jim Pellegrino bring to the construction site is required. The National Science Foundation (NSF), which is concerned with bringing psychometricians and science educators together, sponsored a parallel workshop to the LeaPS conference (see Duncan & Krajcik, 2008) to help young scholars in the field develop psychometric expertise.

Shavelson and Kurpius make several points in their closing remarks that I wish to re-emphasize. I agree that the community needs to prevent force-fitting data to preconceived notions about learning progressions. This won't be easy to avoid, and the community must be self-critical to avoid this tendency. In addition, the community should gather evidence from a variety of classrooms in which curriculum materials and instruction follow ideas linked to levels of a learning progression to determine if the learning progression operates as intended. This is challenging work since few curricula based on learning progressions now exist. When such research doesn't exist, it needs to be conducted or researchers and designers need to take their best guess based on theory and experience to design materials. Through testing these materials, researchers will learn how students' understandings develop. The NSF should also offer strong support of systematic and long-term research on learning progressions. If the current round of funding is not continued in future cycles, knowledge for the field will be developed, but it will be limited. Learning progression research, by its very nature, is longitudinal and it is unlikely that researchers will design optimal progressions and associated curriculum and assessment materials on their first attempt. Researchers will learn much from the data they collect that will feed back into revising the learning progression. Moreover, because learning progressions depend on instruction interacting with learners' prior knowledge, students who advance to higher grades, having experienced instruction in earlier grades based on learning progressions, should reason about phenomena differently compared to learners who have not had such instruction.

Finally, the idea of teaching experiments is consistent with the work in IQWST and the work by Andy Anderson, Ravit Duncan, and others. Working closely with teachers and researchers with different areas of expertise is critical to developing learning progressions and materials that can change how learning occurs in the classroom. Research teams should consist of individuals with expertise in teaching, psychometrics, the cognitive and learning sciences, science education, and the science disciplines. In addition, more teachers should be involved with these teaching experiences so that we can see how these ideas scale and if they can be used in a variety of classrooms.

The chapter by Rich Shavelson and Amy Kurpius points to some important considerations. Generally a curmudgeon is defined as a difficult, cantankerous person or a killjoy—a person who spoils the joy or pleasure of others. Despite their use of the word as a self-description, it certainly does not apply to Shavelson and Kurpius. To me, they are thoughtful and friendly critics or commentators who highlight some important ideas that advance the work of learning progressions. If at one time there was a curriculum and instruction road and a cognition and instruction road, they have now merged as a third road —the super highway. But like any road, and especially a super highway, it takes money, time and the collaborative efforts of individuals with diverse expertise to construct.

ACKNOWLEDGEMENTS

This chapter was written while I was a visiting faculty member at Ewha Womans University. This writing was partially funded by the National Science Foundation (NSF) under grant DRL-0822038 awarded to the Learning Progression Team at the University of Michigan led by Dr. Namsoo Shin. I would like to thank Dr. Namsoo Shin and Dr. Shawn Stevens for their contributions in helping me develop my ideas regarding learning progressions. The opinions expressed herein are those of the author and not of the NSF or of the Learning Progression Team. I would also like to thank Ann M. Novak for her comments on various drafts of this manuscript.

NOTES

[1] http://www7.nationalacademies.org/bose/Standards_Framework_Homepage.html

[2] Wilson (this volume) proposes one possibility for a larger learning progression that includes students' developing understanding of density.

REFERENCES

Anderson, C. W. (2008, February). *Conceptual and empirical validation of learning progressions: Response to "Learning progressions: Supporting instruction and formative assessment."* Retrieved from Consortium for Policy Research in Education website: http://www.cpre.org/ccii/images/stories/ccii_pdfs/learning%20progressions%20anderson.pdf

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence - based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Duncan, R., & Krajcik, J. (2008). *Moving beyond conceptual models: Developing assessments to validate and study learning progressions* (NSF, DRL-0830931).

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th–10th grades. *Journal of Research in Science Teaching, 46*, 655–674.

Krajcik, J., McNeill, K. L., & Reiser, B. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*, 1–32.

Krajcik, J., Reiser, B. J., Sutherland, L. M., & Fortus, D. (2008). *Investigating and questioning our world through science and technology*. Ann Arbor, MI: Regents of the University of Michigan.

Lehrer, R., & Schauble, L. (2008, September). *A modeling foundation for reasoning about evolution*. Presentation to the Consortium for Policy Research in Education Learning Progression Working Group, Ann Arbor, MI.
Margel, H., Eylon, B.-S., & Scherz, Z. (2008). A longitudinal study of junior high school students' conceptions of the structure of materials. *Journal of Research in Science Teaching, 45*, 132–152.
Merritt, J. D. (2010). *Tracking students' understanding of the particle nature of matter*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3441321)
Merritt, J., Krajcik, J., & Shwartz, Y. (2008). *Development of a learning progression for the particle model of matter*. Paper presented at the biennial International Conference of the Learning Sciences, Utrecht, The Netherlands.
Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.
National Assessment Governing Board (2008, September). *Science framework for the 2009 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/publications/frameworks/science-09.pdf
National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.
Roseman, J. E., Stern, L., & Koppal, M. (2010). A method for analyzing the coherence of high school biology textbooks. *Journal of Research in Science Teaching, 47*, 47–70.
Shin, N., & Krajcik, J. (2008). *Developing an empirically-tested learning progression for the transformation of matter to inform curriculum, instruction and assessment design* (National Science Foundation DRL-0822038).
Shin, N., Stevens, S. Y., & Krajcik, J. (2010). *Using construct-centered design as a systematic approach for tracking student learning over time*. London, UK: Taylor & Francis Group.
Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*, 606–609.
Stevens, S., Sutherland, L., & Krajcik, J. S., (2009). *The big ideas of nanoscale science and engineering*. Arlington, VA: National Science Teachers Association Press.

DEFINING LEARNING PROGRESSIONS

KRISTIN L. GUNCKEL, LINDSEY MOHAN, BETH A. COVITT AND CHARLES W. ANDERSON

ADDRESSING CHALLENGES IN DEVELOPING LEARNING PROGRESSIONS FOR ENVIRONMENTAL SCIENCE LITERACY

In a world where human actions increasingly affect the natural systems on which all life depends, we need educated citizens who can participate in personal and public decisions about environmental issues. The effects of global warming have wide-reaching ramifications. No longer can policy decisions be made by a select few. For example, decisions about how to distribute water so that urban, agricultural, and natural ecosystems have adequate water supplies or about whether to tax carbon emissions require that citizens understand scientific arguments about the effects of their actions. Scientists and policy makers are presenting results of scientific research directly to the public—for example, 2007 Nobel Prize Winners Al Gore (2006) and the Intergovernmental Panel on Climate Change (2007). A question that confronts us as science educators is how can we help the public respond to these reports and the debates around them: How can we prepare our citizens to engage directly in the collective human response to global climate change?

The purpose of the Environmental Literacy Project is to contribute to the preparation of citizens for participation in the necessary collective response to global and local environmental issues. We believe that citizens must have an understanding of underlying scientific models and principles in order to evaluate experts' arguments about environmental issues and recognize policies and actions that are consistent with their environmental values. Environmental science literacy requires an understanding of many aspects of science, including chemical and physical change, carbon cycling, water cycling, biodiversity, and evolution by natural selection. Although these phenomena are currently addressed in many state and national standards documents and in school curricula, typically they are addressed in disconnected ways—in different courses or in different units in the same course.

Research in the Environmental Literacy Project is divided into four strands. Research groups for three of these strands are working to develop and validate learning progressions leading toward connected and coherent understandings of three key aspects of socio-ecological systems: water (Covitt, Gunckel, & Anderson, 2009; Gunckel, Covitt, Dionise, Dudek, & Anderson, 2009), carbon (Jin & Anderson, this volume; Mohan, Chen, & Anderson, 2009), and biodiversity (Zesaguli et al., 2009). A fourth research group has investigated students'

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 39–75.

decision-making practices in citizenship roles (Covitt, Tan, Tsurusaki, & Anderson, 2009). The work on citizenship practices is relevant to the other three research strands of the project because it explores students' developing capacities to use their understanding of water cycling, carbon cycling, and biodiversity to make informed and responsible decisions about socio-ecological issues.

The goal of developing an environmentally literate citizenry defines the parameters of our learning progressions. Three parameters are of particular importance. First, developing environmental science literacy involves both cognitive and sociocultural aspects of learning. Citizens must understand the conceptual scientific models related to important environmental issues and be able to draw on their understanding of these models in order to participate in discussions and decision-making processes. These two aspects of learning mean that we must address such questions as: What cognitive models and types of reasoning are necessary for understanding environmental issues? What forms of participation are valued in a community? How does one become an informed and engaged member of a community? We chose to develop learning progression frameworks that describe changes in students' knowledge and practices as they progress toward environmental science literacy.

Second, our learning progressions cover broad scientific content areas. Even within the domain of any single strand of our research (e.g., water or carbon), the content involves multiple conceptual models. Furthermore, environmental science literacy involves making interconnections among the three content domains. Therefore, our learning progression frameworks do not focus on any single conceptual model (e.g., atomic-molecular theory). Rather, we chose to focus our learning progression frameworks on the changes in knowledge and practices that are apparent as students develop scientific, model-based views of the world.

Third, our learning progressions are not tied to a specific curriculum or curriculum materials since we want to describe the current progression of student thinking given the status quo curriculum and the current state of education. We seek to develop learning progression frameworks for a broad range of students—upper elementary through high school students in urban, suburban, rural, and international settings. Our learning progression frameworks have to account for diversity in student backgrounds and grade levels. Therefore, we use a large grain size in defining the steps in our learning progression frameworks in order to highlight the patterns in student thinking across this broad range of students.

In this chapter, we describe our responses to two challenges that these choices present to our work. One challenge is *defining what progresses in a learning progression* that spans a broad grade range of diverse students across three conceptual domains. We found that elementary students' accounts of phenomena rely on a different worldview and different word meanings than the scientific understandings we want students to develop by twelfth grade. This challenge has led us to use a Discourses perspective to define our learning progression frameworks. We describe this perspective in the context of the water cycle learning progression. Another challenge we have faced is *describing the role that instruction plays in defining a learning progression*. We have identified two possible pathways that students may take through our learning progressions and hypothesize that each pathway may be linked to different instructional approaches. We describe these pathways and approaches in the context of the carbon cycle learning progression.

CHALLENGE 1: WHAT PROGRESSES IN A LEARNING PROGRESSION?

Our approach to defining what progresses in our learning progressions has changed over time. During our initial rounds of learning progression framework development, we focused on determining what students did and did not understand about the big ideas in each domain. However, we soon discovered that we had difficulty in connecting the responses that younger students gave us to the responses that older students provided, and in connecting both younger and older students' ideas to the scientific, model-based reasoning of environmental science literacy without taking a deficit perspective on younger students' responses and thinking. For example, when we asked students if the rain near the ocean is salty, we received some responses from young students that included, "No, because the rain taste[s] the same," and "No, because it's filtered by the sky." Older students' responses included "No, because as it [the water] evaporates back into the clouds, the salt molecules are too heavy to evaporate as part of the water molecules." One interpretation may be that the younger students had misconceptions about how water and substances cycle through the atmosphere whereas the older students had conceptions that were closer to correct ones. However, this interpretation only told us what the students did not understand. It did not help us understand how the younger students' reasoning might become more like the older students' reasoning.

We realized that students reasoned about the problems we posed to them in very different ways than scientists do. We initially sought to make sense of the differences by contrasting informal narrative accounts with scientific, model-based accounts of phenomena. However, we eventually realized that both informal and scientific accounts may take the form of narratives; they are merely different types of narratives. Informal narratives tell stories about actors who make things happen. Scientific accounts also tell stories about phenomena. However, scientific narratives are constructed using scientific principles to constrain possible outcomes and explanations. Thus when asked what happens to a puddle, a student who provides an informal narrative may say that the sun dried up the water. A student who uses a scientific narrative will recognize that the water does not disappear but instead changes state. This student will say that the water in the puddle evaporated to become water vapor in the air.

Furthermore, our decision to focus on the sociocultural aspects of learning meant that we had to account not just for students' conceptions and reasoning but also for how they participate in various communities. Members in different communities provide different types of accounts of phenomena that are based on the norms for talking and interacting within a community and on the purposes of the account (Driver, Asoko, Leach, Mortimer, & Scott, 1994). In talking with their parents at home, children may use informal, everyday narratives of the world that would be insufficient and unacceptable to scientists grappling with scientific problems. To address these challenges, we turned to the relationships among Discourses, practices, and knowledge to organize our learning progression frameworks and to track changes in student thinking.

RELATIONSHIPS AMONG DISCOURSES, PRACTICES, AND KNOWLEDGE

From the perspective of Discourses, practices, and knowledge, learning is conceptualized as the process of mastering a new Discourse (Cobb & Hodge, 2002; Wenger, 1998). Discourses are the ways of talking, thinking, and acting that identify a socially meaningful group. Discourses are enacted in communities through the practices in which members of the community engage (Gee, 1991). Participating in the practices of a community, in turn, requires specific knowledge. Figure 1 shows the embedded relationship of knowledge in practices in Discourses. Tracking students' progress as they learn new Discourses requires tracing changes in student knowledge as students engage in new practices.

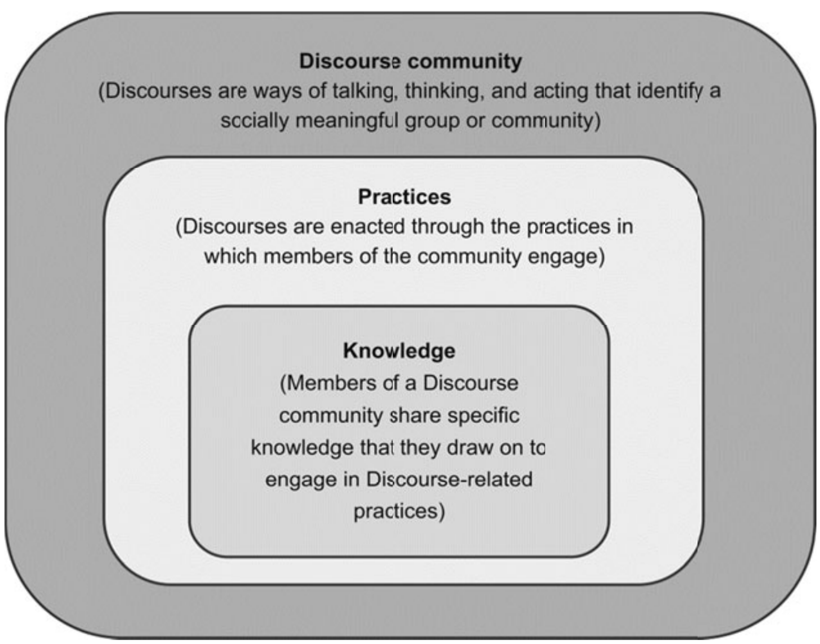


Figure 1. Embedded relationships among Discourses, practices, and knowledge.

Discourses. We describe the patterns of language use that define the perspectives, values, and identities that link people in social networks as Discourses. These Discourses provide the lenses through which people see and make sense of their world.

People participate in many different communities during their lives and can thus draw on many Discourses. They begin life with the primary Discourse of their home communities. "All humans… get one form of discourse free, so to speak… . This is our socioculturally determined way of using language in face-to-face communication with intimates" (Gee, 1991, p. 7). As people expand their communities of participation, they learn new, or secondary, Discourses.

> Beyond the primary discourse, however, there are other discourses which crucially involve institutions beyond the family… Let us refer to these institutions as secondary institutions (such as schools, workplaces, stores, government offices, businesses, or churches)… Thus we will refer to them as "secondary discourses." (Gee, 1991, p. 8)

Students' primary Discourses define the lower end of our learning progression frameworks. The process of learning involves mastering the ways of talking, thinking, and acting associated with secondary Discourses. The target secondary Discourse for the Environmental Literacy Project learning progression frameworks is the Discourse of environmentally literate citizens capable of using scientific reasoning in their roles as democratic citizens (Covitt, Gunckel, et al., 2009; Covitt, Tan, et al., 2009; Mohan, Chen, & Anderson, 2009). For example, with respect to the water cycle, environmentally literate citizens participate in the collective decision-making processes necessary to maintain and protect adequate fresh water quality and quantity for people and the natural ecosystems on which they depend. Below we describe the primary and secondary Discourses that are relevant to our learning progressions.

Primary Discourse: Force-dynamic reasoning. Students' primary Discourses provide insight into how they make sense of their world. Understanding students' primary Discourses is about more than just determining what students do and do not know about the world; this understanding also involves students' world views and experiences.

Although there are many different primary Discourses rooted in diverse sociocultural communities, one common feature that they share is a force-dynamic approach to explaining the events of the world.[1] Linguist Stephen Pinker (2007) and developmental psychologist Leonard Talmy (1988) argue that there is a "theory of the world" built into the basic grammar of all languages. They label this theory of the world force-dynamic reasoning. We must learn this theory in order to speak grammatical English; this theory, in turn, shapes how we view and explain events.

> There is a theory of space and time embedded in the way we use words. There is a theory of matter and causality, too. . . . These conceptions . . . add up to a distinctively human model of reality, which differs in major ways from the objective understanding of reality eked out by our best science and

> logic. Though these ideas are woven into language, their roots are deeper than language itself. They lay out the ground rules for how we understand our surroundings. (Pinker, 2007, p. vii)

Pinker notes that this structure exists in many languages, not just English. Thus characteristics of how students make sense of the world are rooted in the grammatical structure of the language of their primary home Discourse. Recognizing these characteristics in students' primary Discourses allows us to use patterns in students' language structures to look across students' diverse social and cultural home communities and find common patterns in their ways of thinking about the world.

Force-dynamic reasoning explains the events of the world in terms of cause and effect relationships between objects with "intrinsic tendencies and countervailing powers" (Pinker, 2007, p. 219). Characteristics of force-dynamic reasoning include:

- *Actors and abilities*. The events of the world are largely caused by actors in accord with their abilities. Human actors have the most abilities, followed by animals, then plants. Dead things have no abilities, so they are acted on by other actors. Non-living entities such as machines can be actors with limited abilities. Depending on the situation, water can also be an actor, such as when a river carves a canyon.
- *Purposes and results.* Actors have goals or purposes, and the results of events are generally the fulfillment of the actors' purposes. Higher-level actors can have many purposes, so animals grow, move, think, etc. Lower-level actors have fewer purposes—for example, the main purpose of a tree is to grow. While inanimate materials such as water do not have purposes, they do have "natural tendencies" to move toward their appropriate places in the world. One such tendency of water, for example, is to flow downhill.
- *Needs or enablers.* In order to use their abilities and fulfill their purposes, actors have to satisfy certain needs. For example, a tree needs soil, water, air, and sunlight to grow. Conversely, actors can also have inhibitors or antagonists that prevent them from fulfilling their purposes. Thus a concrete sidewalk inhibits water from soaking into the ground. Water can also be an enabler or an inhibitor for another actor, such as the person who needs clean water to drink.
- *Events or actions.* The events of the world occur when actors have all their needs met or all required conditions are present. For example, water can flow from one lake to another lake if a river connects them.
- *Settings or scenes* for the action. Finally, there are settings or scenes for the action. These settings include air, earth, stones, etc. They provide the background landscape or the stage on which actors act or events happen. Water is often the background landscape against which other events happen.

In force-dynamic reasoning, the ultimate outcome of an event, or an action by the actor, is the result of the interplay of what can broadly be described as "forces"—forces that support the action through enablers, and forces that hinder the action through antagonists.

Secondary Discourse: Scientific reasoning. The secondary Discourse of environmentally literate citizens that defines the upper end of our learning progression frameworks embodies a different type of reasoning from the force-dynamic reasoning of students' primary Discourses. The Discourse of environmentally literate citizens relies on scientific reasoning, which views all phenomena as taking place in connected and dynamic systems that operate on multiple scales and are constrained by fundamental scientific principles. Scientific reasoning relies on models that are grounded in observations (data) and that are applied consistently to explain the events of the world (Anderson, 2003; National Research Council, 2007; Sharma & Anderson, 2009). For example, model-based reasoning about water in socio-ecological systems involves recognizing that water and other substances are parts of connected systems (e.g., watersheds, groundwater, municipal water systems, etc.) and that the movement of water and other substances through these systems is constrained by principles such as the law of conservation of matter and the law of gravity.

Practices. Discourses are enacted through the practices of the communities in which people participate (Cobb & Hodge, 2002; Wenger, 1998). We define practice as a pattern of activity that is engaged in repeatedly. Discourses shape or mediate the activities in which members of a group participate (Cobb & Hodge, 2003; Wertsch, 1991).

We are interested in the practices that are essential for environmentally responsible citizenship: investigating, accounting (explaining and predicting), and deciding (Figure 2). These are the practices all citizens engage in when making decisions and acting in public and private roles:

- Public roles: voter, advocate, volunteer, elected official
- Private roles: consumer, owner, worker, learner

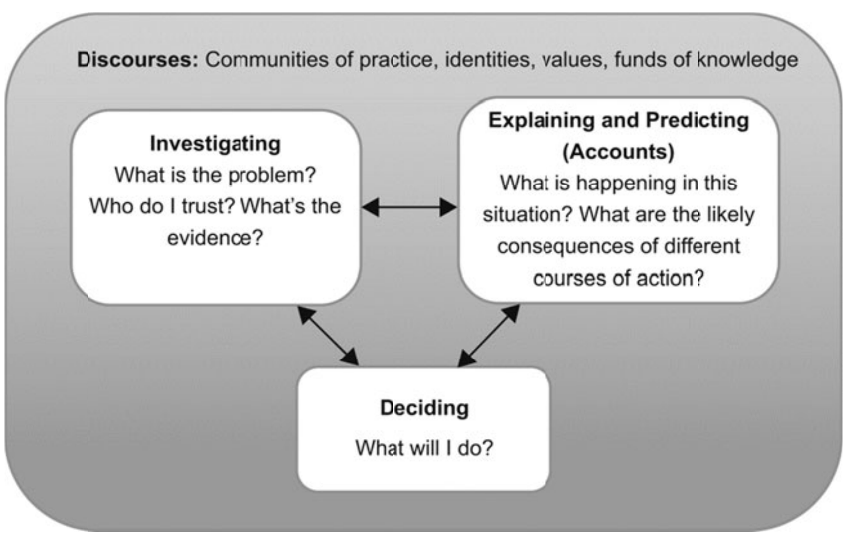


Figure 2. Citizenship practices (Covitt, Tan, et al., 2009).

How people in their various roles engage in these practices depends on the Discourses of the communities in which they are participants. People who participate in communities that use the secondary Discourse of scientific reasoning engage in these practices in ways that represent environmentally responsible citizenship. We would like students to become informed citizens who are aware of the possible environmental consequences of their actions and who will take those consequences into account.

Citizens' decisions and actions can always—and should—be based on considerations and values other than scientific knowledge and environmental consequences. Environmental science literacy is about giving people real choices— helping them understand possible alternative actions and their consequences— rather than leaving them trapped by ignorance. The citizenship practices we are interested in are:

1. *Inquiry (Investigating):* Inquiry involves investigating issues and deciding whom to trust. Environmentally literate citizens learn from experience and use evidence to construct and evaluate explanations. They evaluate both sources of evidence and the evidence itself. For example, citizens must be able to learn about and understand the specifics of particular water quality and supply issues and situations. They must be able to identify and understand pertinent evidence and then analyze and evaluate the quality of the evidence and arguments presented by multiple stakeholders. In contrast, people engaging in inquiry practices embedded in non-scientific Discourses may be limited to investigating issues by considering social rather than scientific information (Fleming, 1986). They may rely on immediate factual claims rather than evaluate those claims in conjunction with scientific theories and content knowledge learned in school (Kolstø, 2006). Furthermore, in deciding whom and what to trust, people using non-scientific Discourses use strategies such as "thinking for themselves" and evaluating the motivations, interests, and biases of different sources (Kolstø, 2001) without considering the relevance and validity of evidence presented by sources with perceived "suspect" interests. From a scientific perspective, one should not assume that just because evidence is, for example, presented by a large corporation that it is, by virtue of its source, automatically invalid.
2. *Accounts:* Accounts involve the practices of explaining and predicting outcomes of processes in socio-ecological systems.
   - *Explaining* processes in systems. Environmentally literate citizens must combine scientific and socio-scientific models and theories (i.e., general knowledge) with specific facts of the case (i.e., local knowledge) to explain what happens to water in the socio-ecological systems in which they live and how these systems are affected by human actions. People using non-scientific Discourses explain processes using informal knowledge rooted in family experience, popular culture, and popular media. As such, their explanations often differ greatly from scientific explanations.

- *Predicting* effects of disturbances or human policies and actions on processes in systems. When making informed decisions, citizens must use their understanding of socio-ecological systems to make predictions about the potential consequences of possible actions on the local water system. While predictions are always complicated by limited information and uncertainty, scientists use specific strategies (e.g., calculating confidence intervals) for dealing with uncertainty. In contrast, in their day-to-day lives, few people consciously engage in weighing uncertainties when making predictions (Arvai, Campbell, Baird, & Rivers, 2004). Instead, people generally rely on heuristic principles (i.e., intuitive judgments; Tversky & Kahneman, 2000) when predicting likely outcomes of different actions. Nevertheless, with instructional support, even children in second grade are capable of conceptualizing multiple types of uncertainty in scientific investigations (Metz, 2004).

3. *Deciding*: Decision-making involves conscious or unconscious choices about personal lifestyles or courses of action in private roles and about people or policies to support in public roles. Decisions related to socioscientific issues do not depend just on science: ultimately, such decisions also depend on personal values (Kolstø, 2006). Thus scientific values cannot determine our decisions, but our decisions can be informed by scientific knowledge and practice. Scientifically-informed decision-making involves using science as a tool to support the practices identified in Figure 2: investigating, explaining, predicting and deciding.

In the Environmental Literacy Project, we have done some work on students' inquiry and deciding practices (Covitt, Tan et al., 2009). Our primary focus, however, has been on students' accounting practices: explaining and predicting.

Knowledge. Citizenship practices for environmental science literacy require that citizens understand and use knowledge. Such knowledge ranges from understanding general principles, such as the law of conservation of matter, to specific knowledge of local situations. Figure 3 shows the domain of general knowledge about water in socio-ecological systems necessary for environmentally literate citizens if they are to engage in the practices described above. The boxes in Figure 3 show the environmental systems and human social and economic systems that comprise a global, connected socio-ecological system. The arrows connecting the boxes show that the systems in the boxes do not exist in isolation. Human social and economic systems depend on natural systems for fresh water; the decisions and actions that take place within the human social and economic systems have a significant impact on the quality and distribution of water in environmental systems.

The Loop Diagram for water in socio-ecological systems (Figure 3) depicts knowledge that we believe students should have upon graduation from high school. How students think about and understand the systems and processes through which water and substances in water move is the focus of our learning progression research. The next section presents the learning progression framework that describes the knowledge and practices that students bring to learning about water in socio-ecological systems and how their knowledge and practices change through their experiences in school.

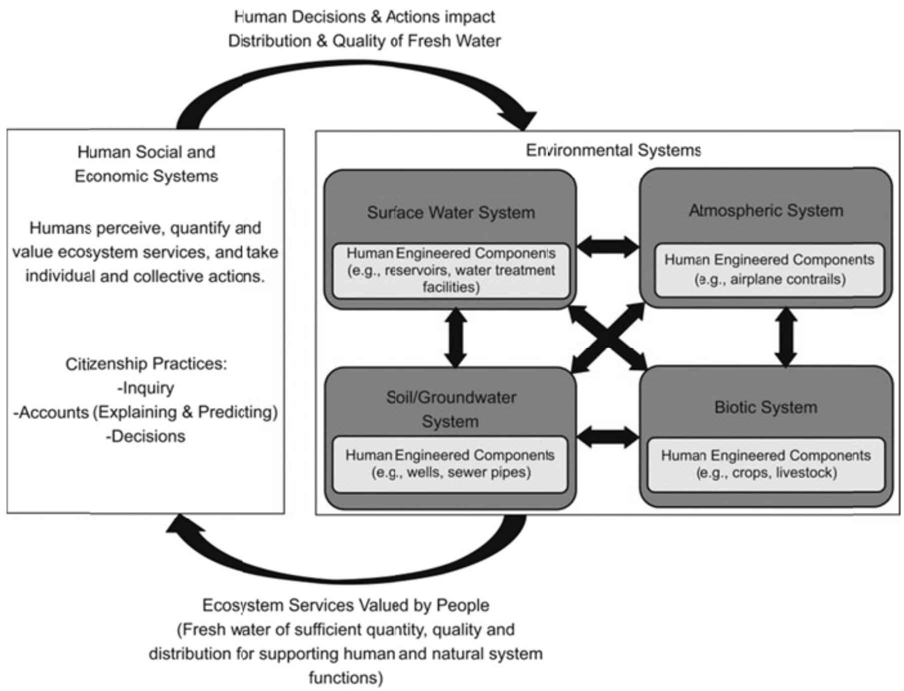


Figure 3. Loop Diagram for water in socio-ecological systems. Adapted from the Loop Diagram from the Long Term Ecological Research Network (Long Term Ecological Research Planning Committee, 2007)

USING RELATIONSHIPS AMONG DISCOURSES, PRACTICES, AND KNOWLEDGE TO BUILD LEARNING PROGRESSION FRAMEWORKS

The perspective of Discourses, practices, and knowledge has been helpful in incorporating students' knowledge and practices into our learning progression frameworks focusing on carbon, water, and biodiversity in socio-ecological systems. Our development work thus far has focused primarily on students' accounts: their approaches to explaining and predicting phenomena (see Figure 2). In this section, we explain how we used this perspective to build and test our learning progression frameworks. We begin by explaining our development methodology. We then provide an overview of the components of our learning progression frameworks and describe the learning progression framework for water in socio-ecological systems. We end with a description of some of the current challenges we face.

Methods. The development of our three learning progression frameworks has followed an iterative design research process (Barab & Squire, 2004; Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Design-Based Research Collective, 2003). For each strand of our research, we began by developing a Loop Diagram (such as the one in Figure 3 for the water strand) and identifying the key conceptual understandings that environmentally literate citizens must have. We call these understandings the Upper Anchors of our learning progression frameworks. We then developed initial assessment items to probe students' thinking about these ideas. Assessments were administered to students in grades 4–12. The students' responses to an assessment item were pooled, and then a sample of responses was ranked from least sophisticated to most sophisticated. We used our Upper Anchors as the standard for ranking responses.

This process involved many discussions among the researchers as we debated what constituted a more or less sophisticated answer. For example, if a student said that the water in a puddle "dried up," was the student describing evaporation? Did the student believe that the water disappeared forever? Was the student describing an observation (e.g., the puddle is gone) as opposed to describing a process (e.g., drying is a process of becoming unwet)? If a student said that the water from a puddle "soaked into the ground," was that evidence that the student was tracing water, or did the student believe that water that soaked into the ground was gone forever?

We realized that the challenge was in understanding how students' language provided clues to their views of the world. We turned to the work of linguists such as Stephen Pinker and Leonard Talmy who have studied the connection between language and cognition. Eventually we were able to use the characteristics of force-dynamic and model-based reasoning as lenses to examine our data. By searching for characteristics of force-dynamic and model-based reasoning we began to see patterns in the rank ordering and to identify groups of student responses with similar characteristics. We were then able to identify features in student responses that changed from less to more sophisticated answers. We used these features to build an initial learning progression framework. Responses in the least sophisticated group were labeled the Lower Anchor and reflect the ideas and accounting practices representative of the primary Discourse that students bring to learning about water, carbon, and biodiversity. Groups of responses that were more sophisticated than the Lower Anchor answers and less sophisticated than the Upper Anchor answers were used to describe changes in student thinking across the initial learning progression framework. Once we had an initial learning progression framework, we continued to conduct successive rounds of assessment design, administration, and analysis to refine the learning progression framework. We followed this procedure for developing each of our learning progression frameworks (i.e., water, carbon, and biodiversity).

Components of a learning progression framework. All the learning progression frameworks in the Environmental Literacy Project have the same general structure—similar to the one presented in Table 1—as the learning progression framework for water in socio-ecological systems (Anderson, 2009). This structure uses a learning performance as a unit of analysis. Learning performances are students' responses to assessment items. They are organized according to progress variables and levels of achievement. The next sections describe these features in terms of the learning progression framework for water in socio-ecological systems and connect these features to relationships among Discourses, practices, and knowledge described above.

Table 1. General Structure of Learning Progression Framework for Water.

<table>
<tr><th rowspan="2">Levels of Achievement</th><th colspan="2">Progress Variables</th></tr>
<tr><th>Movement of Water</th><th>Movement of Substances</th></tr>
<tr><td>4: Qualitative model-based accounts</td><td></td><td></td></tr>
<tr><td>3: School science accounts</td><td rowspan="2" colspan="2">Learning performances for specific Progress Variables and Levels of Achievement</td></tr>
<tr><td>2: Force-dynamic accounts with hidden mechanisms</td></tr>
<tr><td>1: Force-dynamic accounts</td><td></td><td></td></tr>
</table>

Progress variables. Progress variables are aspects of accounts that are present in some form in all students' accounts and can be used to track changes in student reasoning across levels. Determining which aspects of accounts to use as progress variables has been a challenge because our learning progression frameworks involve complex domains. Progress variables are derived partly from theories about how knowledge and practice are organized and partly from our empirical research data (Briggs, Alonzo, Schwab, & Wilson, 2006; Draney, Wilson, Choi, & Lee, 2007; Wilson, 2005). Because knowledge and practices are organized differently for each of our learning progressions frameworks, progress variables of one learning progression framework do not necessarily map onto the progress variables of another learning progression framework. As a result, the progress variables identified for the water learning progression framework, described below, are different from the progress variables identified and described in the second half of this chapter for the carbon learning progression framework.

Accounts of water in socio-ecological systems explain and predict the movement of water and substances through multiple connected systems. In earlier versions of our water learning progression framework, our progress variables distinguished between students' understanding of the structure of connected systems and the processes that move water and substances through these systems. This organization, however, proved unproductive because we could not separate students' understanding of structure from their understanding of processes. We therefore decided to consider elements of a complete account as progress variables. A complete account of water in any of these systems (e.g., a groundwater system) traces both water and substances in water. Therefore, for the water learning progression framework, we chose to examine student progress in tracing water and in tracing substances in water as our progress variables.

1. Movement of Water – Describes how students identify and describe processes that move water across landscape-scale distances through connected systems. This progress variable includes whether students recognize and apply constraints on processes such as conservation of matter, gravitational control of water flow, and permeability of materials. It also includes students' understanding of the structure of the systems through which the water moves (e.g., a groundwater system, a surface water system).
2. Movement of Substances in Water – Describes students' conceptions of water quality and the ways in which students identify and describe processes that mix and move substances with water. It includes students' attention to the microscopic and atomic-molecular scales when describing substances in water and the processes that mix, move, and unmix substances. The progress variable describes whether students recognize and apply constraints on processes, including conservation of matter, gravity, and permeability of materials.

Levels of achievement. Levels of achievement are patterns in learners' performances that extend across progress variables (Mohan, Chen, & Anderson, 2009). We have identified four levels of achievement that track student accounts from a force-dynamic to a model-based view of the world. Levels 1 and 2 describe a force-dynamic Discourse. Level 2 represents a more fully developed force-dynamic account of events of the world than Level 1. Levels 3 and 4 describe the transition to a scientific model-based view of the world. Level 4 describes a more fully developed model-based reasoning than Level 3. The next section describes these levels in detail.

A learning progression framework for water in socio-ecological systems. The following description of the levels of achievement for the learning progression framework for water in socio-ecological systems was developed based on students' written responses to 20 assessment items addressing different aspects of hydrologic systems. Students in grades 2–12, from rural, suburban, and urban schools, responded to these items. The descriptions provided below use examples of students' responses from a subset of these 20 items.

Items focusing on movement of water:

1. *Puddles*: After it rains you notice puddles in the middle of the soccer field. After a few days you notice that the puddles are gone. Where did the water go?
2. *Bathtub*: Could the water (from the puddles) get in your bathtub?
3. *Groundwater*: Draw a picture of what you think it looks like underground where there is water.
4. *Water in Rivers*: How does water get into a river?

Items focusing on movement of substances in water:

1. *Water Pollution*: What are examples of water pollution?
2. *Salt Dissolving*: What happens to salt when it dissolves in water?
3. *Treatment*: Describe the different treatments that are used to make sure water is safe to drink.

4. *Ocean Water*: If you had to make ocean water drinkable, how would you go about doing it?
5. *Salty Rain*: If you live by an ocean, will your rain be salty? Why or why not?

Level 1: Force-dynamic accounts. Level 1 students explain and predict using the language of force-dynamic Discourse. Their accounts include the key characteristics of force-dynamic reasoning about the course of an event, including the setting, the actors, and their abilities, purposes, and needs. Actors can achieve their purposes if they have all the necessary enablers and if there are no antagonists or opposing actors. If there are antagonists, then the outcome depends on which actor has the greater powers.

Moving water: Water in the background landscape. Level 1 students describe water as part of the background landscape. Responses at this level do not account for what happens to visible water after it disappears from view. For example, Level 1 responses to the Puddles item included, “It got dried up by the sun.” Similarly, a response to the Bathtub item was, “No. It already disappeared into the air.” When asked to draw water in places they cannot see, such as underground, Level 1 students imagine water in locations they can see and translate those images to places they cannot see. For example, they draw pictures of groundwater as water in underground pipes or tanks.

Substances in water: Accounts of types or qualities of water. Level 1 students describe water quality in terms of types of water rather than other materials mixed with water. For example, one student’s examples of water pollution included, “Lake water, ocean water, sea water, well water, pond water.” Another student wrote, “black merkey [sic] water.” Level 1 students focus on visible features and on human actors as agents. Thus one response to the Water Pollution item focused on a human action rather than on matter: “Some examples are throughing [sic] bottles and pop cans.” When asked about materials in water that are not visible, Level 1 students tend to say that the materials have gone away. Answering the Salt Dissolving item, one student wrote, “the water overpowers the salt by making it disappear.” Level 1 students think of changes in water quality as something that changes water from one type to another and of water purification as something that human actors do without describing a specific process. For example, in response to the Ocean Water item, one student wrote, “I would not be happy because I would have to drink uncleaned [sic] water.” Another wrote, “Cleaning it and making sure it’s clean.”

Level 2: Force-dynamic accounts with hidden mechanisms. Level 2 students still explain and predict using force-dynamic reasoning but give more attention to hidden mechanisms in their accounts. They recognize that events have causes and often use simple mechanisms to explain or predict events. Students at Level 2 are beginning to trace water and substances, recognizing that water and substances that are no longer visible go someplace else.

Moving water: Natural tendencies with conditions. At Level 2, students still think about water as part of the background landscape, but their conception of the size of the background landscape is larger. Level 2 students think about the water in rivers as connected to water in other rivers and groundwater as layers of water underground. Level 2 students think about the movement of water as a natural tendency of water, and they identify possible enablers and antagonists to movement. For example, Level 2 responses to the Bathtub item included, "Yes. If it was a rainy day and if there were puddles saved from yesterday and you open the door it could go in to the bath tub [sic] then there would be puddles in the bathtub." And, "Yes. If you had a window in your bathroom like I do, if you happened to have it open it would condensate [sic]." These responses identify an action that a person must take to enable water to move from the puddle into the bathtub.

Substances in water: Objects and unspecified stuff in water. At Level 2, students recognize that water can mix with other materials. They think water pollution is the result of harmful things put in water, often by people. These harmful things may be objects (e.g., "garbage," "dead animals," "rotten food") or unspecified materials (e.g., "muck," "cemicals [sic]"). When materials are mixed with water, and the materials are no longer visible (e.g., salt dissolving in water), Level 2 students, like Level 1 students, may explain that the materials have disappeared. However, Level 2 students begin to provide novice explanations for tracing matter. Example responses to the Salt Dissolving item include explaining that the substances stay separated, "The salt will go to the bottom," or explaining that you will see a visible change, "The water changes color." Level 2 students also describe simple, macroscopic scale mechanisms to mix or unmix water and other substances. For example, one student responded to the Treatment item by writing that a filter "Takes the rocks and mud/dirt out of it." Level 2 students have difficulty tracing substances with water through invisible system boundaries. For example, some Level 2 students answered the Salty Rain item by suggesting that salty water evaporates and turns into salty rain. Another student suggested that salty water does not turn into salty rain because the water is "filtered by the sky."

Level 3: School science accounts. Level 3 accounts can be characterized as the re-telling of stories about water that are learned in school. Students recognize that water and substances in water are parts of connected systems, and their accounts include processes that move water and substances through systems. However, there are gaps in students' reasoning, suggesting that students' narratives are not connected into complete models that they use to explain and predict. Level 3 students do not consistently use principles to constrain processes. While they recognize that water and substances can exist at atomic-molecular scales, Level 3 students mostly identify processes (e.g., evaporation) without describing what happens to atoms and molecules.

Moving water: Partially connected systems. At Level 3, students are beginning to trace water through connected systems. However, the nature of the connections among systems is not always clear to students. Hidden connections are most problematic. For example, one Level 3 response to the Bathtub item is, "I think yes because of the fact where else would we get our water from? I know this because after it goes back into the water system it gets cleaned and then it goes to our wells and gets used in our bathtubs." This student left out essential steps in moving water from puddles into the engineered water system. An example of a Level 3 response to the Water in River item is, "through streams, tributaries, and run off." This response suggests that the student is tracing water along multiple pathways along the surface, but is not considering possible underground pathways to the river.

Substances in water: Substances mixed with water. Students at Level 3 understand water quality in terms of identified substances mixed with water. They sometimes use common chemical names (e.g., identifying "chlorine" as a possible water treatment). Their accounts conserve matter through changes in water quality, including invisible changes such as salt dissolving in water. They demonstrate awareness of smaller than visible scales (e.g., they use the word "molecule"), but they do not describe structures and processes at the atomic-molecular scale. For example, one student answered the Salt Dissolving item by writing, "The salt molecules spread out in the water." At this level, students' accounts trace water and substances across invisible boundaries, generally using descriptions that do not account for atoms and molecules. For example, one student answered the Salty Rain item, "No, because when water evaporates it only evaporated as water and leaves the salt behind."

Level 4: Qualitative model-based accounts. Level 4 students use scientific model-based accounts to explain and predict. Their predictions use data about particular situations combined with scientific principles to determine the movements of water and substances in water. Students who use scientific model-based thinking can trace water and substances in water along multiple pathways through connected systems. Furthermore, students at Level 4 can connect phenomena that happen at the macroscopic scale to landscape and atomic-molecular scales.

Moving water: Connected systems. At Level 4, students trace water through connected natural and engineered systems along multiple pathways. For example, Level 4 responses to the Puddle item trace water along multiple pathways. "Runoff into drainage system or seeped into groundwater supply or evaporated into air or combination of all of these." Level 4 responses to the Bathtub item show more detailed connections between natural and human-engineered systems. "Yes: As the water returns to groundwater, it flows into an aquifer. This aquifer could possibly be the one tapped for city water. The water would be purified and delivered via pipes to my house." Furthermore, Level 4 responses apply principles to constrain processes at the landscape scale. For example, one Level 4 response to the Water in River item noted that water could get into a river through the aquifer by following the downhill underground flow and an impermeable layer underground. This response identified how topography and permeability constrain the flow of water in aquifers.

Substances in water: Identified substances mixed with water at multiple scales. Students at Level 4 consistently provide chemical identities for substances and consider relative amounts of substances in reasoning about water quality. Furthermore, identified chemical substances are connected to an understanding of structure at the atomic-molecular scale. For example, one student answered the Salt Dissolving item by writing, "When salt is dissolved into water the salt breaks up into its ions of NA^+ [sic] and CL^- [sic]." In the assessment data, there were some responses that reached Level 4 with respect to simple substances (e.g., salt). However, there were very few responses at Level 4 with respect to more complex substances (e.g., sewage). In addition, few students provided Level 4 accounts by tracing substances mixed with water across system boundaries (especially invisible boundaries) with atomic-molecular scale descriptions.

What is progressing? In our view, growth along a learning progression framework represents movement towards mastering a secondary Discourse. Students' primary Discourses include characteristics of force-dynamic reasoning. As students develop the model-based reasoning of the secondary scientific Discourse, force-dynamic reasoning does not disappear. Students at lower levels of achievement have only their primary Discourse to frame the way they view the world and participate in communities. As students gain mastery over secondary Discourses, they have more tools to use to account for their experiences and make sense of the world. The practices they engage in depend on the Discourses of the communities in which they are participants. Thus students may be capable of providing a model-based account of water in environmental systems, but they may provide force-dynamic accounts if they judge that is what their listeners or readers expect. In fact, force-dynamic accounts can often be sufficient for explaining phenomena. It is not always necessary to explain evaporation in terms of molecules and energy if one just needs to communicate that the puddle in the field is no longer there and the team can now play soccer ("The field dried up; let's go play"). Stating that the puddle is gone is all that is necessary in this situation. However, if one is participating in a community that is trying to figure out why the soccer field is always soggy (e.g., it was built in a place where the water table is close to the surface), one needs to use a model-based account of a scientific secondary Discourse. Students who control secondary Discourses can participate in more communities. Without access to the Discourses necessary for environmental science literacy, students cannot become active participants in evidence-based discussions about local and global environmental issues.

Remaining issues. The perspective of Discourses, practices, and knowledge has been productive in helping us describe and track what progresses in learning progressions that must account for a wide range of students' changing knowledge and practices. It has helped us to organize our data in ways that have allowed us to see important patterns in students' reasoning. However, there are still some challenges that we need to address.

One difficulty is describing the nature of Level 3. We are still trying to determine if students at Level 3 are developing beginning model-based reasoning or if their accounts are the result of layering on more details to their primary Discourse view of the world. This challenge is complicated because the process of developing a new Discourse is a process of adding a secondary Discourse and not replacing the primary Discourse. Thus characteristics of both primary and secondary Discourses are often present in Level 3 accounts. For example, a Level 3 student, when asked to explain how water gets into a river, responded:

> Water gets into a river by a cycle called the water cycle. First, clouds fill up with water droplets and rain onto mountains. The water on the mountains builds up and slides down the mountains into the river. Some of this water evaporates and becomes more clouds.

This student seems to be tracing water from the atmosphere to the surface water system and back. However, the description includes force-dynamic elements, such as clouds filling up with water. Is this student developing a model-based view of the world, or is this student just incorporating school-based narratives about how water cycles into their force-dynamic views of the world? Plans to conduct more clinical interviews that probe students' responses to assessment items may help us tease apart these details.

Another difficulty is writing assessment items that can be answered by students who are at different levels of achievement. We have had to learn how to write assessment prompts that can elicit responses across a range of Discourses. We have found that students who can use model-based reasoning may provide force-dynamic responses to assessment items if the item prompt does not specifically request a model-based response. For example, students who can use a model-based account to describe what happens when salt is mixed with water may not do so unless specifically requested to include descriptions of atoms and molecules in their answer. However, adding these clues to the prompts sometimes makes the prompt seem too difficult to students who have not developed a model-based view of the world. Sometimes these students do not provide any response to the item, even though a force-dynamic response would have been possible. We continue to explore ways to write assessment prompts that can be productive for both force-dynamic and model-based reasoners (Jin & Anderson, this volume).

The relationships among Discourses, practices, and knowledge are useful in helping us meet the challenges that our choices for our learning progressions have introduced. We will continue to leverage the benefits this perspective provides and address the limitations that it presents as we move forward. In the next section, we describe another challenge that our goals for developing interconnected learning progressions for environmental science literacy have presented.

CHALLENGE TWO: DEFINING PATHWAYS AND LINKING TO INSTRUCTION

The research groups of the Environmental Literacy Project are at different stages in the learning progression design process. The water and carbon research group have each developed two critical design products—a learning progression framework and associated assessments. In addition, the carbon cycle research group has developed a third design product aimed at linking its learning progression framework to instruction. This third design product is a set of instructional resources that allows teachers to use learning progression frameworks in their classrooms and allows researchers to investigate how students learn the practices of environmental science literacy.

We recognize that progress through a learning progression is not developmentally inevitable. Instruction influences progress, and students may take more than one path through learning progression framework levels, depending on the instruction they receive. While the majority of students we assessed in the past five years revealed similar types of reasoning, which were characteristic of Levels 1–3 in our learning progression frameworks (as described previously for the water learning progression framework), the carbon research group also collected evidence that suggested some students exhibit notable differences in reasoning compared to their peers at the same grade level (Chen, Anderson, & Jin, 2009; Jin & Anderson, this volume; Jin, Zhan, & Anderson, 2009; Mohan, Chen, & Anderson, 2009). We interpreted these differences as indications of alternative pathways in the learning progression. Our project uses the term "pathway" to describe the paths learners may take between the Lower and Upper Anchors. While pathways share anchor points, the intermediate levels vary, which makes the pathways distinguishable. The variation in these intermediate levels provides an opportunity to explore the role instruction plays in the learning progression.

The teaching experiments conducted by the carbon cycle research group show how we approached the challenge of identifying and defining multiple pathways and the challenge of defining the link between instruction and the learning progression. In this section we provide a brief overview of the carbon cycle learning progression framework. Thereafter we explain how we identified and defined alternative pathways in the learning progression, the approach we took in linking these pathways to instruction, and the limitations of our approach.

OVERVIEW OF THE CARBON CYCLE LEARNING PROGRESSION FRAMEWORK

The Upper Anchor of the carbon cycle learning progression framework identifies three groups of carbon-transforming processes that are necessary for mastering scientific Discourse. These are the processes that *generate* organic carbon through photosynthesis, the processes that *transform* organic carbon through biosynthesis and digestion, and the processes that *oxidize* organic carbon through cellular respiration and combustion. We chose to organize the Upper Anchor around these processes because human and other living systems use them to acquire energy, and environmental systems use them to regulate levels of atmospheric CO_2; thus an understanding of these processes is central to environmental science literacy. This grouping highlights important similarities and differences in how the three processes alter the flow of matter and energy at different scales.

Table 2 shows that progress from the Lower Anchor to the Upper Anchor requires substantial reorganization of knowledge about these processes. The middle row—Macroscopic events—is accessible by individuals using both Lower and Upper Anchor Discourses; thus we can use these events to examine different Discourses (see Jin & Anderson, this volume). The bottom row—Lower Anchor—shows how an individual using primary Discourse might organize and account for macroscopic events, while the top two rows—Upper Anchor—show how an individual who has mastered scientific Discourse accounts for the same set of events (i.e., the Carbon-transforming process row shows patterns in chemical reactions, while the Scientific accounts row shows specific chemical processes learned in school).

Table 2. Contrasting Ways of Grouping Carbon-Transforming Processes.

<table>
<tr><td rowspan="2">Upper Anchor</td><td>Carbon-transforming process</td><td>Generating organic carbon</td><td colspan="3">Transforming organic carbon</td><td colspan="3">Oxidizing organic carbon</td></tr>
<tr><td>Scientific accounts</td><td>Photosynthesis</td><td>Biosynthesis</td><td>Digestion</td><td>Biosynthesis</td><td colspan="2">Cellular respiration</td><td>Combustion</td></tr>
<tr><td colspan="2">Macroscopic events</td><td colspan="2">Plant growth</td><td colspan="2">Animal growth</td><td>Breathing
Exercise
Weight loss</td><td>Decay</td><td>Burning</td></tr>
<tr><td colspan="2">Lower Anchor: Informal accounts</td><td colspan="5">Natural processes in plants and animals, enabled by food, water, sunlight, air, and/or other things</td><td>Natural process in dead things</td><td>Flame consuming fuel</td></tr>
</table>

Note. From "Developing a Multi-Year Learning Progression for Carbon Cycling in Socio-Ecological systems," by L. Mohan, J. Chen, and C. W. Anderson, 2009, Journal for Research in Science Teaching, 46, p. 684. Copyright 2009 by Wiley Periodicals, Inc. Reproduced with permission of Wiley Periodicals, Inc. via Copyright Clearance Center.

Students at the Lower Anchor view macroscopic events as characteristic of organisms and objects. These students organize their world based on actors—plants, animals, and objects—as opposed to processes. They pay particular attention to the different needs and abilities of actors and to the outcomes of events that involve actors struggling to fulfill their natural tendencies. Dead things, which have lost their capacity as actors (students often say dead things "have no energy"), are prone to decay.

When individuals at the Upper Anchor observe the same macroscopic events, they can provide scientific accounts that reflect organization based on scientific principles. Mastering scientific Discourse includes recognizing that every process obeys the following principles:

1. hierarchy of systems and scale (i.e., the world is organized into dynamic systems that have structures and processes that occur at multiple scales);
2. conservation and cycling of matter (i.e., laws of conservation of mass and atoms); and
3. conservation and degradation of energy (i.e., energy, like matter, cannot be created or destroyed; however, unlike matter, energy cannot be recycled).

An individual at the Upper Anchor uses scientific principles—especially matter and energy principles—to construct explanations and to organize processes in the world.

The carbon cycle learning progression framework uses both processes (i.e., generation, transformation, and oxidation) and principles (i.e., scale, matter, and energy) as key dimensions. These dimensions have recognizable face validity in the science and science education communities. We used processes and principles to opearationalize the knowledge and practice components of our carbon learning progression framework. These dimensions guided the development of assessments and the analyses of data. Both dimensions became especially important as we began to explore pathways in the learning progression and the relationship between pathways and instruction.

LIMITATIONS OF THE CARBON CYCLE FRAMEWORK

Before the 2008–2009 academic year we developed a learning progression framework that describes how student reasoning changes, or evolves, without special instructional interventions from researchers. As with the water learning progression framework, the carbon learning progression framework was developed using an iterative approach where learning progression framework development and empirical data from assessments informed each other. What emerged from several years of work was a learning progression framework that documented consistent patterns among student responses across different settings (Mohan, Chen, & Anderson, 2009). The initial design products of this research included a learning progression framework and assessments.

Processes and principles were central to the initial stages of our work. Both dimensions helped us define the knowledge and practice necessary for reasoning about carbon cycling. Furthermore, both dimensions were useful for designing assessments. For example, we designed assessment items in order to elicit students' accounts about at least one process and at least one principle (e.g., the item, "where does the mass of a tree come from?" targets the process of photosynthesis and the principle of conservation of matter; "where does gasoline go when a car's fuel tank is empty?" targets the process of combustion and the principle of conservation of matter). For this reason, our initial work used processes and principles as progress variables in the carbon cycle learning progression framework. We used the macroscopic events from Table 2 to identify types of accounts. Most of our questions elicited students' accounts of individual macroscopic processes (e.g., plant growth), while a few questions focused on comparisons among processes or connections between processes (e.g., how decomposition connects to plant growth). We used principles to identify elements of accounts. A complete account of any process describes changes in matter and/or energy at different scales.

As we continued our development and validation work, however, we saw two limitations to the learning progression framework. The first limitation was a conceptual problem: although using the processes as progress variables to describe types of accounts was useful as a data analysis strategy, using principles proved problematic because these principles are not easily differentiated for many students. The second limitation concerned evidence of failure in our educational system: few students achieved Level 4 reasoning, mostly because their accounts did not consistently conserve both matter and energy.

Conceptual limitations: Matter and energy as progress variables. As we began to look more closely at alternative pathways, we initially hypothesized that alternative pathways in the learning progression would relate to different progress on processes and/or principles. We hypothesized that students may reason at higher levels about particular processes (e.g. photosynthesis) or particular principles (e.g., conservation of matter) compared to other processes and principles. If true, our learning progression framework would have to account for these differences. In fact, we would have to be especially attentive to these differences when designing instructional materials. For example, if students grasped matter principles more readily than energy principles, we would use this information to inform our instructional interventions.

In order to test these hypotheses, we designed assessments to elicit responses about both processes and principles. Our goal was to explore whether students tended to reason at higher levels about particular processes or principles. The assessments were comprised of open-response items about the five macroscopic events identified in Table 2. We used 29 assessment items that asked students to account for what happens to matter and/or energy during these five events. We scored student performance on each item. Although we used 29 items, we assigned 45 scores for each student since some items were scored for more than one process or principle. Of the 29 items, 25 were scored for matter and 20 were scored for energy. It is important to point out that while some items targeted either matter or energy, student responses often addressed both. Therefore coders scored both principles. For example, when students were asked to explain what happens to matter during weight loss, many students used energy in their explanations, prompting coders to score for both matter principles and energy principles.

After scoring individual items we conducted Multidimensional Item Response Theory (IRT) analyses in order to obtain person ability estimates, each representing the average performance of a single student on all items related to a process or principle. We examined whether performance on items about one type of process or principle correlated with performance on items about another process or principle (Mohan, Chen, Baek, Anderson, Choi, & Lee, 2009). To conduct our analyses, we used a sample of assessments from 771 students in 18 classrooms, grades 4–12. We found correlations between the processes were generally moderate to high (.542 or greater). Cellular respiration and growth/biosynthesis events appeared slightly more difficult for students than other processes. Yet, in general, students exhibited consistent levels of reasoning about the processes.

However, we encountered two significant difficulties when we examined principles—matter and energy—as progress variables in our coding and analyses. The first difficulty was conceptual: What do matter and energy mean to students who reason at Level 1 and Level 2? When we coded Level 3 and Level 4 accounts, we could generally identify elements that corresponded to the scientific concepts of matter and could distinguish those from scientific concepts of energy. As we uncovered the force-dynamic reasoning of Level 1 and Level 2 accounts, it became increasingly problematic to identify "matter" and "energy" in these accounts. In explaining the requirements for plant growth, for example, Level 1 and Level 2 students did not distinguish between forms of energy (sunlight), forms of matter (air, water, soil), or conditions, (warmth, care). Although Level 1 and Level 2 students frequently used the word "energy," sometimes they used it to identify powers or abilities of actors (e.g., the girl can run because she has energy) and sometimes they used it to refer to generalized needs or enablers (e.g., water, air, sunlight, and soil all supply plants with energy in different ways).

Our search for developmental precursors to scientific concepts of matter and energy proved intellectually fruitful. We could track connections between younger students' ideas about enablers and results of macroscopic events and older students' ideas about matter and energy (Mohan, Chen, & Anderson, 2009). Similarly, we saw connections between younger students' ideas about cause and action and older students' ideas about energy sources and transformations of energy (Jin & Anderson, 2008). These connections, however, did not really solve our underlying conceptual difficulty. The intellectual precursors to scientific concepts of matter and energy were like tributaries to a stream: There were many of them. It did not really make sense to privilege some over others by labeling them as "matter" and "energy" elements in the accounts of students who really were not thinking about matter and energy.

A second limitation with matter and energy as distinct progress variables emerged from our data analyses. We found the correlation between matter and energy dimensions was high (0.959), indicating students had very similar scores for both matter and energy. This finding reflects both the conceptual difficulty in separating the two principles and the limitation these two principles place on scoring. While not supporting our original hypothesis that students may understand one principle before the other, the results made sense given the characteristics of student accounts, especially at the lower levels. In our earlier studies, students seemed to use energy as a way to account for mass changes attributable to gases. Thus students' developing knowledge about matter—especially gases—and their developing knowledge about energy were deeply intertwined. Trying to separate and code the two principles forced a distinction that most students did not make.

We retained matter and energy as progress variables because both are distinguishable at the Upper Anchor. Moreover, it is likely that if students are given targeted instruction on these principles they may demonstrate pathways with different understandings of these two principles. However, our initial use of matter and energy progress variables proved fruitless given the understanding of students who have experienced traditional, or status quo, instruction.

Practical limitations: Evidence of failure in our educational system. Our definition of the Upper Anchor (Level 4) of the carbon cycle learning progression framework is not overly ambitious. The ideas in the Upper Anchor are included in current national standards (National Research Council, 1996) and in the standards of many states, including Michigan, where we collected much of our data (www.michigan.gov/mde). Our findings showed, however, that few students met these standards: Mohan, Chen, and Anderson (2009) found that only 10% of the high school students in our sample provided Level 4 accounts of processes and principles. These students received similar instruction compared to their peers—

instruction that mainly focused on delivering detail-oriented science information to students.

A close examination of our data showed that the main explanation of students' difficulties with achieving Level 4 reasoning was their struggle to understand and apply matter and energy principles. For example, Level 3 students had difficulty connecting macroscopic events with atomic-molecular models. Their accounts of processes with gaseous reactants or products often converted matter to energy or vice versa.

So these findings posed a dilemma. Matter and energy principles did not work very well for us as progress variables. At the same time, however, students' main obstacle in achieving Level 4 reasoning was their failure to understand and apply those same principles. We also knew that the principles should play a critical role in developing our third design product—instructional resources—especially given the central role of principles in Upper Anchor reasoning. We organized our research in the 2008–2009 academic year to address this dilemma.

ALTERNATIVE PATHWAYS

As we struggled with this dilemma, we identified a potential solution based on research comparing Chinese and American students' accounts (Jin & Anderson, this volume). The evidence that emerged from our analyses suggested we were missing an important dimension in our learning progression framework. Jin, et al. (2009) and Chen et al. (2009) administered written assessments and conducted interviews with middle and high school Chinese students. Their analyses showed that Chinese students could more readily use technical language that accurately identified appropriate processes. Yet when Chinese students were asked to elaborate on their answers, they struggled to construct explanations that followed scientific principles. In our reconsideration of the data from American students we recognized that many of them had the ability to use scientific "names" for systems and processes that exceeded their ability to construct an explanation using scientific principles.

With this insight, we reexamined data from both Chinese and American students in terms of students' ability to provide "names" for systems and processes and their ability to use scientific principles in their explanations. We labeled these dimensions "naming" and "explaining." We used the "naming" dimension to explore students' use of specific key words and phrases characteristic of particular levels of reasoning. We used the "explaining" dimension to examine the structure of students' explanations and their grounding in terms of scientific principles (Jin & Anderson, this volume). In this way, scientific principles remained a centerpiece of our new explaining dimension.

Our re-examination of the data showed that the majority of students demonstrated levels of naming that exceeded their levels of explaining (e.g., Jin et al., 2009). This observation made sense given most students in our sample were receiving traditional science instruction, or what we refer to as status quo instruction. Such instruction focuses on communicating to students the technical language of science (e.g., Lemke, 1990) and on building detailed narratives about specific processes. While these narratives are constrained by scientific principles, students often focus on the details of the narrative rather than on the more general principles. For example, students can memorize chemical equations such as, $C_6H_{12}O_6 + 6O_2 \rightarrow 6H_2O + 6CO_2$ without recognizing that "balancing the equation" is a way of applying conservation of matter as a constraining principle—the process of cellular respiration does not create or destroy atoms. Similarly, students often fail to connect accounts of processes across scales. Students also learn narratives about principles as shown by their recitation of conservation laws; however, the principles are largely invisible to students. They make no connections between narratives of processes and principles. For example, students may be able to describe conservation laws, but they cannot use them as reasoning tools in different contexts.

Our re-examination of the data also revealed that some students had similar levels of naming and explaining, while for others levels of explaining exceeded levels of naming. For example, some students showed a strong commitment to principles, such as conservation of matter or energy, without knowing the technical language and technical details of a chemical process. This additional pattern was an indication of the possibility of an alternative pathway in our learning progression framework. For this reason, naming and explaining dimensions were particularly useful for distinguishing between the pathways students take through the levels of the learning progression framework. We labeled these pathways "structure-first," which focused on naming, and "principles-first," which focused on explaining with principles. Figure 4 depicts these pathways with their shared Lower and Upper Anchor points.

Structure-first pathway and naming. The current carbon cycle learning progression framework described by Mohan, Chen, and Anderson (2009) and the water cycle learning progression framework described earlier in this chapter are largely descriptions of the "structure-first" pathway (the solid line in Figure 4). Students on this pathway acquire new scientific words and phrases but use them in explanations that retain significant force-dynamic characteristics (e.g., students may identify "photosynthesis" as a key process in plants but cannot explain how it changes matter or energy). Students may be able to recite conservation laws but cannot use these laws to explain what happens to matter or energy during weight loss, combustion, or other carbon-transforming processes. Thus we expect students who follow this pathway to have higher levels on the Naming progress variable than on the Explaining progress variable.

We interpret our data and other research on classroom teaching (e.g., TIMSS Video Study, Roth et al., 1999) to mean (a) that the structure-first pathway is currently the norm in American classrooms and (b) that progress to the Upper Anchor through the structure-first pathway is achieved by only a small percentage of students. The transcript below illustrates the structure-first pathway. This high school student, Dan, showed Level 3 in terms of naming and Level 2 in terms of explaining.

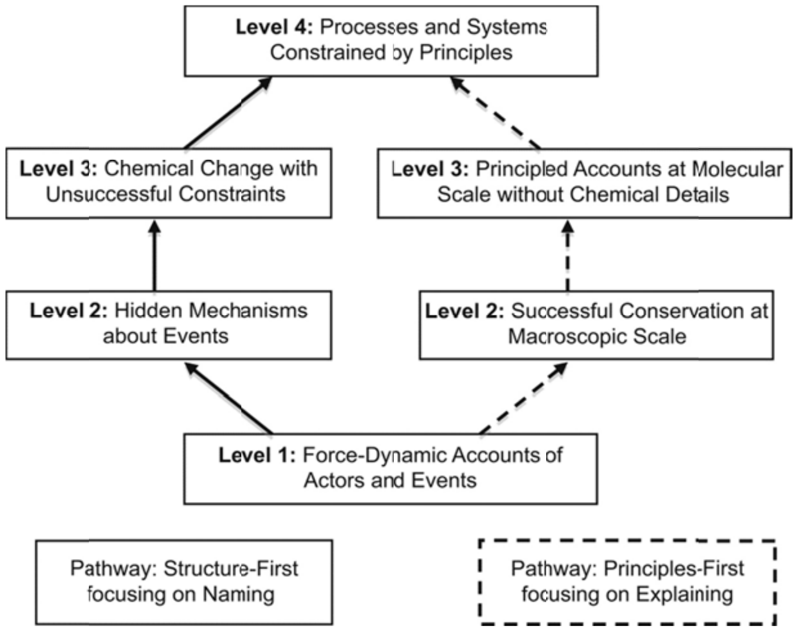


Figure 4. Two different pathways between the Lower Anchor (Level 1) and the Upper Anchor (Level 4).

Example 1: Structure-First Pathway

Interviewer: How does sunlight help photosynthesis?

Dan: The, well like the vitamins and stuff in it, like that's what it uses.

Interviewer: And you also talked about a food...What do you mean by food?

Dan: Like glucose that the tree uses to grow.

Interviewer: Okay. So where does glucose come from?

Dan: The tree makes it from all the different things that it uses.

Interviewer: Could you talk a little bit about what are the different things?

Dan: Like air, vitamins, the soil, nutrients, sun and water.

... ...

Dan: Well, yeah I think that uses like all the same…after it makes its food it uses the glucose for energy.

Interviewer: Glucose is a type of energy?

Dan: Yep.

... ...

Interviewer: Okay. Now, you know, the tree, when the tree grows it becomes heavier, right? It will put on more weight. So where does the mass come from?

Dan: It comes from the, all like glucose that it makes, it like keeps building on and building on until it gets as big as it is.

Interviewer: So what are the energy sources for the tree?

Dan: Well, the same as photosynthesis- vitamins, water, air, light, yeah.

Level 3 reasoning in the carbon cycle learning progression framework involves the incorporation of chemical processes into a student's account. Dan is able to provide scientific names for a chemical process (photosynthesis) and a chemical identity for an important material (glucose), which indicates he has acquired "names" consistent with Level 3 reasoning. He also understands that plants make glucose from other components and that glucose contributes to the increase in mass. Yet Dan cannot differentiate between key materials and energy resources in terms of scientific principles. In addition to light, he also lists vitamins, water, and air as energy sources for photosynthesis.

As in the water learning progression framework, Level 2 reasoning about the carbon cycle is characterized by force-dynamic accounts including actors ("The tree makes it from all the different things that it uses.") and enablers ("vitamins, water, air, light"). While Level 2 students understand that actors accomplish their purposes through hidden mechanisms ("photosynthesis"), they lump enablers into one group. To Dan, materials and "light" are lumped into a group of enablers required by the tree for growth. While he has acquired "glucose" as a new descriptive term, either he is confused about whether it is a form of matter or a form of energy, or he sees no need to differentiate between the two.

Principles-first pathway and explaining. Figure 4 also shows a "principles-first" pathway focused on explaining (the dashed line). This pathway describes students who offer explanations using scientific principles even in instances when they do not have the detailed chemical knowledge and language to provide a full description. While we have examples of students who demonstrate explaining that is aligned with or exceeds their naming, this pattern is rare in our data. The transcript below shows an example of a middle school student, Ryan, who exhibits a pattern in which explaining is aligned with or exceeds naming. This student shows Level 3 on explaining and Levels 2 or 3 on naming.

Example 2: Principles-First Pathway

> **Interviewer**: You said sunlight, can you tell me a little bit about sunlight, how does it supply the tree with energy, do you know how it happens?
>
> **Ryan**: It comes in, obviously as a form of light energy, and that being a form of energy, it then converts through photosynthesis, it converts that to a form of energy that the tree can use.
>
> **Interviewer**: What form of energy is that?
>
> **Ryan**: Either kinetic or stored, I am not sure, probably more stored.
>
> **Interviewer**: Keep going.
>
> **Ryan**: And it would use kinetic for whatever growing it does at the moment, but it would probably use more stored energy to store it away for another time to use.
>
> **Interviewer**: Where does the tree store its energy?
>
> **Ryan**: It stores it mostly in the trunk, since that's the largest area, but in all of the branches of it, in the form of starch.
>
> **Interviewer**: Do you think energy is stored in molecules?
>
> **Ryan**: No.
>
> **Interviewer**: You mentioned a form of starch, do you think starch is a molecule and do you think energy is stored in that?
>
> **Ryan**: It is. I am not sure how it's stored in it. It might be with the molecule's vibrations or something. I am not positive.

Ryan has developed a story about energy transformations in plants that recognizes different forms of energy. Although he admits he does not know how starch stores energy, he does not default to the matter-energy conversions often observed among Level 2 and Level 3 students on the structure-first pathway. Ryan shows a commitment to conservation of energy without fully understanding the chemical nature of molecules; he does not use scientific terms that exceed the explanation provided.

As described earlier in the chapter, our goals for environmental science literacy include the conviction that model-based reasoning is necessary if students are to master scientific Discourse and to participate as environmentally literate citizens. While the principles-first pathway appears to be the exception to the rule, it is our belief that this pathway has potential for supporting students in acquiring model-based reasoning. For this reason we use our learning progression framework, especially the principles-first pathway, to help design our instructional resources.

DESIGNING INSTRUCTIONAL INTERVENTIONS

Our approach to learning progression work and instructional interventions is notably different from other learning progression projects. Some learning progression researchers focus on defining a clear link between instruction and learning progression framework early in their design process and then develop instructional materials that have a very specific and carefully planned instructional sequence that is closely linked to progress from one level to the next (e.g., Schauble, 2009; Wiser & Smith, 2008; Wiser, Smith, Asbell-Clarke, & Doubler, 2009). These projects document what is possible for students given a specific instructional context and what students are capable of doing in those environments. Learning progression frameworks developed within this perspective tend to focus on the boundaries of what could be, given the right set of curricula and support within a given context.

In contrast, we began our work by documenting what is happening in the instructional context of our schools today. We focused on developing learning progression frameworks and assessments that could capture the current reality of schools. We designed assessments that could be used to elicit responses from students of diverse ages, cultures, and social backgrounds (Jin & Anderson, this volume). We needed a learning progression framework and operational system for handling that diversity. We also devoted time to refining our learning progression framework based on what we had learned from those assessments.

Our data suggested that status quo teaching leads many students to achieve Level 3 reasoning on the structure-first pathway in which naming exceeds explaining. However, we thought that an alternative to this instructional approach—one that emphasized principle-based reasoning—would support students on the principles-first pathway and that this pathway would be more successful in helping students reach the Upper Anchor. We recognized the link between instruction and the learning progression framework was not tied to a specific instructional sequence but rather reflected a teacher's general approach to conveying the importance of principle-based reasoning in a variety of contexts. Rather than using our teaching experiments to test the effectiveness of a sequenced set of activities, we chose to design a learning progression system and instructional intervention based on the following goals:

- We wanted to help teachers recognize that scientific Discourse involves careful attention to principle-based explanations and to offer suggestions for how to make these principles more visible to students.
- We wanted to make conservative changes to instruction that would improve student performance without making whole scale changes in curricula.

- We wanted the instructional interventions to span a large age range and to be useful to teachers and students in a variety of settings.
- We wanted the instructional interventions to be flexible so that teachers could use our resources within the curriculum adopted by their district.

Given these goals, we focused on designing Tools for Reasoning that were closely linked to the learning progression framework. These tools had to capture important aspects of different processes, follow scientific principles, and ultimately help students construct explanations instead of focusing on names and school science narratives unconstrained by scientific principles. When working with a broad age span of students from diverse settings, this also meant generating tools that had continuity across the students' ages and that could be used within different instructional and social contexts.

The matter and energy Process Tool is an example of a principle-based tool used in our instructional interventions (see Figures 5 and 6). It is designed to scaffold construction of scientific accounts of carbon-transforming processes. For students who have mastered scientific Discourse, the Process Tool can be used to trace matter and energy inputs and outputs. The accounts for the same process by students who have not mastered scientific Discourse will be very different. For example, Figure 5 shows a comparison account for both primary and scientific Discourses using the Process Tool. Students who reason with their primary Discourse describe needs or enablers (which may include materials, forms of energy, or conditions) that actors must have to accomplish their purposes. The results are usually not in material form; matter is simply allowed to appear or disappear without accounting for its conservation. Students who use their primary Discourse describe the end purpose or results accomplished by actors when they obtain the enablers they need. In contrast, students who use scientific Discourse distinguish inputs in terms of matter and energy for particular processes; the results of events are matter and energy outputs. Thus, in the scientific Discourse, there is a storyline about how matter and energy transform during a particular process.

For classroom use, we designed the Process Tool to help organize students' accounts around the structure of scientific Discourse as shown in Figure 5. In using the Process Tool, students must choose from a given set of matter and energy inputs and outputs. Students are asked to identify the materials entering the system. Students are also asked to identify the energy entering the system. The students use labels to represent these matter and energy inputs. As with inputs, students must choose from the same set of labels in order to identify matter and energy outputs. Figure 6 shows an example of the Process Tool for plant growth, with a set of labels that students choose from. The matter labels in Figure 6 provide space for students to identify specific materials.

The Process Tool can be used to describe macroscopic events (e.g., match burning, plant growing, etc.), landscape-scale processes (e.g., primary production, food chains), and atomic-molecular scale chemical processes (e.g., combustion, photosynthesis). In elementary school the Process Tool can help students begin tracing matter and energy through systems (focusing particularly on distinguishing between different types of enablers and becoming more aware of gases as a form of matter). We believe middle school students can learn to use atomic molecular models to explain transformations of matter and energy with the Process Tool, although without much chemical detail. High school students can master additional chemical details.

The Process Tool was designed to support students in using the laws of conservation of matter and conservation of energy to reason about events or processes. Students are given a limited number of forms of matter—solids, liquids, and gases—that they can use to label either material kinds (e.g., food) or chemical identities (e.g., glucose: $C_6H_{12}O_6$). The Process Tool also uses a limited number of energy forms—light, motion, chemical energy, electrical energy, and heat. Students trace energy transformations between these forms to practice forming accounts that conserve energy. We also wanted to provide teachers with an opportunity to highlight the principle of energy degradation. For example, "heat" uses slightly different labeling to indicate that it is a form of energy no longer usable by organisms or objects.

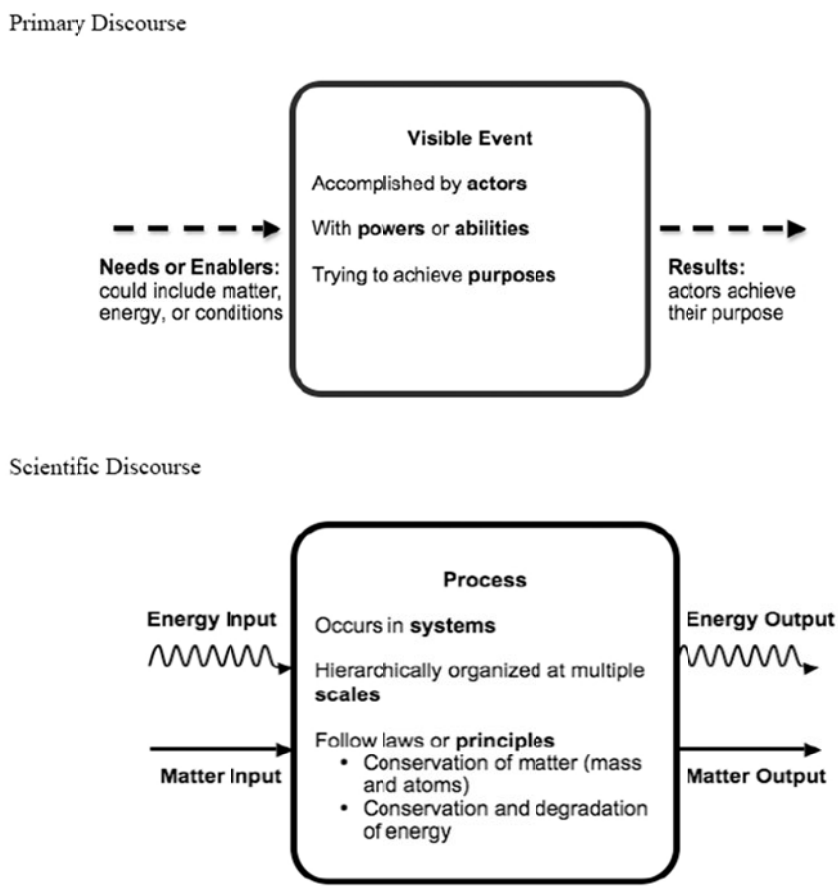


Figure 5. Comparison between the structure and content of the Process Tool for primary versus scientific Discourse.

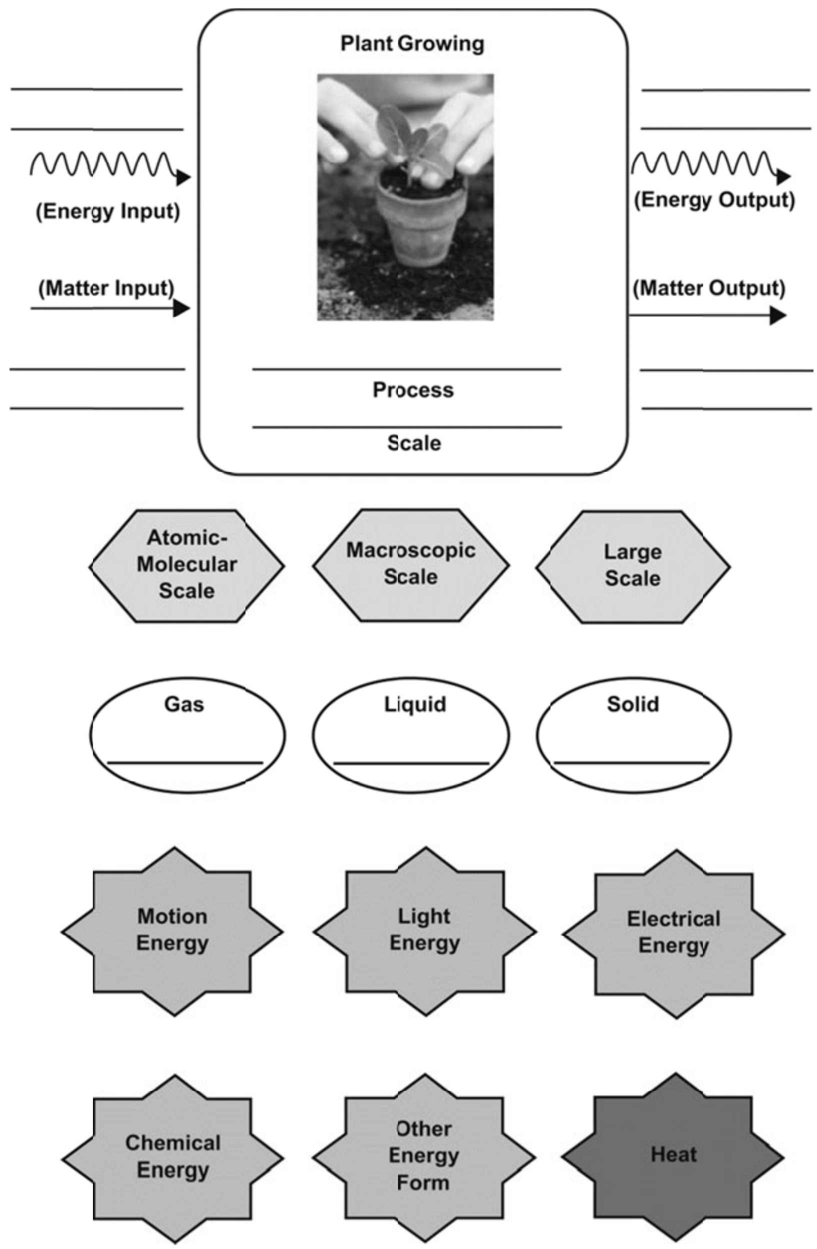


Figure 6. Process Tool for plant growth.

The design of the Process Tool allows students to construct accounts of processes at different scales and to discuss how the labeling of matter and energy inputs and outputs changes as a result of moving up and down scales (e.g., food at the macroscopic scale may become carbohydrates, lipids, and proteins at the atomic-molecular scale). The Process Tool is used in the classroom in three forms: as a 3x4 poster with Velcro or magnetic tabs for matter and energy labels, in student activity pages, and in PowerPoint presentations the teacher uses in whole group instruction.

LIMITATIONS TO OUR LINKS TO INSTRUCTION

The naming and explaining dimensions have been particularly helpful in distinguishing between a structure-first pathway and a principles-first pathway. These pathways share the same anchor points, but transitional levels vary in fundamental ways. The structure-first pathway describes transitional levels for individuals whose ability to name systems and processes exceeds their ability to explain. The principles-first pathway describes transitional levels in which naming and explaining are aligned, or explaining exceeds naming. Differences between these pathways have informed our design of instructional resources, which relies on the principles-first pathway—a pathway that we believe will help students acquire the model-based reasoning necessary for environmental literacy.

We are currently analyzing data from our pilot teaching experiments to explore whether the use of Tools for Reasoning, such as the Matter and Energy Process Tool, appeared to influence student use of principle-based reasoning in their explanations. While we hope to find evidence of improved student learning, we are aware of important limitations in the materials that we are currently testing.

First, while our approach to instructional interventions includes what we refer to as “conservative” changes to instruction, we argue that these changes represent substantial shifts in pedagogy—shifts that place more responsibility on classroom teachers. While we provide teachers with some lesson plans and materials, our approach relies primarily on a set of tools—a learning progression framework, assessments, and Tools for Reasoning—with the expectation that teachers will determine how best to use these tools in their classroom. We still know little about the extent of professional development required to support teachers as active users of the type of learning progression system proposed by our work.

Similarly, for our instructional interventions to achieve real change for students, teachers must also make decisions about when and how to integrate Tools for Reasoning and other instructional resources into their existing curricula. Yet we still know little about the extent to which these tools must be integrated to achieve observable changes in student performance. Formative assessments would help teachers track student progress and would help them make immediate instructional decisions using the learning progression frameworks to inform these decisions. We feel that formative assessments we have developed thus far are limited and perhaps inadequate. In our proposed research we plan to make formative assessments more central to our learning progression system.

CONCLUSIONS

In this chapter we have described our approaches to two core challenges that we face in defining learning progressions leading toward environmental science literacy: defining what progresses in a learning progression and defining alternate pathways linked to instruction. In addressing these challenges, we have developed a coordinated learning progression system that includes a learning progression framework, sets of validated assessments in several domains important to environmental literacy, and tools and instructional resources that can be used flexibly in the classroom.

An important feature of our learning progression frameworks, assessments, and tools is our focus on language and language use. We have grounded our learning progression frameworks in a Discourse perspective that focuses on how language both shapes and represents student reasoning. Language, which shapes the way students view the world, provides a clue to understanding how students reason about phenomena. This focus on language as both a shaper and a product of how students view the world has allowed us to develop learning progression frameworks that account for the sociocultural as well as cognitive aspects of learning across a broad range of students and across broad scientific domains necessary for environmental science literacy.

Furthermore, the focus on student language and practices helps us understand the pathways that students take through the learning progression from their primary Discourse to a secondary Discourse of scientific model-based reasoning. For the carbon cycle learning progression framework we recognize the key role that the principles—the hierarchy of systems at different scales and the conservation of matter and energy—play in scientific reasoning. We suggest on the basis of our research that these principles can be at the core of teaching that helps students take a "principles-first" pathway toward environmental science literacy that will be more effective than status quo teaching. We are currently testing the effectiveness of instructional interventions that support this alternate pathway; we look forward to learning more about their effectiveness.

ACKNOWLEDGEMENTS

This research is supported in part by grants from the National Science Foundation: Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems (NSF 0815993), and Targeted Partnership: Culturally Relevant Ecology, Learning Progressions and Environmental Literacy (NSF-0832173), and CCE: A Learning Progression-based System for Promoting Understanding of Carbon-Transforming Processes (DRL 1020187). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The authors would like to thank several people for their invaluable contributions to the work presented in this chapter. We would like to acknowledge the contributions made by Hui Jin, Jing Chen, Li Zhan, Jonathon Schramm, Kennedy Onyancha, Hamin Baek, Hsin-Yuan Chen, Rebecca Dudek, and Tammy Dionese at

Michigan State University; Karen Draney, Mark Wilson, Yong-Sang Lee, and Jinnie Choi at the University of California, Berkeley; Alan Berkowitz at the Cary Institute; and Ali Whitmer at Georgetown University.

NOTE

[1] We recognize that there is a large body of literature on communities and discourse in education and science education, in particular, and on the differences between students' home communities and school (Heath, 1983; Lee & Fradd, 1998; O'Connor & Michaels, 1993). We acknowledge this literature. In this chapter we refer to specific characteristics of students' language and the relationship between their use of language and how they make sense of the world.

REFERENCES

Anderson, C. W. (2003). *Teaching science for motivation and understanding*. Retrieved from Michigan State University website: https://www.msu.edu/~andya/TEScience/Assets/Files/TSMU.pdf

Anderson, C. W. (2009, April). *Learning progressions for principled accounts of processes in socio-ecological systems.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Arvai, J., Campbell, V. E. A., Baird, A., & Rivers, L. (2004). Teaching students to make better decisions about the environment: Lessons from the decision sciences. *The Journal of Environmental Education, 36*, 33–44.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*, 1–14.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Chen, J., Anderson, C. W., & Jin, X. (2009, April). *American and Chinese secondary students' written accounts of carbon cycling in socio-ecological systems*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Cobb, P., Confrey, J., diSessa, A. A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Cobb, P., & Hodge, L. L. (2002). A relational perspective on issues of cultural diversity and equity as they play out in the mathematics classroom. *Mathematical Thinking and Learning, 4*, 249–284.

Cobb, P., & Hodge, L. L. (2003). Culture, identity and equity in the mathematics classroom. In N. S. Nasir & P. Cobb (Eds.), *Improving access to mathematics: Diversity and equity in the classroom*. New York, NY: Teachers College Press.

Covitt, B. A., Gunckel, K. L., & Anderson, C. W. (2009). Students' developing understanding of water in environmental systems. *Journal of Environmental Education, 40*(3), 37–51.

Covitt, B. A., Tan, E., Tsurusaki, B. K., & Anderson, C. W. (2009, April). *Students' use of scientific knowledge and practices when making decisions in citizens' roles*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA

Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

Draney, K., Wilson, M., Choi, J., & Lee, Y. (2007, July). *Developing progress variables and learning progressions for the carbon cycle*. Paper presented at the Center for Curriculum Materials in Science Knowledge Sharing Institute, Washington, D.C.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, 23(7), 5–12.

Fleming, R. (1986). Adolescent reasoning in socio-scientific issues. Part I: Social cognition. *Journal of Research in Science Teaching, 23*, 667–687.

Gee, J. (1991). What is literacy? In C. Mitchell & K. Weiler (Eds.), *Rewriting literacy: Culture and the discourse of the other* (pp. 3–11). Westport, CT: Bergin & Gavin.

Gore, A. (2006). *An inconvenient truth: The planetary emergency of global warming and what we can do about it.* New York, NY: Rodale Press.

Gunckel, K. L., Covitt, B. A., Dionise, T., Dudek, R., & Anderson, C. W. (2009, April). *Developing and validating a learning progression for students' understanding of water in environmental systems*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. Cambridge, UK: Cambridge University Press.

Intergovernmental Panel on Climate Change. (2007). *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: Author.

Jin, H., & Anderson, C. W. (2008, April). *Developing a long-term learning progression for energy in socio-ecological systems*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Jin, H., Zhan, L., & Anderson, C.W. (2009, April). *A cross-cultural study: Comparing learning progressions for carbon-transforming processes of American and Chinese student.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Kolstø, S. (2001). "To trust or not to trust…" - Pupils' ways of judging information encountered in a socio-scientific issue. *International Journal of Science Education, 23*, 877–901.

Kolstø, S. (2006). Patterns in students' argumentation confronted with a risk-focused socio-scientific issue. *International Journal of Science Education, 28*, 1689–1716.

Lee, O., & Fradd, S. H. (1998). Science for all, including students from non-English-language backgrounds. *Educational Researcher, 27*(4), 12–21.

Lemke, J. (1990). *Talking science: Language, learning and values*. Westport, CT: Ablex.

Long Term Ecological Research Planning Committee. (2007). *Integrative science for society and environment: A strategic research initiative*. Retrieved from Long Term Ecological Research Network website: http://www.lternet.edu/decadalplan/

Metz, K. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22*, 219–290.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

Mohan, L., Chen, J., Baek, H., Anderson, C. W., Choi, J., & Lee, Y.-S. (2009, April). *Validation of a multi-year carbon cycle learning progression: A closer look at progress variables and processes.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

O'Connor, M. C., & Michaels, S. (1993). Aligning academic tasks and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology and Education Quarterly, 24*, 318–335.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* New York, NY: Penguin Group.

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., . . . Gallimore, R. (2006). *Teaching science in five countries: Results From the TIMSS 1999 Video Study* (NCES 2006–011). Washington, DC: U.S. Government Printing Office.

Schauble, L. (2009, April). *What's foundational for science learning?* Plenary address presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Sharma, A., & Anderson, C. W. (2009). Recontextualization of science from lab to school: Implications for science literacy. *Science & Education, 18*, 1253–1275.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*, 49–100.

Tversky, A., & Kahneman, D. (2000). Judgment under uncertainty: Heuristics and biases. In T. Connolly, H. Arkes, & K. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 35–52). Cambridge, UK: Cambridge University Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. New York, NY: Cambridge University Press.

Wertsch, J. W. (1991). *Voices in the mind: A sociocultural approach to mediated action*. Cambridge, MA: Harvard University Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wiser, M., & Smith, C. L. (2008). Learning and teaching about matter in grades K-8: When should the atomic-molecular theory be introduced? In S. Vosniadou (Ed.), *The international handbook of research on conceptual change* (pp. 205–239). New York, NY: Routledge.

Wiser, M., Smith, C., Asbell-Clarke, J., & Doubler, S. (2009, April). Developing and refining a learning progression for matter: The inquiry project (grades 3–5). In C. Smith (Chair), *Developing and refining a learning progression for matter from pre-K to grade 12: Commonalities and contrasts among four current projects*. Symposium presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Zesaguli, J., Wilke, B., Hartley, L., Tan, E., Schenk, C., & Anderson, C. W. (2009, April). *Developing a K-12 learning progression for biodiversity in environmental systems*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

JULIA D. PLUMMER

CHALLENGES IN DEFINING AND VALIDATING AN ASTRONOMY LEARNING PROGRESSION

This chapter focuses on challenges in developing and validating a learning progression in the domain of astronomy. These challenges are not unique to astronomy; consideration of them may be useful for researchers working in other areas. As an astronomy education researcher, I believe learning progression research has the potential to provide needed coherence and direction for the field. In general, astronomy education receives a small amount of instructional time (Plummer & Zahm, 2010). Yet current instruction in astronomy, as in other science disciplines, is often fragmented, focused on breadth rather than depth, and with a greater emphasis on inconsequential facts than on the discipline's core (Kesidou & Roseman, 2002). This suggests that science teachers, in following this fragmented curriculum, may not take full advantage of the limited amount of time allocated to astronomy. In addition, the current research base lacks coherence across conceptual topics and has limited coverage of instructional interventions (for reviews, see Bailey & Slater, 2003; Lelliott & Rollnick, 2010). More work is needed to move astronomy education forward in ways that help teachers, as well as curriculum developers and assessment designers.

Through analysis of the logical structure of astronomy, review of relevant astronomy education literature, and consideration of learning progression research, I examine two areas necessary for defining a learning progression in astronomy: identifying the focus of the learning progression and obtaining empirical support for defining the learning progression, which includes validating the levels of increasing sophistication with respect to the targeted content. Within these two areas, I discuss the following six challenges that arose in defining a learning progression in the domain of astronomy.

1. *Identifying the learning progression focus:* The first challenge is to determine what constitutes a "big idea" in the domain of astronomy. As explained below, I chose the big idea of celestial motion. This choice leads to the second challenge: developing sophistication in celestial motion is specifically tied to learning about a finite set of observable phenomena. These phenomena define the conceptual space for learning celestial motion and place a limit on the available contexts in which students can learn about this big idea. This constraint leads to the third challenge: defining the celestial motion learning progression in a way that values the importance of both understanding observations from an earth-based perspective and learning the explanations for these observations. The fourth challenge deals with making adequate links to other big ideas and therefore other learning progressions.

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 77–100.

2. *Obtaining theoretical and empirical support for defining the learning progression:* While some areas of astronomy education are well-explored in the research literature, many areas have not been extensively researched. This is the fifth challenge: examining the literature on how children learn across time and highlighting the potential research-based pathways in the celestial motion learning progression. The sixth challenge examines how to obtain further empirical evidence to begin the process of validating the hypothetical learning progression, given limited student knowledge of even the most basic concepts and limited inclusion of astronomy in the curriculum.

In this chapter, I explain these challenges and propose solutions that may help the field of astronomy education move forward. The solutions proposed may also be applicable to other disciplines.

IDENTIFYING THE LEARNING PROGRESSION FOCUS

Identifying the focus for a learning progression includes articulating the big idea, a unifying concept that helps make sense of a broad variety of phenomena, situations, and problems. Big ideas have great explanatory power for the world (Smith, Wiser, Anderson, & Krajcik, 2006). Learning progressions describe how a learner may develop more sophisticated understandings of a big idea across time and through appropriate instruction (Smith et al., 2006). At one end of a learning progression is the upper anchor: the level of scientific understanding of the big idea as determined by societal goals for students. At the lower anchor of the learning progression is a description of children's knowledge and reasoning ability as they enter school (National Research Council [NRC], 2007). In this section, I explain the challenges that arose in choosing celestial motion as a big idea in astronomy and in defining a learning progression for this big idea.

Challenge #1: Identifying Big Ideas in Astronomy

Choosing an appropriate big idea for the domain of astronomy is the first challenge discussed in this chapter, as the answer is not obvious or clearly agreed upon by astronomy education researchers. Big ideas are descriptions of powerful explanatory models that have the far-reaching ability to explain a broad range of observable phenomena (NRC, 2007). There are many possible big ideas for the development of learning progressions in astronomy. In addition to the general definition stated above, I add the following criteria for big ideas in astronomy:

1. Big ideas are those that are important to the field of astronomy. Astronomy, as a science, is concerned with describing and explaining the universe as a whole. Thus big ideas in astronomy are those that represent ways of knowing and understanding the universe.
2. Big ideas describe explanatory models that can be learned by beginning with a child's observations of the world. This approach begins to capture the "increasing

in sophistication" criterion generally accepted in the definition of learning progressions (Corcoran, Mosher, & Rogat, 2009; NRC, 2007; Smith et al., 2006).

3. Big ideas can explain multiple, unified astronomical phenomena such that learning to explain an individual phenomenon helps the learner build in sophistication toward the big idea and, thus, toward explanations of other phenomena.

My research focuses on developing a learning progression for the big idea of "celestial motion." The big idea of celestial motion can be described as an answer to the question: How do we explain our earth-based perspectives of astronomical phenomena using the actual motions and properties of celestial objects? Astronomical phenomena observed from an earth-based perspective (such as the patterns of apparent daily motion of celestial objects, seasonal changes, and the phases of the moon) can be explained using the earth's rotation and tilt, the earth's orbit around the sun, and the moon's orbit around the earth (Plummer & Krajcik, 2010).

Ultimately, the big idea of celestial motion combines two concepts: *motions of celestial objects* and the observer's movement between *frames of reference* in order to understand observable phenomena. The various phenomena explained by this big idea are not caused by the same underlying motions; however, explanations of these phenomena are unified by their reliance upon the motion of celestial objects. In this chapter, I use several terms to describe critical features of the big idea of celestial motion. An *earth-based perspective* is generally used to describe more than just a single observation of the sky; however, I use this term to describe the appearance of a particular celestial object viewed from the earth across time (such as the sun rising and setting or the changing phases of the moon). The earth-based perspective contrasts with the *heliocentric model* (also referred to as the *explanatory motion*) that explains our earth-based perspective by describing the actual rotation or revolution of celestial bodies in the solar system. These two perspectives are each *frames of reference* from which we may define our description of a phenomenon. In this section, I explain the choice of the big idea of celestial motion as the learning progression upper anchor and describe other possible big ideas in astronomy.

There were two steps in my selection and definition of this big idea. First, I consulted policy documents (American Association for the Advancement of Science [AAAS], 1993; NRC, 1996), research syntheses (Adams & Slater, 2000; Agan & Sneider, 2003; Bailey & Slater, 2003; Kavanagh, Agan, & Sneider, 2005), and considered the logical conceptual structure of the domain. I used these documents and research studies for a consideration of the logical, conceptual structure of the domain. Second, I chose an explanatory model that provides coherence among the aspects of astronomy that are initially accessible (through personal and cultural experience) by young children. Specifically, the following topics are part of the same big idea of applying the motion of celestial objects to explain observations from an earth-based perspective: daily patterns in the apparent motion of the sun, moon, and stars; lunar phases; yearly patterns in the motion of the sun and stars; the reason for the seasons; and the motions of other solar system objects, such as the planets.[1] Understanding these apparent patterns of motion also requires understanding the earth's shape as well as its size and distance from other objects, both within the solar system and beyond (e.g., the earth's distance to the stars). These topics that relate to the big idea of celestial motion form the foundation of the major concepts of astronomy in the K-8 astronomy curricula (AAAS, 2001; NRC, 1996; Palen & Proctor, 2006).

One goal of developing learning progressions in science is to deepen the focus of science education on central concepts rather than on topical and inconsequential ideas. By focusing on celestial motion as an overarching big idea we take the focus off individual phenomena (e.g., day/night cycle, phases of the moon, seasons) and place more emphasis on connecting observations to underlying explanatory motions across multiple contexts. The expectation is that we will provide a more unified, integrated understanding of motion in the solar system. While not having the status of a universal theory, such as the Big Bang theory or the Universal Theory of Gravitation, celestial motion fits the criteria for big ideas in that it provides organization across a range of concepts in the domain and offers explanatory power with respect to a wide range of phenomena. Celestial motion also provides a useful framework to organize children's initial explorations in astronomy at a level that is accessible to them. Children have beliefs and personal observations about the appearance and apparent motion of the sun, moon, and stars (e.g., Plummer, 2009a; Vosniadou & Brewer, 1994). The big idea of celestial motion also focuses on a specific way of knowing that is important in astronomy: making observations of phenomena and then interpreting them in light of potential, unobserved motions. The concepts of rotation and revolution, which explain phenomena beyond the solar system, thus create a foundation for discussing advanced topics in astrophysics (e.g., binary stars, extrasolar planets, clusters of galaxies, and pulsars).

However, celestial motion is not the only big idea that could be selected. Other researchers have proposed alternative big ideas for astronomy. In their review of the Project 2061 science standards (AAAS, 2001), Lelliott and Rollnick (2010) propose eight big ideas: gravity, the solar system, stars, size and distance, earth shape, the day/night cycle, the seasons, and the earth/sun/moon system. They propose these eight big ideas because they are concepts often taught in school and have been subjects of extensive, ongoing educational research. Nevertheless, while clearly drawn from standards and the literature, these concepts do not help us see how students build in sophistication across the domain. Gravity is certainly a big idea, offering broad explanatory power for an extensive array of phenomena. However, the other seven proposed big ideas are topics rather than explanatory models. They do not offer useful ways of understanding the world and do not provide coherence for the learner.

Slater and Slater (2009) use existing standards and the expertise of astronomers and astronomy educators to create a list of 11 broad categories in astronomy that they link to the overarching big idea of the Big Bang theory. These 11 categories are: moon phases, daily/diurnal patterns, yearly patterns, size and scale, seasons, evolution and structure of planetary systems, stars and stellar evolution, formation of the universe, formation of elements, electromagnetic radiation, and gravity. While the Big Bang theory is undeniably an overarching and extremely important theory in science, a smaller grain-size big idea is needed to inform curriculum and standards that are useful in K-12 schooling. Ultimately, understanding how these concepts link to the Big Bang theory may be a goal in astronomy education. Thus, instead of beginning with the Big Bang theory as a big idea for a learning progression, we might view those categories that build in sophistication across multiple years (such as electromagnetic radiation, gravity, and perhaps stellar evolution) as potential big ideas that are a better fit to the big idea criteria stated above. Furthermore, four of Slater and Slater's categories are part of celestial motion (moon phases, diurnal patterns, yearly patterns, and seasons). Thus, I suggest that increasing in sophistication in celestial motion is a potential step towards the big idea of the Big Bang theory.

Challenge #2: Balancing the Variety of Phenomena Within the Big Idea

Within the commonly used definitions of learning progressions, *learning performances* represent the ways in which students may express understanding of the big idea at different levels of sophistication (Corcoran et al., 2009; NRC, 2007; Smith et al., 2006). Defining learning performances may be constrained by the nature of phenomena appropriate for the progression. In Smith et al.'s (2006) K-8 atomic molecular theory learning progression, learning performances are largely unconstrained by particular phenomena; students may learn about materials and properties of a wide range of objects in nearly any context. Similarly, Duncan, Rogat, and Yarden (2009) developed learning performances for a genetics learning progression that addresses a wide range of phenomena explained by the function of proteins in living organisms.

In contrast, there is a finite set of observable phenomena in celestial motion. These phenomena include the day/night cycle, daily observable patterns of rising and setting, the phases of the moon, eclipses, seasonal and latitudinal changes associated with changes in the sun's path, and seasonal star patterns. Other potential phenomena are the tides and retrograde motions of the planets. For example, explaining the relationship between the earth's rotation and our earth-based observations is limited to an understanding of the apparent motions of the sun, moon, and stars. Thus learning performances related to the effects of the moon's orbital motion are limited by our understanding of the earth-moon system. In addition, each phenomenon is coupled with a distinct set of explanatory motions rather than a single underlying explanation, as is the case for the genetics and molecular theory learning progressions.

Therefore, the second challenge in defining the celestial motion learning progression has to do with specifying the learning performances for celestial motion. This step requires exploring how each specific, finite context contributes to the overall model of observation and motion in the solar system (see Table 1). This method contrasts with the use of numerous phenomena to help the learner generalize the big idea across contexts, as occurs in learning progression research in other domains. If we focus too much on celestial motion as a generalized concept (e.g., "rotation and revolution in the solar system explain observable phenomena"), we lose the focus on how students learn to explain individual phenomena. If students are to use the underlying conceptual model to generate explanations, they must begin with something more concrete than generalized knowledge of rotation and revolution. Their knowledge of celestial motion should begin with instruction that is highly contextualized in familiar, observable phenomena. We hope that, through appropriate instruction, students will eventually reach a broader and more inter-connected view of celestial motion.

Increasing explanatory sophistication, from the lower anchor to the upper anchor of the learning progression for celestial motion, may mean that students have to learn to work with different time scales and combinations of motion as they explain more complicated phenomena. In some cases, this may mean learning to explain the same phenomena with greater sophistication. For example, students may learn initially that the moon rises and sets once every 24-hour period because of the earth's rotation. Later they may learn that the moon appears to rise and set about 50 minutes later every day because of its monthly orbit. The increase in explanatory sophistication depends on their knowledge of additional time scales and new motions. In other situations, students may apply concepts they learned about one phenomenon (e.g., explaining sunrise and sunset by the earth's rotation) using a more sophisticated explanation (e.g., including the earth's rotation in an explanation of the seasons).

The learning progression for celestial motion uses the explanatory motions of celestial objects as its backbone, to provide the coherence necessary in learning progression research. At the same time, clear connections are made to observable phenomena to acknowledge the central importance of specific phenomena associated with the domain. Table 1 shows how these underlying motions (e.g., the earth's rotation and orbit) can be combined to explain earth-based, observable phenomena.

Table 1. An Exploration of the Role of Heliocentric Motions in a Learning Progression for Celestial Motion.

Object-motion	*Relevant phenomena*
Earth-rotation	Day/night cycle; Daily apparent motion of sun, moon, stars, etc.
Earth-rotation + Moon-orbit	Lunar phases
Earth-rotation + Moon-orbit + Earth-orbit	Eclipses
Earth-rotation + Earth-orbit + Planet-orbit	Apparent motion of planets and retrograde motion
Earth-rotation + Spherical Earth[a]	Difference in visible constellations with latitude and circumpolar constellations
Earth-rotation + Earth-orbit	Seasonal stars
Earth-rotation + Spherical Earth[a] + Earth-tilt + Earth-orbit	Change in sun's path across the earth's surface and the seasons to explain the seasons

[a] The concept of a spherical earth is not a motion, as are rotation and orbit. However, our own motion across the spherical surface of the earth causes, for example, the visibility of constellations and differences in seasonal change.

Table 1 is not a learning progression for celestial motion in that it does not show a direct progression from naïve ideas, through increasingly sophisticated ideas, to the big idea. Table 1 presents aspects of the overall scientific model but does not tell us how concepts build on each other. For example, the phases of the moon may be best taught after students understand the day/night cycle (at its simplest, why we see the sun in the day and not at night). Does this teaching sequence mean understanding the phases of the moon requires a greater sophistication level of understanding than that needed for the day/night cycle?

The underlying conceptual model that explains the phases of the moon is more complex than that for the day/night cycle because an understanding of the phases of the moon requires an understanding of the earth's rotation and the moon's orbit. However, the day/night cycle is also part of a larger phenomenon: the appearance of the daily motion of celestial objects to an observer on earth is caused by the earth's rotation (e.g., the sun, moon, stars, and planets). This more detailed explanation of the effects of the earth's rotation, which is another conceptual area, is unnecessary for an understanding of the phases of the moon. Students could learn to explain the phases of the moon independent of learning to explain the stars' daily apparent celestial motion. Understanding the phases of the moon requires understanding how the earth's rotation explains our daily observations of the sun and moon. Therefore, the concept of the phases of the moon builds on the day/night cycle level of understanding but does not require knowledge of other aspects of daily celestial motion. Explaining the lunar phases also requires knowledge of how the moon's orbit affects our observations, making understanding the phases of the moon more than just a more sophisticated way of understanding the day/night cycle.

Wilson's (2009) proposal to build learning progressions from sets of *construct maps* may be a way to address these complexities around the big idea of celestial motion. Each construct map, highlighting a separate astronomical phenomenon, would allow us to focus on a single set of earth-based observations and their associated explanatory motions (e.g., daily apparent celestial motion and the rotation of the earth). Construct maps can be stacked or aligned to create a learning progression that leads to a single big idea that students may reach with appropriate instruction. Figure 1 shows a potential mapping of individual construct maps connected within a single learning progression for celestial motion, including earth-based observable phenomena and associated explanations in the heliocentric frame of reference. This mapping includes earth-based observable phenomena and associated explanations in the heliocentric frame of reference. The explanatory motions for each phenomenon (descriptions of actual motions in the solar system) are on the left side of the figure. Some explanatory motions of celestial objects correspond to multiple phenomena (and, thus, to multiple construct maps), such as the earth's rotation and orbit. This is shown by the grey shaded bands in Figure 1. Other explanatory motions, such as the orbits of the moon and the planets, only appear in a single construct map.

Figure 1, which is a rough sketch of the construct map layout and not a completely articulated or validated learning progression, does not show the intermediate levels or all the necessary links between the construct maps. However, Figure 1 provides a potential structure for future research. The height difference of the columns may represent the difficulty differences in achieving a scientific understanding of that construct. For example, it seems likely that learning to explain the seasons is more difficult than learning to explain the phases of the moon. However, there is limited empirical data that validates this hypothesis.

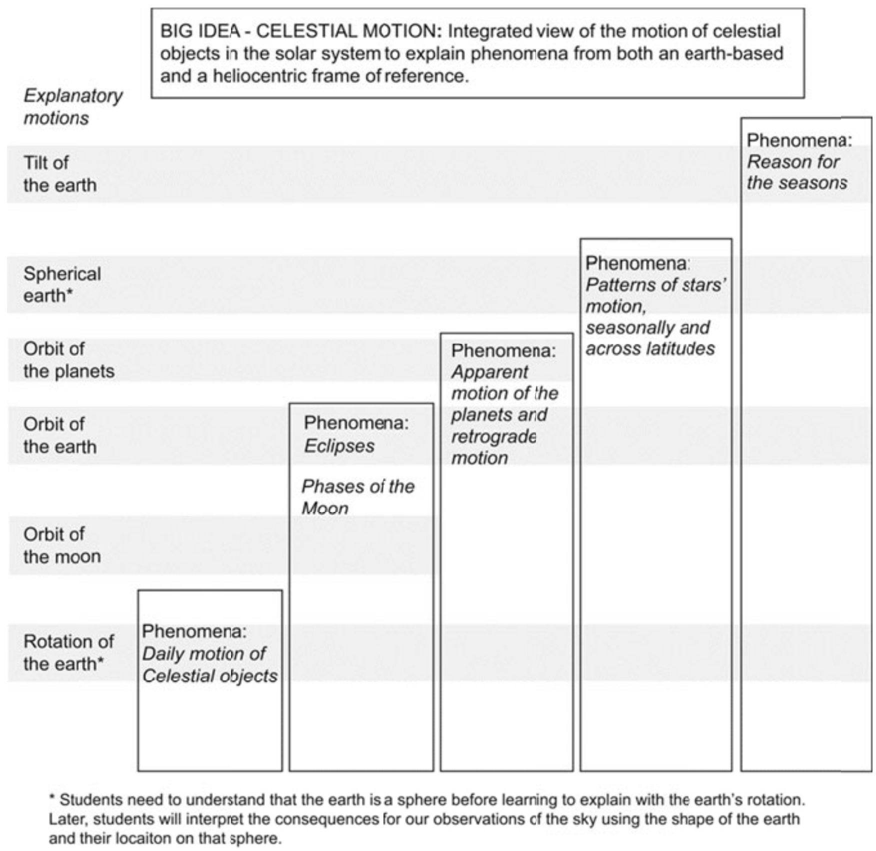


Figure 1. An outline of how earth-based phenomena and the actual heliocentric motions within the solar system can be linked within a learning progression for celestial motion. Each vertical column is a construct map. The gray-shaded bands indicate where explanatory motions (left column) link to each construct map. For example, the gray band for the earth's rotation overlaps all the construct maps because it is part of the explanation for all the phenomena.

In Figure 1, learning progresses upward through levels of increasing sophistication. The lower levels of each construct map represent students' naïve ways of knowing as they begin instruction. Higher levels of each construct map represent their increasingly sophisticated understanding of how celestial motion explains earth-based observable phenomena. A richer understanding of celestial motion is achieved when students explore connections between the different construct maps and see celestial motion not just as a collection of phenomena but as part of a larger pattern of motion. Instruction can begin along any of the construct maps in Figure 1, although future research may find that some starting points may be more fruitful for learners than others.

Challenge #3: Accounting for Both the Earth-Based Perspective (Observable Phenomena) and the Heliocentric Model (Explanatory Motion)

A third challenge arises as we consider the importance of both the *observational phenomena* in the earth-based frame of reference and the underlying *explanatory motion* in the heliocentric frame of reference. Merely understanding that the earth rotates, that the moon orbits, that the earth orbits the sun, that the earth is tilted, etc. is not enough to use these motions to explain earth-based observations. For example, elementary students may be able to state that the earth rotates and may demonstrate that concept with a model. However, when asked to explain why the sun rises and sets, they do not use the earth's rotation in their explanations (Plummer, Wasko, & Slagle, 2011). Thus additional attention must be paid to how students describe observable phenomena and understand the connection between that phenomena and the underlying explanatory motion. In building a learning progression for celestial motion, we must decide how to evaluate students' understanding of the apparent celestial motion. For example, consider a student who describes the sun as rising and setting in the same place on the horizon but explains this with an accurate description of the earth's rotation. His description of the earth-based perspective suggests that he is not reasoning between the frames of reference. Is his answer more sophisticated than one that includes an accurate description of the sun rising and setting in a smooth arc from east to west across the sky but an explanation that the earth rotates twice a day? Each child has a piece of the scientific model, but neither has a sophisticated understanding of the consequence of the earth's 24-hour rotation on our observations of the sun.

These examples suggest that development of a celestial motion learning progression will need to describe increasing sophistication of both descriptions of earth-based observable phenomena and explanations for those motions. My colleagues and I have studied the developing sophistication of children along a portion of the learning progression: the daily celestial motion construct map (the leftmost construct map in Figure 1; Plummer et al., 2011). Consistent with the design of other learning progressions, the daily celestial motion construct map (see Table 2) is anchored by a naïve understanding at one end and the full scientific understanding across both frames of reference at the other end. The construct map is organized on two dimensions. First, I organized students' ideas by their explanation (for example, do they use the earth's rotation or a non-normative explanation for the sun's rising and setting?). Within those groups, I organized the levels according to the accuracy of their descriptions of the apparent motion. In doing so, this construct map goes beyond the simple overview provided in Figure 1.

Each row in Table 2 describes a level of the construct map and represents progress along the construct, increasing in sophistication from bottom to top.[2] The left column is an overview of the construct map level. The middle column identifies the ways students might describe the earth-based observation at that level. The right column describes how students explain the earth-based observations at that level. An increase in sophistication results when students pair accurate descriptions with accurate explanations, showing that they link the two frames of reference. Students at lower levels do not make accurate connections between the frames of reference, but they have adopted aspects of the scientific concept. For example, a non-scientific description of the earth's rotation is an advance in sophistication over a belief that the sun orbits the earth (a more naïve perspective).

Table 2. Construct Map for Daily Celestial Motion.

Levels of the Construct Map	*Earth-based observed motions*	*Explanation for observed motions*
Scientific daily celestial motion: Students at the scientific level use the earth's rotation to explain all earth-based observed patterns of daily celestial motion. [NOTE: This level connects, as pre-requisite knowledge, to the *phases the moon* and *patterns of the stars' motion* construct maps.]	Students give an accurate description of the sun, moon, and stars' apparent daily motion by describing all as rising and setting in the same direction.	Students use the earth's rotation to explain all apparent daily motion.
Upper synthetic: Students use the earth's rotation to explain that the sun appears to rise and set across the sky. However, students do not extend this explanation to both the moon and stars.	All students in upper synthetic give a scientific description for the apparent motion of the sun. Within this level, there are students who may also give the scientific description for the moon and stars' apparent motion as well.	Students accurately describe the earth's rotation. Students may use the earth's rotation to explain only the sun's apparent motion or they may also explain the moon or stars' apparent motion accurately.
Lower synthetic: Students believe that the sun is stationary and that the earth is moving. Students' descriptions and explanations for the moon and stars' apparent motion are likely to retain the inaccuracies of the naïve perspective; this level is primarily determined by how the students explain the sun's apparent motion. There may be limited coherence between the actual motion of the earth and apparent patterns of motion of other celestial objects.	The apparent motion of the sun, moon, and stars may or may not be accurately described	Explanation for sun's apparent motion includes less sophisticated ideas (e.g., the earth orbits the sun once a day) and more sophisticated ideas (e.g., using the earth's rotation in combination with other inaccurate explanations).
Naïve: This level represents where most students enter elementary school. Students at this level believe that the earth-based patterns of motion (or lack of motion) are due to the objects' actual motion (or lack of motion).	Some students may be able to provide relatively accurate descriptions of the sun and moon's apparent motion while others provide only non-scientific descriptions. Most students believe that the stars do not move or only move at the end of the night.	Explanations use the sun, moon, and stars' actual motion.

Challenge #4: Making Links to Other Learning Progressions

A fourth challenge in designing a meaningful learning progression arises in the consideration of students' understanding of related concepts necessary for full understanding of the targeted content and how these related concepts fit within a learning progression framework. Connections between big ideas should be made explicit as we move forward so that learning progressions are useful to curriculum developers, assessment designers, and policy makers. A major criticism of K-12 school instruction is that students are not forming deep and rich connections across science topics (Corcoran et al., 2009; Kesidou & Roseman, 2002; Schmidt, McKnight, & Raizen, 1997). The ability to draw connections within and between such topics distinguishes a novice from an expert; integrated knowledge permits flexible retrieval of information that can be used in problem solving situations (NRC, 1999).

Understanding the big idea of celestial motion requires understanding related big ideas. Several areas associated with the big idea of celestial motion could be developed as separate learning progressions. These areas include size and scale, light and energy, spatial reasoning, and scientific modeling (scientific modeling appears in a learning progression developed by Schwarz, Reiser, Acher, Kenyon, & Fortus, this volume). At lower levels, understanding the size of and distance to celestial objects is important for learning about daily celestial motion. The progress from a naïve perspective (the sun and moon move around the stationary earth while the stars stay still) to the scientific perspective (the earth rotates once a day causing the relatively stationary sun and stars and the slow-moving moon to appear to rise and set) is assisted when students learn that the sun is very far from the earth, the sun is very large compared to the earth, and the stars are similarly large but much farther away. The moon's size and distance from the earth also become useful in understanding why the moon slowly orbits the earth and contributes to understanding the difference between phases of the moon and eclipses. Understanding the properties of light is important as students progress to more advanced topics in astronomy. For example, understanding the phases of the moon requires that students understand that the moon is lit by sunlight, reflected off the moon's surface, that then travels in a straight line to our eyes. Examples such as these demonstrate that building sophistication in astronomy means that students are learning to apply more sophisticated concepts of celestial motion to observable phenomena and to make connections to other concepts.

Ultimately, moving toward more sophisticated levels of astronomy than are depicted in the five construct maps (see Figure 1) involves integration with big ideas in physics, such as gravity. For example, the celestial motion learning progression leads to explanations of earth-based observations of the apparent motions of the sun, moon, stars, and planets; explaining why the planets and moon orbit in the ways that they do, as well as how those orbits first began (the formation of the solar system) requires the use of gravitational theory. These explanations require an understanding of gravitational theory. If learning progressions are developed using structures similar to those used in the interconnected construct maps approach (Wilson, 2009), then perhaps making links between learning progressions is a process of finding alignments between segments of the concept maps that make up larger learning progressions.

OBTAINING THEORETICAL AND EMPIRICAL SUPPORT FOR DEFINING THE LEARNING PROGRESSION

Learning progressions are not natural or developmental progressions of understanding; they describe what may be attained through appropriate instruction. After unpacking the concepts through a domain analysis, development of a learning progression relies on the potentially productive pathways that research identifies between naïve and scientific levels of understanding. We can draw on the literature that describes students' alternative conceptions about celestial motion to help define the entry points—what students believe as they enter school. These beliefs are the lower anchor of the learning progression. Cross-sectional research may tell us about likely progressions of concepts based on traditional instruction. To advance further, design-based research is needed to test these potential pathways that result from instruction designed to support students' movement along the progression. This research may allow us to identify productive instructional sequences that move students toward the upper anchor. By examining the existing literature, we can uncover the ways in which that literature can help us hypothesize about productive sequences and where additional research is needed to provide a comprehensive, multi-year understanding of what progress towards the big idea might look like.

In this section, I discuss the challenges presented by limits of the existing literature base in celestial motion and my research group's attempts to extend the research in these areas. This discussion addresses the challenges researchers face in investigating how students may reach the upper level of sophistication of the learning progression for celestial motion given the limited place of astronomy in school science curricula and upper level students' lack of foundational knowledge of astronomy.

Challenge #5: Using the Existing Literature Base

In this section, I review the literature that addresses students' naïve scientific understanding when they begin formal instruction on astronomy. I also examine the potential research-based pathways along the construct maps in the celestial motion learning progression. This discussion also highlights areas where additional research is needed to overcome challenges in defining a hypothetical learning progression using the current literature base.

Extensive research has been conducted on children's naïve beliefs as they begin school, especially concerning the shape of the earth and the reason for the day/night cycle (see the review by Lelliott & Rollnick, 2010). For example, several researchers have described and refined a developmental progression for the earth's shape and its role in children's personal cosmologies. This progression begins with the commonly held belief that the earth is flat and objects fall "down" toward some nonphysical cosmic “bottom” (e.g., Nussbaum & Novak, 1976; Nussbaum & Sharoni-Dagan, 1983; Sneider & Pulos, 1983; Vosniadou & Brewer, 1992). Research on children’s explanations of the day/night cycle demonstrates that children begin school believing day and night are caused by the sun’s actual motion or objects blocking the sun (Samarapungavan, Vosniadou, & Brewer, 1996; Vosniadou & Brewer, 1994). Research on the phases of the moon suggests that many early elementary students believe that clouds cause the phases of the moon while older students commonly believe the earth’s shadow causes the phases of the moon (Baxter, 1989). Literature on students’ conceptions has also examined various topics associated with celestial motion, such as the aspects of the seasons (e.g., Baxter, 1989), the solar system (e.g., Sharp, 1996), and the nature of the stars (e.g., Agan, 2004). This research on students’ early cognition in astronomy helps us understand the lower levels of the learning progression on celestial motion; however, validating a learning progression that includes the upper levels of sophistication requires understanding the role of targeted instruction in developing student understanding.

In astronomy education, research on the impact of instruction is limited in its scope (Bailey & Slater, 2003). Most astronomy education research focuses on students’ and teachers’ knowledge of concepts and their mental models (Lelliott & Rollnick, 2010); there are few longitudinal studies and little focus on the effect of instruction or on the connections between learning various astronomical concepts and building concepts over time. The research is also often limited by its focus on single concepts rather a focus on students’ development of an integrated understanding of astronomical phenomena. While there has been more research on astronomy instruction in recent years (Kavanagh, 2007), much more is needed.

Despite these limitations, astronomy education research provides some evidence on how students learn the phases of the moon and the seasons. This research can be used to inform development of a learning progression for celestial motion. While learning to describe the observable pattern of the phases of the moon is relatively straightforward for children, using the relative positions and movements in the sun-earth-moon system to explain these phases is challenging for learners of all ages (Lelliott & Rollnick, 2010). Early elementary students can learn to describe and illustrate the phases of the moon; there is no indication that specific pre-requisite knowledge is needed for learning this pattern (Hobson, Trundle, & Sackes, 2010; Trundle, Atwood, & Christopher, 2007). In a study of students in a New Zealand intermediate school, Taylor, Barker, and Jones (2003) describe instruction designed to promote the development of a scientific mental model of the sun-earth-moon system by allowing students to offer their own prior knowledge and then to critique the teacher’s use of a physical model. While 90% of the students could accurately describe the orbital motion of the moon and earth, only 15% could accurately explain the phases of the moon. This finding suggests that lunar phases are sufficiently challenging that an awareness of the actual motions (such as the earth’s rotation and the moon’s orbit) does not necessarily lead students to construct a scientific explanation by themselves or with minimal instruction. Other studies suggest that increased sophistication in explanations for lunar phases requires support in describing the observable pattern of change in the phases followed by instruction that directly engages students in generating explanations, using either physical models or computer simulations (Barnett & Morran, 2002; Trundle et al., 2007; Trundle, Atwood, Christopher, & Sackes, 2010). However, research has not yet shown how an understanding of the earth's rotation or of the size and scale of the sun-earth-moon system affects students' ability to explain the phases of the moon.

Seasonal change is another key phenomenon of celestial motion. Extensive research has demonstrated that most people cannot accurately explain the seasons; the most common non-normative explanations are that the earth is moving closer to and farther from the sun (e.g., Atwood & Atwood, 1996; Baxter, 1989; Kikas, 1998; Schoon, 1995; Sharp, 1996). In addition, a lack of understanding that the sun's apparent daily path changes across the seasons (Plummer, 2009a) and a non-normative belief that the earth's orbit is highly elliptical (Kikas, 1998; Schneps & Sadler, 1988) contribute to the difficulty that children have in learning to explain the seasons. Recent studies have documented successful instructional approaches for teaching the explanations for the seasons (Hsu, 2008; Slater, Morrow, & Slater, 2008; Tsai & Chang, 2005). However, these studies present a limited explanation of how students understand the seasons from both an earth-based perspective and a heliocentric perspective. They also do not address how prerequisite knowledge might influence students' learning of this challenging concept or how understanding this concept might be influenced by other aspects of the celestial motion big idea.

While a significant amount of research has explored instruction related to the seasons and the phases of the moon, research on instruction related to other phenomena associated with celestial motion is relatively limited. Only a few studies have analyzed children's knowledge of the motions of the solar system as a whole (e.g., Sharp, 1996; Treagust & Smith, 1989). Sharp and Kuerbis's (2005) study is perhaps the only study that investigates instruction on motion in the solar system. Their study reveals that students showed improvement in describing the motions of the planets in the solar system. However, these students were not assessed on how they used the actual motions to explain observable phenomena. Research on the effect of instruction related to the apparent motion of the stars, as well as their size and distance, is also limited. A few studies have examined children's explanations of the daily motion of the stars (Baxter & Preece, 2000; Dove, 2002), but there is not much research that describes how students learn to explain more advanced aspects of the stars' apparent motion, such as seasonal changes or how apparent motion changes based on one's location on earth.

Understanding and using celestial motion requires students' use of spatial abilities: mental rotation, spatial perception, and spatial visualization (Black, 2005; Linn & Petersen, 1985; Wilhelm, 2009). Although a few researchers have begun to investigate the importance of spatial reasoning in instructional-based studies of celestial motion (Sherrod & Wilhelm, 2009; Wilhelm, 2009), much remains to be done as this research only addresses the phases of the moon. In addition, as mentioned above, a major concept embedded in learning about celestial motion is size and scale. While several studies have investigated students' ability to make comparisons of relative sizes and distances of celestial objects (e.g., Agan, 2004; Bakas & Mikropoulos, 2003; Sharp, 1996), few studies have reported on attempts to teach astronomical size and scale and to build on these concepts for the understanding of celestial motion. One exception is a study examining the "Powers of Ten" video (http://powersof10.com) that has been shown to increase the accuracy of students' use of relative size and to improve their ability to match objects to their actual metric sizes (Jones, Taylor, Minogue et al., 2006).

These and other studies in astronomy education research focus primarily on the individual features of the celestial motion conceptual domain rather than on students' understanding across multiple aspects. Few studies include longitudinal data that would allow us to investigate improvement in these connections across time (Briggs, Alonzo, Schwab, & Wilson, 2006). These connections across the associated phenomena are needed to define and validate the learning progression in such a way that the process is more than an unpacking of the domain. Furthermore, understanding how and why students develop in sophistication along and between the construct maps will require a description of successful instructional practices. There is little research that shows the pathways from children's initial understanding of apparent celestial motion to a fully articulated model of celestial motion. In addition, there is limited research on instruction that helps students connect earth-based descriptions of phenomena to explanatory motions. Children are rarely asked to compare different frames of reference; when observable phenomena are addressed in research on instruction, studies seldom address how students use celestial motion to predict and explain observations. Because of the limited research on using instruction to develop integrated knowledge of celestial motion phenomena, the design and validation of this learning progression requires multiple studies across many grade levels and instructional conditions.

MOVING THE AGENDA FORWARD WITH LEARNING PROGRESSION RESEARCH

My colleagues and I have begun to conduct research that fills in a few of the gaps in the literature on celestial motion in order to move towards a more comprehensive learning progression. Our specific goals are to investigate (a) how students learn to move between frames of reference, (b) how instruction can support students in building in sophistication upwards with respect to the progression, and (c) how instruction supports connections across constructs (between phenomena) within the progression. Next I describe two studies we used to define levels for the daily celestial motion and the seasons construct maps within the celestial motion big idea.

Daily Celestial Motion

Because of the limited research on children's ability to describe observed phenomena from the earth-based perspective, we designed the first set of studies to improve our understanding of children's descriptions of the apparent motion of celestial objects (Plummer, 2009a, 2009b; Plummer & Krajcik, 2010). These studies were undertaken (a) to provide a portion of the lower anchor for the learning progression, (b) to offer cross-sectional data to illuminate the ways in which traditional instruction and experiences with the world influence students' initial ideas, and (c) to investigate the effect of a targeted intervention on students' understanding of the earth-based perspective.

Learning to describe celestial motion from an earth-based perspective is just the first step in improving understanding of the big idea. Sophistication increases as students learn to explain their observations in the earth-based frame of reference with the actual motions of celestial objects (Plummer et al., 2011). To understand daily celestial motion from both frames of reference, we hypothesize that children need (a) to experience visual and/or kinesthetic descriptions of the apparent patterns that are then explicitly connected to explanations that use the earth's rotation and (b) to confront the common, non-normative understanding of the moon's orbit to explain the moon's daily apparent motion. Building on these ideas, we used a design-based approach (Collins, Joseph, & Bielaczyc, 2004) for instruction that supports children in moving between frames of reference. We started with a small group of gifted third graders in a pilot study (N=16; Plummer et al., 2011). The results support our hypothesis that movement along the construct map can be accomplished by instruction that combines visual and kinesthetic instruction with the previously described methods for learning the apparent motions (Plummer, 2009b).

Building on these results, we have begun to analyze the results of integrating these strategies into the regular third grade astronomy curriculum in a suburban school district (N = 99; Plummer, Kocareli, & Slagle, 2010). To understand the nature of student improvement with instruction, we analyzed outcomes of four instructional conditions that varied in the level and type of instructional support for students. Our analysis of children's improvement in the four conditions suggests that children who experience instruction that focuses primarily on heliocentric motions (e.g., rotation of the earth and orbit of the moon) show limited improvement in their understanding of the earth-based frame of reference; similarly, a sole focus on the earth-based perspective does not allow students to automatically connect those observations to the earth's rotation. We analyzed how students' understandings changed and improved based on the instructional conditions. We examined frequencies in the transitions student made from pre- to post-instruction in order to identify aspects of daily celestial motion that appeared necessary for more sophisticated levels of understanding. This analysis supported our hypothesis that understanding how the earth's rotation explains the sun's apparent motion is an important intermediate level in the progress toward more sophisticated understandings, such as explaining the apparent celestial motion of the moon and the stars.

Our next step in defining and validating the learning progression is to look for the ways that students combine aspects of celestial motion in their explanations of more advanced phenomena (looking both horizontally and vertically in the learning progression in Figure 1). The construct map for daily celestial motion (Table 2) describes increasing sophistication in the use of the earth's rotation to explain observable phenomena. This daily celestial motion construct map connects to the other construct maps in the learning progression because the earth's rotation is part of the explanation of other phenomena. However, a full understanding of daily celestial motion is not a precursor to the other constructs; rather, aspects of daily celestial motion link to the other construct maps as prerequisite knowledge. Figure 2 shows these links between the daily celestial motion construct map and the other celestial motion construct maps. For example, a full understanding of lunar phases and eclipses includes understanding how the earth's rotation causes the moon's daily pattern of motion. This understanding is necessary for explaining the correlation between the moon's appearance and the times of its rising and setting. Understanding the daily celestial motion of the moon also helps students distinguish between the scientific explanation of the lunar phases and a common misconception that they are caused by the earth's rotation (Trundle et al., 2010). In our continued analysis, we will investigate this connection in terms of the patterns of improvement observed in the third grade student data.

Reason for the Seasons

Building on our understanding of how children learn to explain daily celestial motion, we have also begun to investigate older students' explanations of how patterns in the sun's apparent motion cause the seasons (Plummer & Maynard, 2012). Our research examines eighth grade students learning both patterns associated with an earth-based perspective and the explanations for those patterns. Using a Rasch modeling approach, we identified a potential ordering of concepts relating to the seasons, from least difficult to most difficult. Based on this quantitative analysis, we identified a set of levels describing increases in sophistication that define a construct map for the seasons. We further refined the construct map using the tentative levels from the Rasch analysis as a tool to classify specific students' knowledge. To do so, we examined how higher levels of the construct map build on previous levels using a Guttman scale approach (assuming that understandings at a given level include those in the previous levels). Using their responses to the assessment, we assigned the students to the identified levels using the Rasch analysis. The analysis revealed that students may reach intermediate levels of the construct map without being able to accurately explain the sun's daily motion. This analysis led us to tentatively link the daily celestial motion construct map to the reason for the seasons construct map at the scientific level.

However, this is a tentative description of our findings. Additional research is needed to test and validate the seasons construct map and its connection to other construct maps.

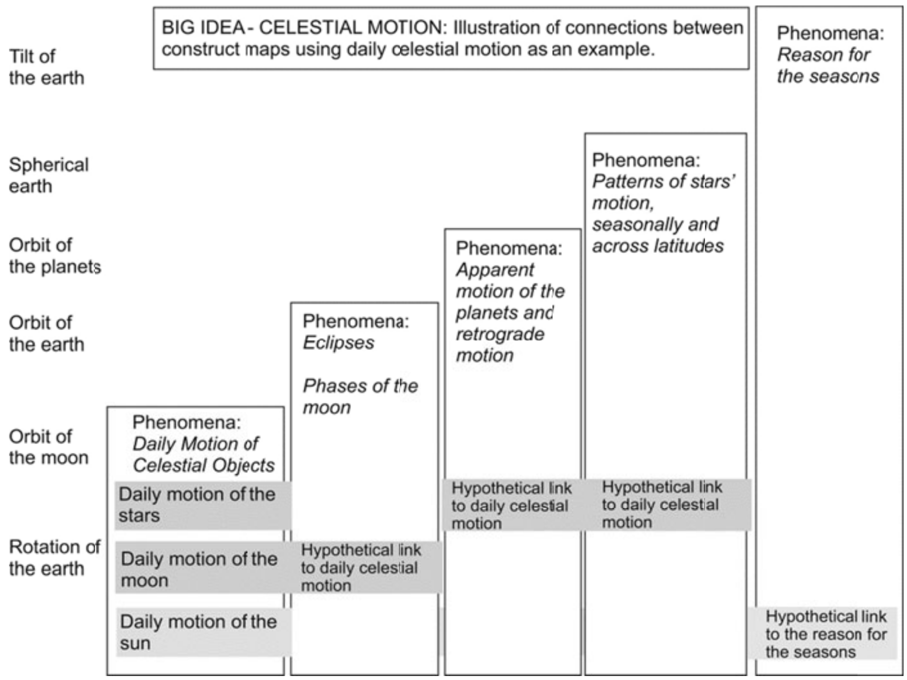


Figure 2. An outline of the aspects of the daily celestial motion construct map necessary for explaining more advanced concepts (in other construct maps).

Very few students reached the scientific level of the progression. The reason may be, in part, that many students had not achieved a robust mental model of the sun-earth portion of the daily celestial motion construct map, a prerequisite for the top level of the seasons construct map. Without a complete understanding of the sun's apparent motion, many students did not have the basic knowledge required to advance to more complex concepts on the seasons construct map (Plummer & Maynard, 2012). Another possible explanation for the low percentage of students achieving the scientific level relates to the amount of classroom instruction; only one day, in a 10-day curriculum, dealt with integrating the students' observational knowledge with the tilt-model that explains the seasons. Possibly with more classroom time and instruction, more students would have reached the scientific level. The research described here would be strengthened by testing the construct maps with students in different contexts. Such testing might include exploring learning pathways across different cultures and examining the roles of geographic location and local temperature patterns in developing an understanding of celestial motion.

Challenge #6: Obtaining Empirical Support for Hypothetical Learning Progressions

Our research reveals an additional challenge in obtaining empirical support to validate the celestial motion learning progression. Significant research has shown that many, if not most, children and adults do not have the foundational knowledge– those concepts that form the initial levels of the construct maps– needed to support more sophisticated levels of understanding (e.g., Atwood & Atwood, 1995; Baxter, 1989; Brunsell & Marcks, 2004; Mant & Summers, 1993; Plummer, 2009a; Plummer & Maynard, 2012; Plummer, Zahm, & Rice, 2010; Schoon, 1995; Sharp, 1996; Trumper, 2006). This lack of foundational knowledge of astronomy means that, for older students, we cannot begin instruction at the more intermediate levels of the progression; advancing to the scientific levels requires beginning with some of the more elementary concepts of astronomy (e.g., daily patterns of motion and the earth's rotation as an explanation of those patterns).

While the students in our study showed overall improvement in their understanding of the seasons, their learning may have been hindered by their lack of fundamental knowledge of astronomy (Plummer & Maynard, 2012). This problem suggests that, for many teachers, reaching the end goal of the learning progression may also mean teaching the foundational concepts, at least until school curricula are designed to address these foundations sufficiently at younger grades. Testing and validating additional aspects of a celestial motion learning progression are therefore problematic because of the amount of classroom time required to help students reach advanced levels beyond their naïve level of understanding. Students could move to these advanced levels more readily if more time and effort were devoted to providing them at an earlier stage with at least some foundational knowledge of celestial motion.

One reason for students' lack of foundational knowledge is that coverage of astronomy is limited across K-12 schooling. Though I have been unable to find studies directly measuring the coverage of astronomy at the elementary level, research suggests that many students are not studying astronomy in middle or high school (Plummer & Zahm, 2010). As a result, researchers are limited in their ability to test theories in the context of classroom-based instruction. Secondary schools that teach astronomy often do so in very short time frames (Plummer & Zahm, 2010). If students have not learned the foundational concepts in elementary school, the fast-paced coverage in secondary schools is unlikely to result in a scientific understanding of the target concepts.

How can we meet the challenge of obtaining empirical support? First, researchers should stress the importance of pre-assessment of foundational concepts in how we articulate learning progressions. We should emphasize that learning progressions do not describe students' knowledge at particular grades; they describe intermediate steps that can be accomplished through well-crafted instruction. In other words, it is important to emphasize that progress is not inevitable and that instruction at higher levels of the learning progression should not proceed unless students have acquired the necessary foundational knowledge (from the lower levels of the progression). Second, researchers should identify school districts that provide multiple opportunities for students to study astronomy at increasing levels of sophistication. These research settings would allow researchers to explore how students develop sophistication in astronomy through repeated explorations of concepts, possibly using longitudinal studies. Reform-based curricula, based on research findings about teachers' pedagogical content knowledge and common alternative conceptions in astronomy, along with a clear plan to support teachers through professional development, are also needed. Examples of schools or districts that demonstrate successful science instruction based on learning progressions may encourage other districts to adopt similar instructional strategies.

CONCLUSIONS

In this chapter, I have presented initial research conducted to define a learning progression in astronomy and have articulated several challenges. The solutions presented to these challenges may be of use to researchers developing learning progressions around other big ideas of science.

The first set of challenges discusses the focus and organization of the learning progression. Other researchers may find it fruitful to consider the benefits of using construct maps to organize smaller elements of their learning progressions. This could be done to describe how students may increase in sophistication in their understanding of various phenomena or to demonstrate ways that learning can occur along different pathways (such as showing that learning to explain the phases of the moon is a separate, not a prerequisite, knowledge from learning to explain the seasons), both of which occur in the celestial motion learning progression. Construct maps may also serve as useful organizational tools as learning progression researchers consider ways to define learning progressions across both content and scientific reasoning abilities, such as in Songer, Kelcey, and Gotwals' (2009) complex reasoning in biodiversity learning progression. The choice to use construct maps, as well as the organization of the construct maps, will depend on the nature of the big idea..

The first set of challenges also refers to the connections between learning progressions. The identification of how concepts are connected between learning progressions (such as the importance of understanding properties of light in developing a rich understanding of celestial motion) may lead to a more sophisticated understanding of the big idea (such as extending the celestial motion learning progression to the big idea of gravity). Ultimately, moving to more sophisticated levels of understanding astronomy requires that students deepen their understanding of physics as well as their scientific reasoning skills.

The second set of challenges identified in this chapter concerns the testing and validation of the celestial motion learning progression. Research in this area is limited by gaps in the current astronomy education research base —a problem that is likely to arise in many other areas of science education. While it is clear that much additional research is needed on astronomy education, identifying the most appropriate instruction and conditions for testing and validating the learning progression will require extensive time and effort. For example, while longitudinal studies may help us answer questions about the validity of the learning progression, are such studies possible? Research that tests and validates the celestial motion learning progression is difficult because of the position of astronomy education in schools today. Because astronomy is rarely taught in K-12 schools, it is difficult to identify school settings where the learning progression can be studied (Plummer & Zahm, 2010). Many districts require that teachers "teach to the test" and/or follow a standard curriculum in step-by-step fashion; other school districts schedule limited instructional time for astronomy. Therefore, external pressures will make large-scale validation projects a challenge in this domain as well as other aspects of sciences, which receive limited inclusion in school curricula. Limited instructional time challenges us to consider what is considered "good enough" in the context of this big idea and how to communicate potential trade-offs to teachers, curriculum developers, and policy makers.

The research described in this chapter presents the initial steps toward developing a learning progression for celestial motion. Additional empirical evidence is needed to define and validate the levels of the construct maps that make up this learning progression. In addition, the big idea, as described in this chapter, is only one possible approach in learning progression research in astronomy. Other big ideas in astronomy, leading to robust knowledge appropriate to K-12 education, should be explored. Such exploration may lead to the definition and validation of additional learning progressions that support improvement of K-12 astronomy education through the development of more coherent standards and more research-based curricula.

NOTES

1. Explaining the earth-based perspective with the actual motions of other objects can ultimately be used to understand other phenomena in the universe, such as our observations of pulsars and the shapes of planetary nebulae.
2. A more detailed description of the levels is in Plummer et al. (2011).

REFERENCES

Adams, J., & Slater, T. (2000). Astronomy in the National Science Education Standards. *Journal of Geoscience Education, 48*, 39–45.

Agan, L. (2004). Stellar ideas: Exploring students' understanding of stars. *Astronomy Education Review, 3*(1), 77–97.

Agan, L., & Sneider, C. (2003). Learning about the Earth's shape and gravity: A guide for teachers and curriculum developers. *Astronomy Education Review, 2*(2), 90–117.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. Washington, DC: Author.

Atwood, V. A., & Atwood, R.K. (1995). Preservice elementary teachers' conceptions of what causes night and day. *School Science and Mathematics, 95*, 290–294.

Atwood, R. K., & Atwood, V. A. (1996). Preservice elementary teachers' conceptions of the causes of seasons. *Journal of Research in Science Teaching, 33*, 553–563.

Bailey, J. M., & Slater, T. F. (2003).A review of astronomy education research. *Astronomy Education Review, 2*(2), 20–45.

Bakas, C., & Mikropoulos, T. (2003). Design of virtual environments for the comprehension of planetary phenomena based on students' ideas. *International Journal of Science Education, 25*, 949–967.

Barnett, M., & Morran, J. (2002). Addressing children's alternative frameworks of the Moon's phases and eclipses. *International Journal of Science Education, 24*, 859–879.
Baxter, J. (1989). Children's understanding of familiar astronomical events. *International Journal of Science Education, 11*, 502–513.
Baxter, J. H., & Preece, P. F. W. (2000). A comparison of dome and computer planetarium in teaching of astronomy. *Research in Science and Technology Education, 18*, 63–69.
Black, A. A. (2005). Spatial ability and earth science conceptual understanding. *Journal of Geoscience Education, 53*, 402–414.
Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.
Brunsell, E., & Marcks, J. (2004). Identifying a baseline for teachers' astronomical concept knowledge. *Astronomy Education Review, 3*(2), 38–46.
Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, *13*, 15–42.
Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.
Dove, J. (2002). Does the man in the moon ever sleep? An analysis of student answers about simple astronomical events: A case study. *International Journal of Science Education 24*, 823–834.
Duncan, R. G., Rogat, A., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th-10th grades. *Journal of Research in Science Teaching, 46*, 655–674.
Hobson, S. M., Trundle, K. C., & Sackes, M. (2010). Using a planetarium software program to promote conceptual change with young children. *Journal of Science Education and Technology*, *19*, 165–176.
Hsu, Y.-S. (2008). Learning about seasons in a technologically enhanced environment: The impact of teacher-guided and student-centered instructional approaches on the process of students' conceptual change. *Science Education, 92*, 320–344.
Jones, M. G., Taylor, A., Minogue, J., Broadwell, B., Wiebe, E., & Carter, G. (2006). Understanding scale: Powers of ten. *Journal of Science Education and Technology, 16*, 191–202.
Kavanagh, C. (2007, September). *Astronomy education research bibliography*. Paper presented at the Astronomy Education Research Symposium, Tufts University, Medford, MA.
Kavanagh, C., Agan, L., & Sneider, C. (2005). Learning about phases of the Moon and eclipses: A guide for teachers and curriculum developers. *Astronomy Education Review 4*(1), 19–52.
Kesidou, S., & Roseman, J. E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching, 39*, 522–549.
Kikas, E. (1998). Pupils' explanations of seasonal changes: Age differences and the influence of teaching. *British Journal of Educational Psychology, 68*, 505–516.
Lelliott, A., & Rollnick, M. (2010). Big ideas: A review of astronomy education research from 1975–2008. *International Journal of Science Education*, *32*, 1771–1799.
Linn, M.C.,& Petersen, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479–1498.
Mant, J., & Summers, M. (1993). Some primary-school teachers' understanding of Earth's place in the universe. *Research Papers in Education, 8*, 101–129.
National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.
Nussbaum, J., & Novak, J. D. (1976). An assessment of children's concepts of the earth utilizing structured interviews. *Science Education, 60*, 535–550.

Nussbaum, J., & Sharoni-Dagan, N. (1983). Changes in second grade children's preconceptions about the earth as a cosmic body resulting from a short series of audio-tutorial lessons. *Science Education, 67*, 99–114.

Palen, S.,& Proctor, A. (2006). Astronomy in the K-8 core curriculum: A survey of state requirements nationwide. *Astronomy Education Review, 5*(1), 23–35.

Plummer, J. D. (2009a). A cross-age study of children's knowledge of apparent celestial motion. *International Journal of Science Education, 31*, 1571–1606.

Plummer, J. D. (2009b). Early elementary students' development of astronomy concepts in the planetarium. *Journal of Research in Science Teaching, 46*, 192–209.

Plummer, J.D. & Maynard, L. (2012). Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. Manuscript submitted for publication.

Plummer, J. D., Kocareli, A., & Slagle, C. (2011). Children developing an ability to move between frames of reference in astronomy: Towards a learning progression in celestial motion. Paper presented at the Annual Meeting of the Jean Piaget Society, Berkeley, CA.

Plummer, J.D.,& Krajcik, J.S. (2010). Building a learning progression for celestial motion: Elementary levels from an earth-based perspective. *Journal of Research in Science Teaching, 47*, 768–787.

Plummer, J. D., Wasko, K., & Slagle, C. (2011). Children learning to explain daily celestial motion: Understanding astronomy across moving frames of reference. *International Journal of Science Education, 33*, 1963–1992.

Plummer, J. D., & Zahm, V. (2010). Covering the standards: Astronomy teachers' preparation and beliefs.*Astronomy Education Review, 9*(1). doi:10.3847/AER2009077

Plummer, J. D., Zahm, V., & Rice, R. (2010). Inquiry and astronomy: Preservice teachers' investigations in celestial motion. *Journal of Science Teacher Education, 21*, 471–493.

Samarapungavan, A., Vosniadou, S., & Brewer, W. F. (1996). Mental models of the earth, sun, and moon: Indian children's cosmologies. *Cognitive Development, 11*, 491–521.

Schmidt, W. H., McKnight, C. C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schneps, M. H., & Sadler, P. (1988). *A private universe* [DVD]. Available from www.learner.org

Schoon, K.J. (1995). The origin and extent of alternative conceptions in the earth and space sciences: A survey of pre-service elementary teachers. *Journal of Elementary Science Education, 7*, 27–46.

Sharp, J. G. (1996). Children's astronomical beliefs: A preliminary study of year 6 children in south-west England. *International Journal of Science Education, 18*, 685–712.

Sharp, J. G., & Kuerbis, P. (2005). Children's ideas about the solar system and the chaos in learning science. *Science Education, 90,* 124–147.

Sherrod, S. E., & Wilhelm, J. (2009). A study of how classroom dialogue facilitates the development of geometric spatial concepts related to understanding the cause of Moon phases. *International Journal of Science Education, 31*, 873–894.

Slater, S. J., Morrow, C. A., & Slater, T. F. (2008). *The impact of a kinesthetic astronomy curriculum on the content knowledge of at-risk students*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Slater, T. F. & Slater, S. J. (2009, August). *A science discipline based perspective on core idea*. Commissioned paper presented at the National Research Council Conference on Core Ideas in Science, Washington, DC. Retrieved from the National Academies website: http://www7. nationalacademies.org/bose/Slater_CommissionedPaper.pdf

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

Sneider, C., & Pulos, S. (1983). Children's cosmographies: Understanding the earth's shape and gravity. *Science Education, 67*, 205–221.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*, 610–631.

Taylor, I., Barker, M., & Jones, A. (2003). Promoting mental model building in astronomy education. *International Journal of Science Education, 25*, 1205–1225.
Treagust, D. F., & Smith, C. L. (1989). Secondary students' understanding of gravity and the motion of planets. *School Science and Mathematics, 89*, 380–391.
Trumper, R. (2006). Teaching future teachers basic astronomy concepts – seasonal change – at a time of reform in science. *Journal of Research in Science Teaching, 43*, 879–906.
Trundle, K. C., Atwood, R. K., & Christopher, J. E. (2007). Fourth-grade elementary students' conceptions of standards-based lunar concepts. *International Journal of Science Education, 29*, 595–616.
Trundle, K.C., Atwood, R.K., Christopher, J.E., & Sackes, M. (2010). The effect of guided inquiry-based instruction on middle school students' understanding of lunar concepts. *Research in Science Education, 40*, 451–478.
Tsai, C.-C., & Chang, C.-Y. (2005). The effects of instruction guided by the conflict map: Experimental study of learning about the causes of seasons. *Journal of Research in Science Teaching, 42*, 1089–1111.
Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.
Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science, 18*, 123–183.
Wilhelm, J. (2009). Gender differences in lunar-related scientific and mathematical understandings. *International Journal of Science Education, 31*, 2105–2122.
Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716–730.

CHRISTINA SCHWARZ, BRIAN J. REISER, ANDRÉS ACHER, LISA KENYON AND DAVID FORTUS

MoDeLS

Challenges in Defining a Learning Progression for Scientific Modeling

The MoDeLS project, *Modeling Designs for Learning Science,* has been developing and refining a learning progression that represents successively more sophisticated levels of engagement in the practice of scientific modeling (Schwarz et al., 2009). Our view of modeling practice draws on areas of agreement in current studies of learning about modeling (Harrison & Treagust, 2000; Lehrer & Schauble, 2000, 2006; Lesh & Doerr, 2003; Lesh & Lehrer, 2003; Treagust, Chittleborough, & Mamiala, 2002). We define a scientific model as an abstract, simplified representation of a system that makes its central features explicit and visible and that can be used to generate explanations and predictions (Harrison & Treagust, 2000). Examples of different kinds of scientific models include the Bohr model of the atom, the particle model of matter, a light ray model for how we see objects, the water cycle model, and a food web model indicating interactions between organisms. Working with scientific models involves constructing and using models as well as evaluating and revising them. The goal of this practice is to develop a model consistent with theoretical and empirical evidence that can be used to explain and predict multiple phenomena.

Developing and using models is central to authentic scientific practice. Involving learners in the practice of scientific modeling can help them construct subject matter knowledge, epistemological understanding, and expertise in building and evaluating scientific ideas (Lehrer & Schauble, 2006; Lesh & Doerr, 2000; Schwarz & White, 2005; Stewart, Cartier, & Passmore, 2005). The opportunity to engage in scientific modeling is important for developing and evaluating explanations of the natural world. Scientific modeling, however, is rarely incorporated into the educational experiences of elementary or middle school students. When modeling is part of school experiences, it is often reserved for secondary students and is primarily used for illustrative or communicative purposes, thus limiting the epistemic richness of the scientific practice (Windschitl, Thompson, & Braaten, 2008).

Our goal is to develop a learning progression that characterizes the aspects of modeling that can be made accessible and meaningful for students and teachers in upper elementary and middle school classrooms – ideally, a learning progression that can be used across multiple science topics and can support development of the practice across multiple years of learning. Consistent with other chapters in this

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 101–137.

book, we view a learning progression as a framework for articulating successively more sophisticated versions of knowledge and practice that is built on the understandings and ways of knowing that learners bring to the classroom (National Research Council [NRC], 2007; Smith, Wiser, Anderson, & Krajcik, 2006). Learning progressions offer the opportunity to explore and characterize paths through which students can build their knowledge and practices over time across a variety of important contexts such as different curriculum materials and classroom environments. Learning progressions are useful for designing effective instructional materials, designing formative and summative assessments, and supporting instruction that can help learners meaningfully engage with science ideas and practices over time. We do not assume that students become more sophisticated at engaging in the practice of scientific modeling in a particular fixed sequence. Progress may take different paths as students build sophistication with respect to the various elements of modeling practice. We expect that learners' enactment of modeling practice is critically dependent on instruction and scaffolding (Lehrer & Schauble, 2006; NRC, 2007). Thus we do not claim that the elements of modeling practice we describe are context-independent.

We have chosen to foreground the scientific practice of modeling in our learning progression. While we recognize the content-dependent nature of scientific practices, our research project has chosen (1) to focus on an important scientific practice that, with some exceptions (e.g., Lehrer & Schauble, 2000), is not typically highlighted in most elementary and middle school classrooms, (2) to determine whether students can abstract aspects of the practice across science topics over time, and (3) to conduct research in elementary and middle school science contexts that typically include teaching multiple science topics (in physical, life, and earth science) each year.

We developed our learning progression through an iterative process involving theoretical and empirical work. We began with consideration of prior theoretical analyses, empirical investigations, and work in philosophy of science; subsequently, our work has been guided by our empirical research from classroom enactments of modeling-oriented curriculum materials and assessments in upper elementary and middle school classrooms (Schwarz et al., 2009). This iterative process of designing a learning progression started with defining an initial framework in conjunction with designing curriculum materials and assessments used in the first year of the classroom enactments. The framework was then fleshed out with our initial empirical data and revised to become our initial learning progression (Schwarz et al., 2009). We subsequently refined our curriculum materials and assessments for classroom enactments in the second year and used the outcomes of these enactments to further revise our learning progression.

There are many challenges associated with building a learning progression for a complex practice such as scientific modeling. The development of an empirically-supported learning progression for a scientific practice is a paradigmatic example of research problems suited to design research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Edelson, 2001). Characterizing and comparing possible paths for learning a particular target concept are not solely matters of empirical investigation. Defining the target of learning involves design considerations. The goal in bringing scientific practices into classrooms is not simply to replicate professional practices; rather, it requires exploring which aspects of a given practice are both feasible and productive for learners (Edelson & Reiser, 2006). Investigations of pathways through which students can develop a scientific practice require empirical explorations of challenges and successes in reaching particular learning goals and entail developing design arguments for the learning goals. Design research is also essential because the target practice rarely exists in typical classrooms; investigations of the practice require design decisions aimed at creating the conditions that can support it (Cobb et al., 2003).

In this chapter, we explore design challenges that emerged in attempting to define, investigate, and revise a learning progression for scientific modeling. We consider issues in *defining* the progression, including which elements of modeling practice are critical for learners, which dimensions of the practice should be represented in a learning progression, and what grain size is needed to capture change. We also consider issues involved in *investigating* the learning progression such as designing effective curriculum and professional development materials to support teachers and students in their enactments of the practice. Investigating the learning progression also requires effective assessments across multiple science topics and appropriate analytical tools for interpreting outcomes. Finally, *revising* the learning progression requires analyzing students' work as they engage in modeling practices. We outline the challenges we faced in our process of defining, investigating, and revising our progression. We also present the learning progression we developed through this iterative design process.

CHALLENGES IN DEFINING A LEARNING PROGRESSION FOR THE PRACTICE OF SCIENTIFIC MODELING

Challenges in Defining the Aspects of Scientific Modeling for Classrooms

Scientific modeling is a rich practice that overlaps with other practices such as conducting investigations, constructing explanations, and engaging in argumentation (Passmore & Svoboda, 2011). Educators could choose to involve learners in a variety of candidate forms of scientific modeling. Hence one challenge in designing a learning progression for modeling is to identify which forms of modeling (with their underlying knowledge) are appropriate and productive for learners in school science settings. In particular, what aspects of modeling can (and should) learners understand and be able to do? Research on bringing scientific practices into classrooms has focused on practices such as argumentation, explanation, and scientific modeling. This work contains common elements, such as comparing alternatives, community building of knowledge involving argumentation and consensus, and evaluating scientific ideas against evidence. However, the research differs in the specific analytical frameworks used to characterize the target practice. Some modeling approaches focus on models as embodying patterns in data (Lehrer & Schauble, 2000) while others discuss models as embodying causal explanatory mechanisms for phenomena (White, 1993;

Windschitl et al., 2008). There are also open questions about conceptualizing learning goals such as whether to prioritize knowledge about the practice (Lederman, 2007) or the practical work of knowledge building through use of the practice (Sandoval, 2005).

To address the challenge of deciding which elements of the practice are suitable and productive for learners, we made several design decisions that influenced our learning progression. These included decisions about highlighting particular forms of models and elements of modeling practice, balancing metaknowledge and practice, and studying the practice across science topics. Our design decisions were based on prior research (Carey & Smith, 1993; Lehrer & Schauble, 2000; Schwarz & White, 2005; Stewart et al., 2005), theoretical arguments for what is most productive for learners (Schwarz, 2002), and contextual constraints such as what is possible in classrooms with existing curriculum materials and teachers.

Highlighting particular forms of models. We chose to design a learning progression that engages learners in modeling components, processes, and mechanisms that can explain and predict phenomena. The focus of modeling in our learning progression contrasts somewhat with a focus on modeling data or patterns in data (e.g., Lehrer & Schauble, 2000). Data modeling is central in science (e.g., much of modern science involves data and computational modeling) and enables a focus on the mathematical and representational aspects of making sense of phenomena. We chose a complementary focus – namely models that embody some aspects of causal and often non-visible mechanisms or explanatory components (Clement, 2000; White, 1993). Like data modeling, this type of modeling involves learners in creating, debating, and reaching consensus on inscriptions that represent their thinking about scientific phenomena (Lehrer & Schauble, 2004, 2006). We focus on explanatory models to emphasize accounting for patterns in phenomena by proposing theoretical explanatory constructs, such as processes and mechanisms, which are a critical part of building knowledge in science (Russ, Scherr, Hammer, & Mikeska, 2008). Examples of these explanatory constructs include (a) using reflection, absorption, and transmission of light rays as mechanisms to explain shadows, reflection, color, and other phenomena and (b) using the presence and movement of particles of matter as mechanisms to explain phase changes, diffusion, and related chemical phenomena.

Targeting explanatory constructs and mechanisms in students' models is certainly ambitious. Students' ideas about causes of scientific phenomena are typically under-developed. At times their ideas are consistent with their perceptions from everyday experiences but inconsistent with scientific data and canonical explanations. However, modeling allows students to externalize and reflect on evidence and experiences and to relate them to possible mechanisms and additional scientific information. In this way, students can move toward higher-level explanations and predictions of phenomena.

Selecting elements of modeling practice. Our learning progression reflects the commitment that learners need to engage in the modeling practice itself: embodying key aspects of a theory and evidence in an expressed representation; using the representation to illustrate, predict, and explain phenomena; and evaluating and revising the representation as it is used. While teachers may provide learners with models for observing aspects of phenomena or have learners construct models to apply what they have been taught, engaging learners in the evaluation and revision of models is rare. Yet this entire model construction and revision process—also called model evolution (Clement, 2008) or model refinement (Acher, Arcà, & Sanmartí, 2007)—is critical for a better understanding of and participation in science. Students need the opportunity to evaluate and revise their models if they are to understand the relationship between models and data, as well as the social aspects of the modeling process (Clement, 2008; White & Frederiksen, 1998). As students engage in elements of modeling practice, they must attend to the role of empirical evidence in constructing, evaluating, and revising models. The social context is critical to motivating and supporting the evaluation and revision of models since the generation of competing alternative models creates a need for criteria to evaluate the strengths and weaknesses of candidate models. In addition, students use the audience of their classroom peers to test the effectiveness of their models for persuading others and helping them to understand targeted scientific ideas.

Drawing on prior work related to epistemologies and the nature of science (Carey & Smith, 1993) and on student learning about modeling (Grosslight, Unger, Jay, & Smith, 1991; Snir, Smith, & Raz, 2003; Spitulnik, Krajcik, & Soloway, 1999; Stewart et al., 2005), we operationalized the practice of modeling to include four elements:

- *Constructing* models consistent with prior evidence and theories to illustrate, explain, and predict phenomena;
- *Using* models to illustrate, explain, and predict phenomena;
- Comparing and *evaluating* the ability of different models to accurately represent and account for patterns in phenomena and to predict new phenomena; and
- *Revising* models to increase their explanatory and predictive power, taking into account additional evidence or aspects of phenomena.

These four elements represent modeling tasks that can be differentiated by their goals and guiding criteria, but they are not discrete steps performed in sequence. For example, students may compare and evaluate candidate models as they attempt to revise their current model. These four elements provide the foundation for the modeling practice in our learning progression.

Balancing and integrating metaknowledge and practice. Our goal is to engage learners in reflective practice in which scientific activity is meaningful to them (Edelson & Reiser, 2006; Lehrer & Schauble, 2006; Lehrer, Schauble, & Lucas, 2008). Achieving this goal requires supporting both the doing and the understanding of the practice (metaknowledge). To guide instruction and assessment, the learning progression must articulate learning goals that involve both performance of the elements of the practice and the associated metaknowledge. An important design tradeoff emerges when selecting and integrating metaknowledge into a practice. On the one hand, we do not want to teach modeling as a scripted routine in which students perform steps simply because they were encouraged to do so in instruction. In contrast to learned routines, engagement in a practice is governed by shared understandings, norms, and goals – the form of practice is meaningful to the community engaged in that practice (Bielaczyc & Collins, 1999; Brown & Campione, 1996). This motivates the need for including elements of metaknowledge as part of modelling instruction. The underlying epistemological understanding explains why the practice takes the form and develops the way it does. For example, constructing scientific explanations is more sensible when learners understand how everyday explanations differ from scientific explanations; while plausibility is a sufficient criterion for everyday explanations, scientific explanations must also be consistent with empirical evidence (Brewer, Chinn, & Samarapungavan, 1998). This argument suggests that understanding how and why models are used, as well as their strengths and limitations, may help learners construct, use, evaluate, and revise their own and others' models (Schwarz, 2002; Schwarz & White, 2005; Schwarz et al., 2009).

On the other hand, there are many ways of exploring how science works and how scientists think that challenge learners' prior understandings (Abd-El-Khalick et al., 2004; Lederman, 2007). Making modeling practice meaningful is our foremost goal. Thus we do not want to target knowledge about the nature of science as a learning goal for its own sake. Teaching nature of science ideas without embedding them in practice risks them becoming decontextualized knowledge that is not grounded in students' own experience of engaging in the practice. Thus there is a tension between including those elements of metaknowledge that can help make the practice meaningful (rather than a scripted routine) and including metaknowledge that may become a set of decontextualized facts simply to be learned.

Our design strategy involves a pragmatic instrumental approach to metaknowledge in which we focus on the metaknowledge that is useful in helping learners resolve the obstacles they encounter (Sandoval, 2005). This is a learning-for-use argument in which the utility of scientific ideas arises as they are introduced to help learners solve problems they are investigating (Edelson, 2001; Kanter, 2010). In this way, learners construct scientific ideas as tools for solving meaningful problems, rather than ideas to be learned for their own sake. We adapt this idea by situating the teaching of the metaknowledge in modeling problems that the metaknowledge can help guide. Therefore, our learning progression specifies the elements of metaknowledge that we theorize influence the elements of modeling practice. We look for and support growth of metaknowledge as it applies to the performance of elements of modeling practice.

Thus in articulating metaknowledge in practice (*metamodeling knowledge*; Schwarz & White, 2005), we focus on elements that help students make modeling decisions. For example, when comparing competing ideas to develop group consensus, knowledge about criteria for evaluating models, such as fit with empirical evidence about phenomena and coherence of explanation, is needed to tackle the modeling work. Other metaknowledge may be possible to target—such as the understanding that models can take different forms; knowlege of relationships among and differences between models, theories, and laws; and more general ideas about scientific disciplines involving "creativity" (Lederman, 2007)—but these ideas have less clear utility in guiding students' construction, comparison, evaluation, and revision of models.

Conceptualizing modeling as a general scientific practice. A related issue in balancing the performance of elements of practice and metaknowledge is the level of abstraction of the reflective practice targeted for instruction and hypothesized to improve with experience. For example, to what extent can students use the same reflective practice in constructing, using, evaluating, and refining models about heredity, the nature of matter, and ecosystems dynamics? The influence that specific scientific disciplines and types of investigations have on learning scientific practices is critical (Lehrer & Schauble, 2006; Tabak & Reiser, 2008). Engaging in a practice is always situated in particular disciplinary investigations that vary along important dimensions. For example, designing experimental investigations (such as studies of the relationship between force and motion of physical objects) is different from designing investigations that require the analysis of naturally occurring datasets to determine relationships between variables (such as longitudinal studies of species interactions in ecosystems). This difference suggests that the approach taken in defining learning progressions could focus on the combination of a practice and a domain, such as scientific modeling in physical science.

We view these discipline-specific practices as important specialized forms of more general practices, such as the scientific evaluation of the relationships between two variables. Thus, for both instruction and analysis, we target general aspects of constructing mechanistic models that cut across specific scientific disciplines – e.g., evaluating fit with evidence, focusing models on the salient details that support explanations, and generalizing the model to explain a range of cases. This focus allows us to explore how the practice can become more sophisticated across a range of scientific ideas. To determine whether the reflective practice of modeling transfers from one setting to another, we examined whether that practice could be applied across a wide range of scientific phenomena.

These decisions to integrate performance of elements of practice and metaknowledge across multiple science topics have important implications for defining, investigating, and revising the learning progression. In the following sections we discuss the challenges that arose in making these decisions.

Challenges in Defining an Initial Learning Progression for Scientific Modeling

We represented our initial learning progression for scientific modeling as a set of related construct maps (Wilson, 2005, 2009), each of which represents a trajectory of the elements of this practice and associated understandings that we expect students to exhibit in classroom modeling activities. Each level in our construct maps represents a more sophisticated version of the previous level; thus these construct maps represent the theoretical articulation of the elements of performance and understanding of the practice that become more sophisticated with experience and instruction. The construct maps then guided the design of supports for learning and analyses of student growth over time. We used these construct maps to develop more specific rubrics for analyzing a range of data from classrooms. Our data included student interviews, written assessments, and classroom discourse in small group work and in whole class discussions. Analyses from multiple enactments helped refine the theoretical articulation of the construct maps.

One important challenge in defining our initial learning progression involved the dimensions of the learning progression (i.e., which construct maps to include) and the grain size for each dimension. We initially considered creating construct maps that represent the combination of each of the four elements of practice and several relevant elements of metamodeling knowledge (such as the purpose and nature of models and the criteria for evaluating models). However, there would have been overlapping elements of practice and similar metamodeling knowledge in these maps. For example, the use of models to explain phenomena motivates both the construction and the use of models. The fit of the model to evidence applies to both the evaluation and the revision of the model. Furthermore, the elements of the practice refer to tasks that may overlap. Constructing models and using models to explain phenomena may be separate steps, for example, when using an existing model to explain a new although similar phenomenon. However, constructing models and using models can also be coordinated tasks since considerations of whether a model sufficiently explains the data about a phenomenon should guide decisions in constructing that model. Similarly, evaluating a model's strengths and weaknesses and revising that model may either be two steps in the same task or two separate tasks.

Thus, rather than creating separate construct maps for each element of the practice combined with associated metamodeling knowledge, we identified two clusters of issues that synthesize metamodeling knowledge and that influence the four elements of practice. Those clusters were represented by the two dimensions of our initial learning progression and are illustrated by two construct maps that describe these aspects of reflective practice. Thus each construct map describes aspects of students' modeling performances (such as their decisions about revising models, properties of their constructed models, or changes in a revised model) and their reasoning about these performances (as reflected in their classroom discourse or written explanations). The *generative* dimension (Table 1) describes the reflective practice for how models predict or explain aspects of phenomena when models are constructed or used. Each target is defined as a combination of students' performance of a modeling task (both process and product) and the underlying metaknowledge that makes the activity meaningful. The *dynamic* dimension (Appendix) describes reflective practice for when and how models need to change when students evaluate and revise them (see Schwarz et al., 2009 for additional details about our initial learning progression). These two dimensions provided the initial framework that guided our empirical investigation of modeling practice in the classrooms.

In our initial generation of the construct maps, we defined levels associated with significant, rather than incremental, differences in reflective practice, resulting in broad descriptions. These levels were based on prior theoretical and empirical work (e.g., Carey & Smith, 1993; Schwarz & White, 2005) that outlined broad categories of meaningful change. Since we were uncertain about what we might detect in students' work as they engaged in modeling practice, we developed broad descriptive levels. Our intent was to revise the level descriptions in response to empirical data. As a result, differences between levels in our construct maps represent significant shifts in reflective practice based on different epistemological notions of modeling. We included several levels, or general descriptions of reflective practice, that are qualitatively different and are associated with aspects of practice that might be valuable in a scientific context. The first level of each construct map embodied the initial description of modeling practice – envisioned as engaged in by learners in early elementary school or by learners with no prior modeling experiences. The final (fourth) level was associated with advanced ideas about and sophisticated use of modeling – envisioned as the practice of experienced college science students, graduate students, or scientists – and helped us outline the possible levels of modeling practice. We also included two intermediate levels: a second level for learners who have moved beyond a beginning phase but still engage in epistemologically limited modeling practice and a third level for learners who have shifted toward more sophisticated understanding and modeling practice.

Table 1. A Construct Map for the Generative Dimension: Understanding Models as Generative Tools for Predicting and Explaining.

Level	*Descriptions of Reflective Practice (Including Performance of Elements of the Practice and Associated Metaknowledge)*
4	Students *construct and use models, extending them* to a range of domains to help their own thinking. Students *consider how the world could behave according to various models*. Students construct and use models to generate new questions about the behavior or existence of phenomena.
3	*Students construct and use multiple models to explain and predict additional aspects of a group of related phenomena*. Students view models as tools that can support their thinking about existing and new phenomena. Students consider alternatives in constructing models based on analyses of the *different advantages and weakness* for explaining and predicting these alternative models possess.
2	*Students construct and use a model to illustrate and explain how a phenomenon occurs*, consistent with the evidence about the phenomenon. Students view models as a *means of communicating their understanding of a phenomenon* rather than as tools to support their own thinking.
1	*Students construct and use models that show literal illustrations of a single phenomenon*. Students do not view models as tools to generate new knowledge but do see models as *a means of showing others what the phenomenon looks like*.

Note. Adapted from "Developing a Learning Progression for Scientific Modeling: Making Scientific Modeling Accessible and Meaningful for Learners," by C. V. Schwarz, B. Reiser, E. A. Davis, L. Kenyon, A. Acher, D. Fortus, Y. Shwartz, B. Hug, and J. Krajcik, 2009, Journal for Research in Science Teaching, 46, p. 640. Copyright 2009 by Wiley Periodicals, Inc. Reproduced with permission of Wiley Periodicals, Inc. via Copyright Clearance Center.

Consider the construct map for the generative dimension, shown in Table 1. This dimension describes how students construct and use models that embody explanatory constructs (e.g., mechanisms, processes) and whether students view models as useful for advancing their own knowledge as well as for helping communicate what has been learned to others. One challenge we faced in constructing this map was aggregating aspects of the practice that generally represent similar epistemological ideas while addressing different decisions surrounding the practice, such as the audience or the type of model constructed or evaluated. We aggregated several aspects under each level based on preliminary classroom data that suggested aspects that either clustered or seemed to play an important role at a particular level.

Each of the four levels shown in Table 1 is defined by two related descriptions. The first description is intended to capture the sophistication of students' construction and use of models and the associated metaknowledge that guides their decisions about the model components and the relationships between these components. The second description focuses on students' understanding of the reasons for constructing and using models. All indicator statements include performance of the elements of the practice guided by associated metaknowledge. We next describe these four levels using examples of student work from classroom enactments in which modeling curriculum materials were used.

We designed the Level 1 descriptions to capture students' reflective practice characterized by considering modeling and the purpose of models to be literal illustrations of a particular phenomenon. For example, Figure 1 shows a fifth grade student's pre-instructional model of water in a covered and uncovered cup. The student drew a model that shows the water level before and after a period of time but that includes no explanatory components. Students at Level 1 make their models as similar to the real thing as possible. For example, a fifth grade student commented (after an introductory modeling unit): "A model would [be] like an actual Coke can with water on the side, or a picture of it, that is more detailed and colored …"

We designed the Level 2 descriptions to capture students' reflective practice characterized by (1) constructing and using models to explain how phenomena occur and (2) making their models consistent with evidence about the phenomena. This level is more sophisticated because students construct and use models that include non-observable processes and mechanisms. Students at this level also consider sources of information such as empirical evidence or information from teachers and books. However, students at this level do not yet view models as tools to support their thinking. After the modeling unit, the same fifth grade student (who drew the model in Figure 1) drew the model in Figure 2. In Figure 2 the student drew microscopic particles in a "zoomed-in" fashion to illustrate how the water condenses and escapes. The student also drew arrows that indicate a process and direction for the evaporation and labels that show the water level has changed in the uncovered cup (i.e., a change over time). Students at this level justify changes to their models in general terms of the model's ability to explain phenomena. For example, this fifth grade student commented after the modeling unit: "The models are helpful because they explain how evaporation and condensation works." This example illustrates how a student's level in the construct map can change after instruction.

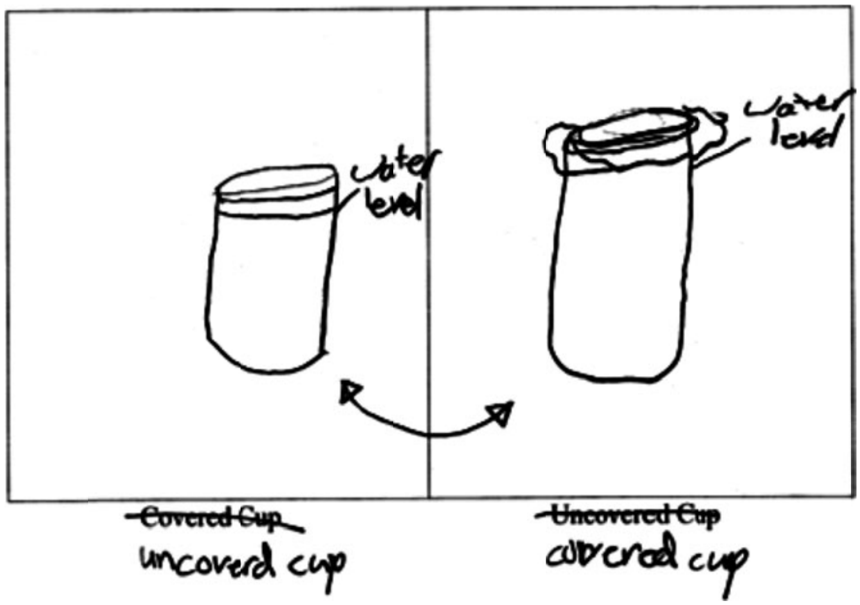


Figure 1. A fifth grade student's pre-instruction model to show what happens to water level in a covered cup and in an uncovered cup (Level 1 on the generative construct map).

We designed the Level 3 descriptions to capture students' reflective practice characterized by using models not only to explain phenomena they observe but also to make predictions about new phenomena. At this level, students view models as thinking tools. Since students' models explain a cluster of related phenomena, multiple models may be required. Students may consider alternatives when constructing models and choosing model components. We see some aspects of Level 3 reflective practice when students apply explanatory models to predict other phenomena that go beyond the situations they used to construct them. Consider this excerpt from an interview with a sixth grade student that occurred during a chemistry modeling unit:

> **Interviewer**: Could a model like the models you have done help you to understand something that you don't know yet? Just to predict?
>
> **Student**: Yes, I think that we could use these models to show about a different gas. Because I think air is also a type of gas. So if we know what air does and has in it and how it moves when other things come in its way and everything, we could find out what other gases also do because they could be similar to air.

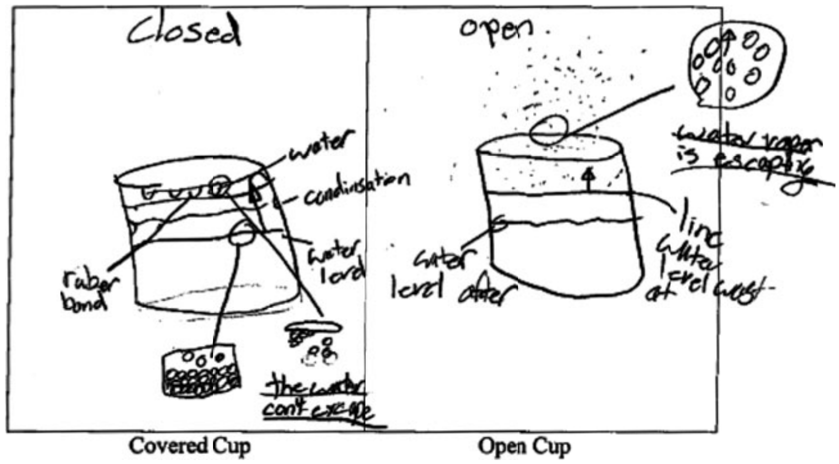


Figure 2. The same fifth grade student's post-instruction model of water in a covered cup and in an uncovered cup (Level 2 on the generative construct map).

We designed the Level 4 descriptions to capture students' reflective practice characterized by constructing and using models that extend to a range of domains in order to advance their thinking. Students consider how the world *could* behave according to various models, and they construct and use models to generate new questions about phenomena. We have found no examples of this type of reflective practice in our elementary and middle school classroom enactments. However, using models to advance scientific knowledge by generating questions to guide research is fundamental to knowledge building in science (Carey & Smith, 1993; Lederman, 2007; Lehrer & Schauble, 2006) and forms an important placeholder for envisioning more advanced forms of modeling practice. Examples of such reflective practice have been shown to exist in other settings (Fortus, Rosenfeld, & Shwartz, 2010; Smith, Maclin, Houghton, & Hennessey, 2000).

CHALLENGES IN INVESTIGATING THE LEARNING PROGRESSION

We engaged in an iterative process of investigating and revising our initial learning progression with data from multiple classroom enactments as part of an empirical validation process (Mohan, Chen, & Anderson, 2009). We analyzed classroom enactments that occurred over two years in fourth, fifth, and sixth grades in three Midwest states, using scientific modeling curricula we developed or adapted to support teachers' and students' engagement in the practice. We collected and analyzed evidence of students' reflective practice from written pre-post assessments, videos and written artifacts from classroom enactments, and student interviews.

Scientific modeling can be complex and difficult for teachers and students (Justi & van Driel, 2005; Schwarz et al., 2009; van Driel & Verloop, 1999; Windschitl et al., 2008). Teachers frequently struggle with understanding what constitutes a scientific model, how to productively incorporate modeling into the curriculum, and how scientific modeling overlaps with other scientific practices emphasized in state and national standards, such as scientific inquiry and generating explanations. Consequently there have been several major challenges in our use of classroom enactments to generate evidence to investigate our learning progression. In particular, we faced challenges (1) in designing and adapting curriculum materials to investigate a learning progression for a practice, (2) in working with teachers to better understand and incorporate scientific modeling in their teaching, and (3) in designing assessments that adequately assess students' engagement in the practice.

Challenges in Designing and Adapting Curriculum Materials

The investigation of learners' engagement in a scientific practice and of the elements of performance and understanding that become more sophisticated over time requires sustained opportunities to engage in the practice with targeted support. Unfortunately, students rarely have the opportunity to engage in knowledge-building scientific practices such as explanation, modeling, and argumentation (NRC, 2007; Weiss, Pasley, Smith, Banilower, & Heck, 2003). Thus investigating a learning progression for a scientific practice requires a research program that studies what is possible for learners to accomplish with support, rather than one that studies what "naturally" emerges as learners participate in typical instruction and out-of-school experiences. Learning progressions for scientific practices do not describe developmentally inevitable stages; rather, they describe what learners can accomplish with suitable learning opportunities for engaging in a given practice and with appropriate support for engagement and reflection. Thus, to investigate a learning progression for scientific modeling, we had to create suitable instructional contexts. We had to make several choices in this effort: the selection of scientific topics in which to embed the practice; the decision of whether to design or to adapt the curriculum; and the identification of the practice elements to highlight.

Selecting scientific topics for modeling. As discussed earlier, we define modeling, as other researchers do (Harrison & Treagust, 2000; Lehrer & Schauble, 2000, 2006; Lesh & Doerr, 2003; Lesh & Lehrer, 2003; Treagust et al., 2002), as an abstract, simplified representation of scientific phenomena that makes the central features of the phenomena explicit and that can be used to generate explanations and predictions (Harrison & Treagust, 2000). Other work has explored how to help students develop explanations for phenomena and support those explanations with scientific arguments (Berland & Reiser, 2009, 2011; McNeill, 2009; McNeill, Lizotte, Krajcik, & Marx, 2006). We build on this prior work with practices related to modeling by adding the notion of models as explanations of phenomena that we can then apply in multiple contexts. Our instructional approach stresses the goal of representing, comparing, evaluating, and reaching consensus on explanations of phenomena as models. We also emphasize the usefulness of constructing explicit external representations of models (such as diagrams of force and motion, food webs, atomic and molecular structure, and so on).

This emphasis within scientific modeling has implications for the choice of topics that can be investigated with modeling practices. Some scientific topics are more suitable than others for making explanatory components and non-visible mechanisms of phenomena visible at a level consistent with grade-level state benchmarks and standards. Although a phenomenon may be observed and empirically investigated, it may be difficult to help students construct mechanistic understandings of that phenomenon from the evidence. For example, a unit on the function of plant parts that asks students to identify those parts and their general function, if redesigned to incorporate modeling practice, would need to model the non-visible components and mechanisms of plant growth. Such a redesign would be difficult as it would require helping younger students understand photosynthesis at the molecular level without building on a particle model of matter.

Therefore, we have focused on science topics for which students can construct and revise diagrammatic models of non-visible explanatory components over time. For middle school, we used units under development as part of the Investigating and Questioning our World through Science and Technology project (IQWST; Krajcik, McNeill, & Reiser, 2008; Shwartz, Weizman, Fortus, Krajcik, & Reiser, 2008). We embedded more explicit supports for modeling in two sixth grade IQWST units. Both units target explanatory models with non-visible components that describe processes occurring over time: a physics unit on light in which students model light rays travelling through space and their interaction with materials, and a chemistry unit about the nature of matter in which students model particles of gas and their movement. We have also extended this curriculum approach to an eighth grade unit on natural selection in which students develop evidence-based explanations of population change. They begin with two individual cases of natural selection, finches in the Galapagos Islands and peppered moths in the United Kingdom. The students then adapt their explanations of natural selection to construct a more general model that accounts for population change through differential survival in a population containing natural variation. We also created an elementary fifth grade unit on evaporation and condensation that introduces the notion of particles and their movement without addressing the full molecular model presented in the IQWST sixth grade unit (Baek, Schwarz, Chen, Hokayem, & Zhan, 2011). The fifth grade unit enabled us to consider the foundations of a more sophisticated model of matter developed in the sixth grade unit.

Deciding how to bring modeling practice into classrooms through curriculum materials. The lack of agreement among school districts and states about which science topics should be taught each year prevented us from developing new units for the multiple school districts in Illinois, Ohio, and Michigan that partnered with us. Thus, in addition to developing a modeling-focused unit on evaporation and condensation, we also developed a set of principles to help teachers embed the practice of scientific modeling in suitable existing curriculum units (Kenyon, Schwarz, & Hug, 2008). We worked with teachers to use these principles to modify commercially published and district-authored units. The topics of these units included form and function in the human body and the properties of electrical circuits.

Identifying elements of modeling practice to target. As described earlier, we selected constructing, using, evaluating, and revising models as the central elements of modeling practice. However, there are many potential approaches to engagement with these elements. We focused on contexts that emphasize comparative evaluation of models, leading to their revision, in order to highlight the nature of models as explanatory tools that can be improved with new evidence, rather than as "final form" answers to scientific questions. Therefore we developed problem contexts requiring iterative model construction and revision through metamodeling discussions as well as social contexts such as peer evaluation and consensus model-building (Kenyon et al., 2008; Schwarz et al., 2009). As the problem and social contexts were enacted somewhat differently for the various curricular materials and classrooms, there was likely an influence on the learning opportunities and engagement in scientific modeling. For example, some materials (and the teachers using them) engaged students in comparing and contrasting representations and their meanings in diagrammatic models, while others emphasized consistency with empirical evidence. Some approaches emphasized metamodeling discussions and language, while others primarily emphasized the construction and revision of models. Each context and approach affected outcomes that guided revisions of our learning progression.

Challenges in Supporting Teachers' Instruction of Scientific Modeling

An important set of challenges in empirical investigations of a learning progression for a practice arises when helping teachers to better understand the practice and incorporate it into their teaching. Most teachers perceive that the goals of science teaching, reinforced by the pressure of high-stakes assessment, require a focus on science content. These perceptions manifest themselves in several forms, sometimes reducing opportunities for engaging in the full range of targeted elements of modeling practice. Yet if students do not have adequate opportunities to engage in modeling, we have little evidence for evaluating and revising the construct maps that comprise our learning progression.

Viewing practice and content as competing goals. One instructional challenge emerged in the teachers' approach to the relationship between scientific practices and content. While we, as researchers and curriculum designers, see engaging in modeling practice as an effective way to develop deeper explanations of scientific phenomena (i.e., content), teachers may think scientific practices and scientific content represent separate and competing goals. Or teachers may lack the pedagogical strategies that can support students in using scientific practices to develop scientific ideas. Most teachers have never used models and modeling other than in a demonstration of a correct scientific idea (Justi & van Driel, 2005; Schwarz et al., 2009; van Driel & Verloop, 1999; Windschitl et al., 2008). Thus viewing modeling as a process that involves constructing, using, evaluating, and revising models as tools for advancing students' thinking may be counter to typical school, curriculum, and teacher norms. In our initial curriculum enactments, we found that some teachers involved in teaching the MoDeLS curriculum materials focused on the use of models to convey scientific content to their students. This focus was at the expense of engaging students more fully in the practice of modeling through repeated opportunities to work with partially correct models and to compare, evaluate and revise them.

Integrating metaknowledge into support of the practice. Another challenge for science instruction incorporating scientific practices arises in the integration of metaknowledge. This challenge stems from the tension between an integrated strand model of science literacy (NRC, 2007, 2009), in which practices are viewed as vehicles for the development and use of scientific knowledge, and the idea that process, content, and the nature of science are separate learning goals. Some teachers had difficulty incorporating the metamodeling aspect of the practice, particularly at appropriate and critical points in the curriculum, since this aspect added another layer of complexity. Other teachers ignored the metaknowledge aspect of modeling or used it as additional information to be learned (e.g., the definition of a model). Not surprisingly, it was also challenging for teachers to develop pedagogical strategies for both elements of the practice and the associated metamodeling knowledge – particularly in their first use of the materials. For example, asking students to engage in discussions that compare models to better understand important criteria for evaluation and using those discussions for improving models were not straightforward endeavors. Social and school norms (e.g., giving the teacher the correct answer or favoring the model presented by a more likable classmate) conflicted with the norms of the scientific practice. These conflicts posed difficulties for highlighting metamodeling knowledge that should guide performance of the elements of the practice, such as the consistency of the model with empirical evidence or the utility of a model in explaining and predicting phenomena. These difficulties influenced students' learning outcomes with respect to the learning progression. For example, if the teacher and school norms promote using models to "show the correct science answer," then a learner might have more difficulty exhibiting the higher levels of reflective practice in the generative construct map that include the idea that models can be tools for thinking about the world.

Supporting teachers through professional development and curriculum design. To address these challenges, we focused our professional development and curriculum design efforts on achieving a more seamless integration of the metaknowledge guiding the elements of modeling practice with the practice of constructing and using models to explain phenomena. We studied our curriculum approach in several different classroom settings, provided teachers with professional development, and modified the curriculum materials for use in additional classrooms. Some of our enactments were with teachers who followed an entire modeling unit; other enactments were with teachers who integrated modeling activities into their existing classroom curricula and engaged students in modeling practices in several contexts. Teachers who engaged their students in modeling practice over several units, and sometimes over multiple years, with project support, reported finding a more productive balance between content and reflective practice in their pedagogical strategies. In addition, teachers' feedback and our analysis of their enactments provided information for use in modifying the curriculum materials. As a result, we revised the elementary materials to support teachers and learners in more effectively relating empirical evidence to model revision, in clarifying the generative purpose of models, and in choosing more effective evaluation criteria and structures to support productive social norms for peer model evaluation. We modified the middle school materials to streamline the metamodeling discussions that may have been overwhelming to some teachers. We are currently studying teacher enactment and changes in student's modeling practice to determine how teacher-student interactions may have affected student-learning outcomes.

Challenges in Obtaining Evidence of Students' Reflective Practice

Designing appropriate written assessments and interview protocols that can be compared across grade levels and topics to inform redesign of the learning progression has revealed other research challenges: how to assess reflective practice that involves both engagement in the practice and understanding of the practice; how to provide scaffolding in assessments; and how to tease apart content learning gains from improved sophistication in the practice.

Analyzing modeling performance and metaknowledge in assessments. We developed our learning progression with a commitment to the integration of metamodeling knowledge with the elements of the practice (Kenyon et al., 2008; Schwarz et al., 2009). The target is reflective practice in which learners engage in the practice with understanding, not simply as a routine. For example, teachers and curriculum resources should help students see the need for labeling from the perspective of the audience that is trying to understand the phenomena, rather than stressing the importance of labeling diagrams and encouraging students to label their models, essentially as an end in itself. Thus, by itself, increased labeling is not necessarily evidence of reflective practice. We did not want to teach and then assess only the performance of aspects of the practice (with little rationale) or only metamodeling knowledge (with no engagement in the practice). Our construct maps focused not only on students' actions (such as labeling) but also on their rationales for choosing particular actions.

In order to assess both students' performances of modeling practice and their rationales for these performances, we wanted to observe their spontaneous justifications for modeling tasks. We also wanted to observe whether they answered questions about metaknowledge in terms of their specific decisions in performing the practice. Thus we used questions in both written assessments and interviews that varied in their focus on the performance of elements of the practice, on metamodeling knowledge, and on the connection between the two as vehicles to analyze the connection between the performance of elements of the practice and metaknowledge. For example, we developed tasks that asked students to perform an element of modeling practice, such as making a prediction using a model: "Use your model [of evaporation] to predict what will happen when a marker dries out." We asked other questions that focused on metaknowledge, such as "How are models useful?" For both types of assessment item, we looked for evidence of metaknowledge as applied to engaging in the practice according to our construct maps (Table 1, Appendix). The metaknowledge questions were sometimes useful in eliciting students' ideas that they rarely mentioned when working on modeling tasks (such as definitions of models). In general, however, we found that the questions that addressed both performance and knowledge of modeling practice were the richest ones for generating examples of student reflective practice. Thus most tasks explicitly required practice with metaknowledge as justification (e.g., "construct a model to explain this phenomenon, and discuss how it is useful for explaining what is going on.")

One approach that merges practice and metaknowledge is *grounded assessment* – interviews or written assessments that are contextualized in students' classroom work. For example, after students had revised their models of the movement of odors in a room based on experiments with air, we asked them how their revised models differed from their previous models and why the revised models were better. Contextualized questions such as these can be more productive in eliciting students' ideas than more abstract questions such as "why do you (or why does a scientist) revise models?" Students can draw on their experiences to justify why their revised models are better. Our scoring requires the integration of performance and metaknowledge – students who understand the movement of the particles and can build a better model but lack a sophisticated understanding of why their new model is better (other than "it's more correct") are distinguishable from students who can connect their modeling decisions to the importance of features such as the generality of the model, its utility in making new predictions, or the presence of a more complete causal mechanism.

Providing varying degrees of scaffolding in assessments. Another important challenge in designing written items or interview questions is how much scaffolding should be provided to support students' responses. Even in grounded assessments, students may not justify their decisions explicitly without prompting. Indeed, it may be challenging for students to give rationales for their performance of elements of modeling practice even when directly questioned in interviews. Students are much more familiar with questions targeting particular scientific content knowledge than with questions asking for a rationale for their decisions about a scientific practice. Interview questions about what the students considered in the construction of their models may not elicit all their construction decisions. Because our learning progression is about reflective practice, which comprises both doing and understanding modeling practice, information about only one of these aspects is insufficient to firmly locate students' at particular levels.

However, specific prompting about decision rationales has disadvantages as well. For example, following an open-ended interview question with the question "Did you use any of the information from the scientific experiments in constructing the model and if so, how?" prompts students to think about the relationship between the model and the empirical data, even if that consideration was not part of their original decision-making. Thus there are trade-offs between the more open-ended question (with a wider range of possible responses) and the more targeted question that may over-represent what students thought about in performing the practice. The best strategy is to attempt to triangulate across students' responses to written assessments and reflective grounded interviews, the rationales they provide in classroom discourse, and improvements that appear in their modeling work over time (Baek & Schwarz, 2011; Zhan, Baek, Chen, & Schwarz, 2011).

Teasing apart content and practice. A central challenge in assessment design for a scientific practice concerns the relationship between content and practice. There are several key issues here. First, it is important to differentiate improvement in the substance of students' models from more general improvement in modeling practice. For example, as students learn more about the particle nature of matter, they shift from representing matter as continuous to representing matter as particles with empty space between them (Smith et al., 2006; Stevens, Delgado, & Krajcik, 2010). These models are indeed more sophisticated and reflect important learning. However, improvements in the model may not be linked to improvements in the modeling practice; rather, students may have learned more about what matter is made of. Thus it is important to distinguish learning about specific scientific explanations, such as learning about the particulate nature of matter, from learning something more general about modeling practice, such as the importance of representing process and mechanism in a model.

Second, it is possible that limitations in proficiency with the scientific content may lead to under-representation of students' growth in modeling. For example, perhaps students have clear ideas about the importance of providing a mechanism that accounts for changes over time, but they do not understand a particular phenomenon well enough to speculate on a mechanism that could explain their observations. This is a more difficult problem. One strategy is to code for the presence of a mechanism in students' models independently of whether there are scientific inaccuracies. However, this is only a partial solution because lack of candidate explanatory concepts and lack of confidence in their explanations may keep students from speculations about a mechanism in their models.

Our construct maps attempt to address these issues by looking for improved performance guided by appropriate rationales. Students should advance beyond saying "our new model is better because it has particles" to justifying how a model of particulate matter fits the evidence better. Thus, if the growth we uncover is apparent, not only in the substance of the model but also in the rationales students provide, we can attribute these improvements to increased understanding of modeling.

The use of grounded interview and assessment questions helps tease out which model improvements seem linked to improvements in the reflective practice of modeling. For example, many elementary students move from literally depicting an open container that lost water through evaporation to representing small, invisible particles of water moving into the air (Baek et al., 2011). When this improvement in modeling is accompanied by changes in students' views of what models do, there is evidence for more than simply a gain in content knowledge. For example, compare these two rationales (from two fifth grade students' post-unit interviews about models). One student commented that "a model would [be] like an actual Coke can with water on the side, or a picture of it, that is more detailed and colored …” In contrast, the second student’s response is more representative of those after the modeling unit since it focuses on the explanatory power of the model: “because it just doesn't show a picture or a diagram. … It doesn't just show it, it explains.” The second student appears to consider different criteria for what a model needs to do and offers these criteria as rationales for how the new model is an improvement. The grounded questions allow us to avoid students’ unfamiliarity with new content or the challenges of decontextualized questions that might not fully reveal what they can do. However, this strategy does not solve the problem of students’ answers on pretests that under-represent their knowledge about modeling if their unfamiliarity with content keeps them from speculating on possible explanatory mechanisms.

Another strategy we have explored is “neutral context” questions that are set in contexts outside the topic of study. For example, we asked students to describe the strengths and limitations of several models a scientist uses to explain what a plant needs to grow. These questions, combined with the other measures, allow us to distinguish learning about content from learning about the practice. If students can more effectively critique models of plant growth after a modeling unit on unrelated science content, such as light or the nature of matter, this improvement appears due to what they have learned about developing and evaluating models.

CHALLENGES IN ANALYZING STUDENTS’ WORK AS THEY ENGAGE IN MODELING PRACTICE TO REVISE THE LEARNING PROGRESSION

We are using findings from classroom enactments to test and revise our initial learning progression. We have attempted to triangulate findings from analyses of written assessments, student interviews, classroom artifacts, and classroom conversations to assess students’ level of modeling practice and to examine whether and how this level improves with more experience with the practice. Our analyses revealed some limitations in our original construct maps arising from our attempt to use only two dimensions to characterize modeling practice. In this section, we describe the challenges that emerged in our attempts to use the learning progression to analyze students’ work as they engage in modeling practice. We also discuss the strategies we used to revise the construct maps that comprise the learning progression in response to these analyses.

Challenges in Analyzing Students’ Work as They Engage in Modeling Practice

Students’ work matching indicators from different levels. Students’ work as they engaged in modeling practice sometimes appeared to be situated across levels, matching indicators from both levels. For example, we found that some students’ work matched at least one Level 2 descriptor since these students constructed and used models that included non-observable processes, mechanisms, or components (one aspect of Level 2 in the generative construct map). Yet some of the same students appeared to consider the model’s purpose as a veridical illustration of phenomena (e.g., including particles that can only be seen with a powerful microscope), an indicator at Level 1. We found other difficulties in aligning indicators across levels. For some students who used non-observable processes or mechanisms in their models (Level 2), the use of evidence to support their models was not explicit in either their models or their associated rationales, which is an important factor in distinguishing between Levels 2 and 3 of the generative construct map.

In general, the level of students' responses to modeling tasks and accompanying rationales sometimes varied, depending upon the indicator considered – the source of the evidence (authority or empirical evidence), the nature of the explanation (description, a vague sense of explanation, or a more precise sense of process and mechanism), or the generality (versus literalness) of the model. Individual student responses in assessments and interviews did not always provide enough information to allow us to distinguish between levels across these indicators. Their responses sometimes matched different levels for different indicators. Thus it was difficult to decide where to locate students' work as they engaged in modeling practice and to identify the key aspects of each level. This difficulty suggests that we combined too many indicators in each construct map. We conclude that different aspects of understanding within the same dimension should be teased apart to better analyze students' modeling practice.

Practice and metaknowledge not always evident together. Students' work as they engaged in modeling practice sometimes included only elements of practice or only elements of metaknowledge, but not both. For example, Level 1 student models occasionally included literal components as well as some non-visible components or processes. If the students provided no rationale for the inclusion of various components and there was no follow-up to augment the item or interview prompt, it was difficult to distinguish between Levels 1 and 2. These cases reinforce the importance of including probes to reveal the reasoning that underlies students' actions and the importance of not relying only on the substance of students' models to ascertain their level of proficiency with the practice.

Ambiguity in language of students' justifications. Sometimes there was ambiguity in students' language (such as in their use of the words "explain" and "detail"). This ambiguity makes it difficult to interpret students' work as they engage in modeling practice since the meaning of these words varies widely across contexts and over time. Many students discussed how they developed their model to "explain." However, "explain" sometimes meant illustrate and at other times meant provide a mechanism. For example, when one fifth grade student was asked after instruction, "Would you call this [example drawing] a model?" he responded: "Because it does explain something in its own way. … It explains like solids, liquids and gas. It explains the key. It explains what's happening to the strawberry banana orange juice." This response demonstrates how the meaning of "explaining" ranged from "illustrating" (the model depicts particular components like solids, liquids, and gases; the model includes a key) to "depicting what is happening" – a Level 2 response. In other cases, students use "explaining" to mean showing a process or mechanism, contrasted with simply "showing something." Similarly, for some students, the words "more detail" at first meant simply providing more information, but later meant more relevant detail that was useful in explaining.

Such ambiguity in students' language about their practice is not surprising. Part of engaging in a practice involves developing particular norms, expectations, and ways of behaving, all of which require specialized use of language. In science, this specialized language overlaps with everyday language, for example, in the use of words such as "evidence," "explain," "argue," "know," and so on. However, these words begin to acquire more specialized and nuanced meanings as students engage more deeply in particular scientific practices. Thus analyzing students' reflective practice requires going beyond the use of particular words (such as "explain" versus "show") and analyzing their work for evidence of particular shades of meaning.

Variation in the extent to which questions encouraged more sophisticated practice. Students' work as they engaged in modeling practice seemed influenced by the opportunities for reflective practice in the assessment items. Some questions and activities explicitly required students to apply their models in new contexts, an important characteristic of Level 3. However, even if students successfully applied their models to these new contexts, it was not always clear whether they fully understood the role of models as tools for generating thinking. Additionally, when students used more general components, as well as relationships between components in the model, in their applications to new contexts, it was not clear that they understood that the model provided more power to explain additional aspects of a phenomenon. We also found some curricular materials gave students the opportunity to consider alternative and multiple representations for modeling components while other materials did not. As a result, there were limited opportunities to engage in modeling practice at Level 3 through classroom artifacts and conversations. These findings suggested the importance of triangulating interpretations of student performance across a range of tasks that include sufficient opportunity to engage in the more sophisticated aspects of the practice and probing to examine students' rationales for making potentially more sophisticated decisions.

Revising the Learning Progression to Address These Challenges

One particularly salient challenge is that students' work as they engaged in modeling practice included multiple features that were not captured by our initial two construct maps. To address this challenge, we continued to unpack the construct maps to describe several subdimensions of student work. We extracted the primary features that seemed to distinguish between levels of student performance and used them as subdimensions for both construct maps. These subdimensions included prevalent themes in students' work as they engaged in modeling practice, such as attention to audience or to evidence. We created four subdimensions that fit both the generative and dynamic construct maps. Our addition of the subdimensions addressed some challenges in analyzing students' work that included multiple features. In order to revise the construct maps and determine the important subdimensions, we looked for exemplar student work for each subdimension. This search allowed us to revise and clarify categories and levels. In this section, we illustrate the motivation for and the nature of these new subdimensions by focusing on Levels 2 and 3 of our revised generative construct map. The summary descriptions of each level from the original construct map (Table 1) are shown in Table 2, followed by their unpacking through four subdimensions.

Table 2. Four Subdimensions of Levels 2 and 3 of the Generative Construct Map.

	Level 2	*Level 3*
Original Description of Level	Students *construct and use a model to illustrate and explain how a phenomenon occurs*, consistent with the evidence about the phenomenon. Students view models as *means of communicating their understanding of a phenomenon* rather than as tools to support their own thinking.	Students *construct and use multiple models to explain and predict additional aspects of a group of related phenomena*. Students view models as tools that can support their thinking about existing and new phenomena. Students consider alternatives in constructing models based on analyses of the *different advantages and weaknesses* for explaining and predicting these alternative models possess.
Underlying Subdimensions	*Students construct and use models…*	
A. Attention to the model's level of abstraction: i. Literalness vs. salience ii. Specificity vs. generality	i. …as a means to show things that are inaccessible to their senses because of scale differences. ii. …to apply the model to new cases, without making the model applicable to a range of phenomena.	i. …by combining components from more than one model. ii. …to make the model applicable to a range of phenomena.
B. Attention to audience and clarity of communication	…considering how well the model reflects the model creator's thinking or how well the model can be understood by others (no attention to the specific type of understanding).	…considering how well the model communicates ideas of evidence and mechanisms to others.
C. Attention to evidence or authority	….drawing support from learned content knowledge, authority, or empirical evidence.	….drawing support from empirical evidence, with justification for how the evidence supports particular claims about the model's fit with the phenomena.
D. Nature of relationship between model and phenomena	…based on a vague sense of explaining or predicting, without specific attention to capturing process or mechanism.	…attempting to represent mechanism or process to explain and predict phenomena.

Attention to the model's level of abstraction. The first subdimension evident in students' work concerned the issue of literal similarity versus generalization of the phenomena. In other words, students moved from viewing models as literal depictions to understanding the importance of showing what is important, even if these aspects are not apparent. As an example of the latter, during instruction a fifth grade student was asked about his model of the interaction between light and matter:

> **Interviewer:** What makes something a model?
>
> **Student:** I think it's a model if it has the real stuff. Like it's not fake. It needs to be accurate. It needs to show what's happening. Like this here - it's showing that all four criteria are met to see [a light source, a detector of light, an object to reflect the light, and an unbroken path]. And it could be a replica of it. So this is a replica of seeing.
>
> **Interviewer**: What do you mean by replica?
>
> **Student:** A scaled-down model. Smaller. Or even see stuff that you can't.

Here the student refers to the criteria that the class decided were important for seeing an object. The student goes beyond viewing models as literal depictions, and refers (indirectly) to what models need to show ("the four criteria met to see"). The student also indicates models can be "a scaled-down" version used to "see stuff that you can't [see]."

We have also seen developing attention to generality in students' understanding that models should apply to new cases. In the following example, a fifth grade student during instruction suggested removing a particular feature in the evaporation model (a light) and replacing it with something more general (because objects other than a light source can produce heat):

> **Interviewer**: Are there any changes you could make to your model (of evaporation) to make it better? … What kind?
>
> **Student**: Remove the light and add an explanation because anything can produce heat.

Attention to audience and clarity of communication. A second subdimension that emerged in students' work is attention to audience and to the clarity of communication. This dimension reflects how students attend to potential audiences for their models and to how well the model communicates to those audiences. At Level 2, this dimension includes students' consideration of how well the model reflects the model creator's thinking or how well the model can be understood. At Level 3 the students consider more specific criteria for helping others understand, namely communicating evidence and mechanisms. For instance, a sixth grade student during instruction stated: "I would use my model to explain my ideas…. Well, I do use my model to explain my ideas because it's something I created that I think how I, in my opinion, would think that something might look." This response reflects Level 2 on the audience subdimension, showing that the student sees that the purpose of the model is to reflect her thinking. Similarly, a fifth grade student during instruction commented that scientists “use models to help them figure out something or help other people understand it.” After the evaporation/condensation modeling unit, another fifth grade student referred to her use of the model, in a more sophisticated Level 3 fashion, as communicating a mechanism (the speed of the moving particles) that explains phase change:

Interviewer: What were you doing when you were using your model? How did you use your model in class?

Student: Um, well, I used it to explain evaporation.

Interviewer: OK. How did you use it to talk to your group members about it?

Student: Um, well, I showed the molecules and pointed out how they weren't moving as fast as these ones.

Attention to evidence or authority. A third subdimension observed in students’ work relates to the nature of support students used to construct their models. This support included prior knowledge, observational or empirical evidence, and authoritative sources such as a reading or the teacher. As an example of Level 2 of this subdimension, a fifth grade student after instruction described how he constructed the model based on the experiments conducted in class during the evaporation/condensation modeling unit:

Interviewer: Did you include any of the evidence of the experiments in any of those models?

Student: …Condensation, I think … we did, because we had evidence because it did collect onto the bottle from the air because it’s in a sealed type thing where it had only air. It was just a cold bottle.

Students often did not prioritize empirical evidence, apparently viewing information from books and teachers as on par with empirical evidence. For example, a sixth grade student stated during the unit: “After you get the information, then you can make the model… The information that comes from the book or whatever you learn about.”

Some students were able to provide more advanced (Level 3) rationales related to this subdimension, including justifying why empirical evidence supported constructions of or revisions to a model. For example, after the unit, a fifth grade student justified his model changes by referring to an experiment with hot and cold water in which the class discovered that both hot and cold water evaporated, but that hot water evaporated more quickly:

Interviewer: What made you decide to do those changes?

Student: Because we did a few experiments like that hot and cold water one. For the hot and cold water I saw that even after that it was still evaporating. So then when I looked back to that one I knew it was wrong. So for the next one I drew the water evaporating the whole time.

This subdimension presents challenges for students learning the practice of scientific modeling. While this scientific practice prioritizes the use of evidence to build knowledge over authoritative sources that provide answers, this privileging of evidence is unfamiliar to students. Thus it is not surprising to find that students lump evidence with information from teachers and textbooks. Students in traditional science classrooms are more accustomed to reporting *what* they know (i.e., the answers) than *how* they know it (i.e., the justifications) (Berland & Reiser, 2011; McNeill et al., 2006).

Nature of the relationship between models and phenomena. We have also observed students reflect on the relationship between their models and phenomena, as well as on the explanatory nature of their models. This is reflected in the fourth subdimension. At Level 2, students highlight the purpose of a model as an explanation of something about a phenomenon but with an unarticulated sense of what it means to explain. For instance, during a focus group interview, a student toward the end of the IQWST sixth grade chemistry unit stated: "Using a model, you can explain something that's going on, like we did with the air. We made models of what's going on in the air." At other times, students at Level 2 on this subdimension were more specific about the relationship between the components of the model. The following exchange occurred in an interview during the IQWST sixth grade physics unit:

> **Interviewer:** Can you tell me a little bit about the drawing that you made here [the student's light model]?
>
> **Student**: Well it is showing, it's explaining everything that light can do. Here it's reflecting, here it's transmitting, and later it absorbs. Here it is reflecting off a tree, and this person standing here can actually see it because the light is entering her eyes. Over here it shows shadowing because light is being blocked there.

We coded responses as Level 3 when students incorporated a mechanism into their models that made explicit how the relationships between model components led to the observed phenomena. In the next example, in a class that had just begun to identify empty space (which they called "nothing") as important, a student in the IQWST sixth grade chemistry unit justified the importance of empty space between particles in a model of odors traveling through air:

> **Teacher** [pointing to the space between particles in the model on the whiteboard]: Nothing! So why is nothing [empty space] there, in the middle of the molecules, able to help me realize why the particles can be compressed, how there's nothing there? [Student]?
>
> **Student**: Because since there's *nothing* there [it] can show like the spaces in-between the particles. And when they're compressed, they go together, and then there's like less of that space. But when they expand, they spread apart and there's more of the "nothing."

There are several challenges in analyzing students' work as they engage in modeling practice with respect to the model-phenomena relationship depicted.

While sometimes it is feasible to distinguish between models constructed using a vague sense of "explaining something" (Level 2) and models involving more precise explanations through inclusion of components representing processes or mechanisms (Level 3), the distinction is not always clear. Students' rationales are frequently limited and sometimes ambiguous. For example, students may provide a more sophisticated mechanism or process in their model but give a limited rationale for the use of the correspondingly sophisticated model-phenomenon relationship. As in other cases, it is important not to assume that the improved model means a more sophisticated understanding of the practice.

Future work with the subdimensions. We continue to try to better understand how to clarify the different subdimensions as we analyze students' modeling work with respect to these subdimensions. We also continue to analyze the role of content (for example, explanatory components in some science topics might make it easier for students to improve in the model-phenomena relationship subdimension) and scaffolding (probing for adequate rationales) in students' work. Thus we are exploring students' work as they engage in modeling practice across assessment measures (written assessment items, interviews, class talk) in different contexts (different science topics and levels of scaffolding). Finally, the distinction between levels (such as that between Levels 2 and 3) remains relatively coarse. We continue to work on determining how to trace shifts in students' practice within a particular level as well as across levels by conducting finer-grained analyses of students' engagement in modeling practice. We are also examining the empirical relationships between these four subdimensions. We continue to test the subdimensions of the construct maps against the data from classroom enactments.

Overall, the subdimensions present a more detailed consideration of the reflective practice than that represented in our initial construct maps. The subdimensions are important for capturing changes in students' reflective practice. While there are still many challenges with evaluating students' work as they engage in modeling practice, this finer-grained framework allows us to determine how to provide better support for learners through the design of more effective curriculum materials and instruction.

SUMMARY OF CHALLENGES IN DEFINING A LEARNING PROGRESSION FOR A SCIENTIFIC PRACTICE AND IMPLICATIONS FOR FUTURE WORK

In this section, we summarize the challenges we have faced and the implications for learning progression research on scientific practices, as well as more general implications. Within the challenges and implications, we discuss two clusters of related issues – (a) analyzing a practice that combines performance and metaknowledge and (b) the design research nature of learning progression research.

Supporting and Assessing Reflective Practice

Theoretical and methodological challenges arise when developing a learning progression for a scientific practice. A particular challenge arises in the commitment to developing a learning progression for reflective practice – the integration of the performance of elements of the practice with underlying metaknowledge. We assess the combination of students' performance of elements of modeling practice and metamodeling knowledge so as to avoid teaching and assessing routine procedures on the one hand and decontextualized understandings about the nature of science on the other hand. However, the focus, breadth, and number of elements in our construct maps make the associated tools for analyzing student work more complex. We have outlined several difficulties in analyzing gains students make in the practice. It is not convincing to rely solely on students' general articulations of metaknowledge in response to surveys or interviews that focus on general epistemological and nature of science questions. Instead, we look for performance accompanied by indicators of related understanding as reflected in our construct maps (Tables 1 and 2, Appendix). We try to guard against unduly crediting some types of improved student modeling work as indicating improvement in reflective practice. We do not want to interpret improved rote performance as improvement in reflective practice. Nor do we want to assume improved model substance necessarily indicates increased understanding about the practice. We also deal with challenges related to the opportunities for students to engage in reflective practice in current classroom contexts.

Distinguishing reflective practice from rote performance. Our goal is to support and evaluate science as reflective practice. Thus we analyze metamodeling knowledge as used in performance of modeling practice. This analysis requires investigating how students construct, use, evaluate, and revise models in particular content domains. We have described several cases where we attempt to avoid crediting students with improved reflective practice if they show improvement in performance of the practice without understanding. One example is teachers' repeated instructions to students about labeling their diagrams in particular ways. Thus students' labels – without accompanying justification – would not receive credit with respect to our construct maps. In addition to the labels, we also look for students' comments about how labels clarify important components of the mechanism or how labels help the audience construct a chain of cause and effect.

Distinguishing improvements in practice from increased content knowledge. Similarly, we attempt to distinguish improvement in the substance of the model from more general understanding of the practice. For example, as students move from viewing air as continuous matter to viewing it as consisting of particles, they reflect this change in their models. The new models clearly show improvement in their ability to account for phenomena. A student's model might explain that an odor spreads across a room because the odor particles collide with air particles and eventually spread (diffuse) across the room. However, this model does not necessarily mean students have developed more sophisticated ideas about the importance of including mechanisms in their models. We also want to see students justify that the new model represents an improvement because it explains, in a step-by-step fashion, what happens to the odor. One strategy for addressing this issue is to use neutral content assessment items where the modeling task is embedded in scientific phenomena that are not the content focus of the modeling unit.

The opposite interpretation problem may also occur. Because of the link between content knowledge and students' modeling practices, we may underestimate students' proficiency with the latter. It is possible that unfamiliarity with the target domain makes students hesitant to speculate on possible mechanisms even though they may realize that mechanistic explanations are a goal of modeling. A partial response to the problem is to code the scientific accuracy of students' mechanisms separately from whether they attempt to include a mechanism in their model and whether they can describe the importance of the mechanism. However, the possibility remains that students may hesitate to speculate if they lack confidence in their knowledge of the domain.

Tension between the scientific practice and prevalent classroom norms. Scientific practices entail a system of norms, expectations, and ways of acting that may conflict with classroom and school norms. Many researchers have written about the need to move from structuring classrooms in which students are consumers of knowledge provided by authority to structuring classrooms in which students are members of a community of learners who build, evaluate, and refine knowledge according to the practices of an intellectual community (e.g., Bielaczyc & Collins, 1999; Brown & Campione, 1994; Driver, Newton, & Osborne, 2000; Herrenkohl, Palincsar, DeWater, & Kawasaki, 1999). Challenges arise when encouraging learners to engage in scientific practices that require that they constructively argue with peers, take the role of authors of knowledge, and recognize that knowledge is continually refined.

The divergence of scientific practices from traditional classroom practices also creates challenges for analyses of students' work as they engage in modeling practice. First, the emphasis on models as the target of sense-making moves beyond simply predicting or capturing regularities in equations or recounting what happens in a particular scientific phenomenon. Instead, modeling focuses students on developing mechanistic explanations of *how* and *why* something happens. Many reforms call for this type of deep understanding. However, such reforms require a shift in the perspective of learners in terms of what constitutes an answer to a scientific question. Second, our attempt to document reflective practice relies on students' rationales for their decisions. Rather than focusing solely on answers, students need to justify why they made particular decisions while engaging in the practice. Furthermore, documenting students' modeling work relies on the idea that students' engagement in the practice is driven by what makes sense and has value in the community rather than by what they are directed to do by the curriculum and by teachers. Thus the nature of students' attitudes within a scientific practice differs from those within traditional classroom practices in which students view learning science as repeating fixed answers to questions.

Viewing practice as "doing with understanding" (Barron et al., 1998) and as reflecting attitudes and expectations creates challenges for support and assessment. However, a reduction of the practice to simply knowledge or skills would diminish the importance of engaging learners in the meaningful knowledge building in science (NRC, 2007, 2009).

Developing Learning Progressions Is Design Research

A second set of challenges concerns the necessary nature of research to define, investigate, and revise a learning progression. A learning progression can be viewed as a hypothesis (Alonzo & Steedle, 2009; Wilson, 2009). However empirical investigations of a learning progression are not best viewed as "hypothesis testing" in the typical sense of this research approach. Learning progressions are conjectures about the *potential* paths learning can take. The learning progression is a hypothesis about how complex learning goals (e.g., an understanding natural selection or the nature of matter) can be built from constituent understandings. The learning progression presents the important elements of a target idea or practice and identifies productive, intermediate stepping stones that lead to advances in reasoning, upon which more sophisticated versions can be built with appropriate support.

The assumption explicit in work on learning progressions is that multiple pathways are possible in movement toward more sophisticated understandings. Furthermore, a learning progression is not a hypothesis about necessary stages of understanding through which learners inevitably progress. Learning progressions are contingent on (1) learning experiences students encounter and (2) support for making sense of those experiences (Lehrer & Schauble, 2009). The learning progression hypothesis includes the elements of knowledge and practice and descriptions of how these elements can build on one another through productive pathways.

Moreover, part of the work of defining a learning progression involves identifying the particular substance of the learning target. Even if there is agreement (e.g., from national and state standards) that particular scientific ideas, such as evolution or the particle nature of matter, should be taught, the research program must include arguments for which aspects of these ideas are essential. These arguments may draw on empirical findings but may also rely on value-based considerations. Empirical evidence can illuminate which challenges may arise in reaching a particular understanding of an idea and can identify important component ideas that may be implicit in the target idea and, therefore, should be targeted for instruction (Krajcik et al., 2008). However, which understandings should be targeted is not just an empirical question; choices about which aspects of these understandings are important must also be made. For example, should learners construct an understanding of evolution by natural selection in advance of understanding the molecular basis of inheritance? Or is the construct of natural selection only useful if it can be built on understanding the molecular basis of inheritance?[1] Empirical research can identify what is feasible, as well as the potential advantages or disadvantages of particular pathways. But empirical evidence cannot be used alone in arguments when value-based considerations must also be taken into account.

The need for both empirical and value-based considerations is even more apparent in learning progressions for scientific practices. National and state standards are much less explicit about which aspects of scientific practice should be targets of instruction; such standards identify developing and investigating explanations, models, and theories quite broadly. There are many variations of scientific practice that could be proposed as relevant for classrooms. These variations might emphasize different aspects of developing knowledge in science, focusing, for example, on designing investigations, analyzing data, developing arguments, producing explanatory texts, and so on. Within any individual element, such as argumentation, there are multiple ways of defining the element, each emphasizing different criteria, such as logical consistency, empirical evidence, coherence, etc. (Sampson & Clark, 2008).

Thus defining a learning progression for a practice involves more than simply investigating the "best way" to reach the learning goal of a particular scientific practice. Part of the research program should develop research-based arguments for what it means to engage in that scientific practice in the classroom. This work combines attempts to design classroom contexts that support the practice with empirical investigations about what is possible and what challenges arise. The hypotheses investigated necessarily include commitments to what should be learned and initial conjectures about what reasonable stepping stones might look like.

A key challenge in this design work arises because it is not a simple instructional task to change the practices through which learners build knowledge in classrooms. As we have said, there is tension in between what counts as "knowing" something in science and students' expectations about knowledge and authority (Herrenkohl et al., 1999; Hogan & Corey, 2001). Thus we are faced with the study of a practice that differs in important ways from what is currently present in classrooms. Dramatic changes in the way knowledge is built through classroom interactions can only occur incrementally and over time (Lehrer & Schauble, 2006).

The design research nature of work on learning progressions has implications for the nature of theories that can be built in this domain. The theories constructed from studies of learning progressions are arguments, supported by evidence, about possible pathways and their associated challenges. It would be over-interpreting the evidence to argue for a necessary sequence. Furthermore, it is important to contextualize the evaluation of these design research arguments with respect to the assumptions that were made about the learning target. Different conceptualizations (e.g., emphasizing argumentation rather than model building or different approaches to supporting modeling) may reveal different challenges and different pathways for learning.

ACKNOWLEDGEMENTS

This chapter was written while Brian Reiser was a Weston Visiting Professor in the Department of Science Teaching, Weizmann Institute of Science. This research was funded by the National Science Foundation (NSF) under Grant ESI-0628199 to the MoDeLS project at Northwestern University and by grants ESI-0439352 and ESI-0439493 to the IQWST project at the University of Michigan and Northwestern University. The opinions expressed herein are those of the authors and do not necessarily reflect those of the NSF. We are grateful to our colleagues Elizabeth A. Davis and Barbara Hug for their helpful insights and feedback on the learning progression. We are also indebted to the graduate students and postdocs who have provided input into these ideas through their work on the MoDeLS project. We thank Hamin Baek, Jing Chen, Hayat Hokayem, and Li Zhan (Michigan State University); James Hagerty and Yael Bamberger (University of Michigan); Brandy Buckingham (Northwestern University); and Yael Shwartz, Malka Yayon, and Dana Veder Weiss (Weizmann Institute of Science).

NOTE

[1] This is the decision reflected in the American Association for the Advancement of Science (AAAS; 2001, 2007) strand maps.

REFERENCES

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, . . . Tuan, H.-l. (2004). Inquiry in science education: International perspectives. *Science Education, 88*, 397–419.

Acher, A., Arcà, M., & Sanmartí, N. (2007). Modeling as a teaching learning process for understanding materials: A case study in primary education. *Science Education*, *91,* 398–418.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

American Association for the Advancement of Science. (2001, 2007). *Atlas of science literacy,* (Vols. 1–2). Washington, DC: Author.

Baek, H., & Schwarz, C. (2011, April). *How teachers and students make sense of and perform model revision and consensus model construction in two 5th grade classrooms.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Orlando, FL.

Baek, H., Schwarz, C., Chen, J., Hokayem, H., & Zhan, L. (2011). Engaging elementary students in scientific modeling: The MoDeLS fifth-grade approach and findings. In M. S. Khine & I. M. Saleh (Eds.), *Models and modeling in science education* (Vol. 6, pp. 195–218). Dordrecht, The Netherlands: Springer.

Barron, J. S., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., & Bransford, J. D. (1998). Doing with understanding: Lessons from research on problem- and project-based learning. *Journal of the Learning Sciences, 7*, 271–311.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*, 26–55.

Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education, 95*, 191–216.

Bielaczyc, K., & Collins, A. (1999). Learning communities in classrooms: A reconceptualization of educational practice. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (pp. 269–292). Mahwah, NJ: Lawrence Erlbaum Associates.

Brewer, W. F., Chinn, C. A., & Samarapungavan, A. (1998). Explanation in scientists and children. *Minds and Machines, 8*, 119–136.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah, NJ: Lawrence Erlbaum Associates.

Carey, S., & Smith, C. (1993). On understanding the nature of scientific knowledge. *Educational Psychologist, 28*, 235–251.

Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education, 22*, 1041–1053.

Clement, J. (2008). Student/teacher co-construction of visualizable models in large group discussion. In J. Clement & M. A. Rea-Ramirez (Eds.), *Model based learning and instruction in science* (pp. 11–22). New York, NY: Springer.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*, 287–312.

Edelson, D. C. (2001). Learning-for-use: A framework for integrating content and process learning in the design of inquiry activities. *Journal of Research in Science Teaching, 38*, 355–385.

Edelson, D. C., & Reiser, B. J. (2006). Making authentic practices accessible to learners: Design challenges and strategies. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 335–354). New York, NY: Cambridge University Press.

Fortus, D., Rosenfeld, S., & Shwartz, Y. (2010, March). *High school students' modeling knowledge.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *Journal of Research in Science Teaching, 28*, 799–822.

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education, 22*, 1011–1026.

Herrenkohl, L. R., Palincsar, A. S., DeWater, L. S., & Kawasaki, K. (1999). Developing scientific communities in classrooms: A sociocognitive approach. *Journal of the Learning Sciences, 8*, 451–493.

Hogan, K., & Corey, C. (2001). Viewing classrooms as cultural contexts for fostering scientific literacy. *Anthropology & Education Quarterly, 32*, 214–243.

Justi, R., & van Driel, J. (2005). A case study of the development of a beginning chemistry teacher's knowledge about models and modelling. *Research in Science Education, 35*, 197–219.

Kanter, D. E. (2010). Doing the project and learning the content: Designing project-based science curricula for meaningful understanding. *Science Education, 94*, 525–551.

Kenyon, L., Schwarz, C., & Hug, B. (2008). The benefits of scientific modeling. *Science and Children, 46*(2), 40–44.

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*, 1–32.

Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 831–879). Mahwah, NJ: Lawrence Erlbaum Associates.

Lehrer, R., & Schauble, L. (2000). Modeling in mathematics and science. In R. Glaser (Ed.), *Advances in Instructional Psychology: Vol. 5. Educational design and cognitive science* (pp. 101–159). Mahwah, NJ: Erlbaum.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*, 635–679.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy: Supporting development in learning in contexts. In W. Damon & R. M. Lerner (Series Eds.) & K. A. Renninger & I. E. Sigel (Volume Eds.), *Handbook of Child Psychology: Vol. 4. Child psychology in practice* (6th ed., pp. 153–196). Hoboken, NJ: John Wiley and Sons.

Lehrer, R., & Schauble, L. (2009). Images of learning, images of progress. *Journal of Research in Science Teaching, 46*, 731–735.

Lehrer, R., Schauble, L., & Lucas, D. (2008). Supporting development of the epistemology of inquiry. *Cognitive Development, 23*, 512–529.

Lesh, R., & Doerr, H. M. (2000). Symbolizing, communicating, and mathematizing: Key components of models and modeling. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 361–383). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Doerr, H. M. (2003). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 3–33). Mahwah, NJ: Erlbaum.

Lesh, R., & Lehrer, R. (2003). Models and modeling perspectives on the development of students and teachers. *Mathematical Thinking and Learning, 5*, 109–129.

McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education, 93*, 233–268.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences, 15*, 153–191.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

National Research Council. (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, DC: The National Academies Press.

Passmore, C. M., & Svoboda, J. (2011). Exploring opportunities for argumentation in modelling classrooms. *International Journal of Science Education*. Advance online publication. doi: 10.1080/09500693.2011.577842

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education, 92*, 499–525.

Sampson, V., & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*, 447–472.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education, 89*, 634–656.

Schwarz, C. V. (2002). Is there a connection? The role of meta-modeling knowledge in learning with models. In P. Bell, R. Stevens, & T. Satwidz (Eds.), *Keeping learning complex: The proceedings of the fifth International Conference of the Learning Sciences (ICLS)*. Mahwah, NJ: Erlbaum. Retrieved from http://schwarz.wiki.educ.msu.edu/file/view/Schwarz_ICLS_metapaper.pdf

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., . . . Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*, 632–654.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*, 165–205.

Shwartz, Y., Weizman, A., Fortus, D., Krajcik, J., & Reiser, B. (2008). The IQWST experience: Using coherence as a design principle for a middle school science curriculum. *The Elementary School Journal, 109*, 199–219.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade students' epistemologies of science: The impact of school science experiences on epistemological development. *Cognition and Instruction, 18*, 349–422.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

Snir, J., Smith, C. L., & Raz, G. (2003). Linking phenomena with competing underlying models: A software tool for introducing students to the particulate nature of matter. *Science Education, 87*, 794–830.

Spitulnik, M. W., Krajcik, J., & Soloway, E. (1999). Construction of models to promote scientific understanding. In W. Feurzeig & N. Roberts (Eds.), *Modeling and simulation in science and mathematics education* (pp. 70–94). New York, NY: Springer-Verlag.

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2010). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching, 47*, 687–715.

Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics, and science in the classroom* (pp. 515–565). Washington, DC: The National Academies Press.

Tabak, I., & Reiser, B. J. (2008). Software-realized inquiry support for cultivating a disciplinary stance. *Pragmatics and Cognition, 16*, 307–355.

Treagust, D. F., Chittleborough, G., & Mamiala, T. L. (2002). Students' understanding of the role of scientific models in learning science. *International Journal of Science Education, 24*, 357–368.

van Driel, J. H., & Verloop, N. (1999). Teachers' knowledge of models and modelling in science. *International Journal of Science Education, 21*, 1141–1153.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K–12 mathematics and science education in the United States*. Retrieved from Horizon Research website: http://www.horizon research.com/insidetheclassroom/reports/looking/complete.pdf

White, B. Y. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction, 10*, 1–100.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3–118.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716–730.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education, 92*, 941–967.

Zhan, L., Baek, H., Chen, J., & Schwarz, C. (2011, April). *Investigating patterns and rationales for 5th grade students' scientific models*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Orlando, FL.

APPENDIX

A Construct Map for the Dynamic Dimension: Understanding Models as Changeable Entities

Level	*Description of Reflective Practice (Including Performance of Elements of the Practice and Associated Metaknowledge)*
4	*Students consider changes in models to enhance the explanatory power prior to obtaining evidence* supporting these changes. Model changes are considered to develop questions that can then be tested against evidence from the phenomena. Students evaluate competing models to *consider combining aspects of models* that can enhance the overall explanatory and predictive power.
3	*Students revise models in order to better fit evidence that has been obtained and to improve the articulation of a mechanism in the model*. Thus models are revised to improve their explanatory power. Students compare models to *see whether different components or relationships fit evidence more completely and provide a more mechanistic explanation of the phenomena*.
2	*Students revise models based on information from authority (teacher, textbook, peer) rather than evidence gathered from the phenomenon or new explanatory mechanisms*. *Students make modifications to improve detail or clarity or to add new information*, without considering how the explanatory power of the model or its fit with empirical evidence is improved.
1	*Students do not expect models to change with new understandings*. They talk about models in absolute terms of right or wrong answers. Students compare their models by assessing if they are *good or bad replicas of the phenomenon*.

Note. Adapted from "Developing a Learning Progression for Scientific Modeling: Making Scientific Modeling Accessible and Meaningful for Learners," by C. V. Schwarz, B. Reiser, E. A. Davis, L. Kenyon, A. Acher, D. Fortus, Y. Shwartz, B. Hug, and J. Krajcik, 2009, Journal for Research in Science Teaching, 46, p. 647.

LINDSEY MOHAN AND JULIA PLUMMER

EXPLORING CHALLENGES TO DEFINING LEARNING PROGRESSIONS

Learning progressions (LPs) in science have received considerable attention in the last five years, with the expectation that they will receive even more attention in future years (e.g., Corcoran, Mosher, & Rogat, 2009; National Assessment Governing Board, 2008; National Research Council [NRC], 2007, 2011; National Science Foundation, 2009, 2010a,b). This emerging field has been touted as one with great promise for improving science instruction in schools (NRC, 2007). Yet there are concerns about the promise of LPs. The pendulum often swings in education as new ideas replace once-promising ideas. The diversity among LP projects suggests how difficult it is to use LPs as a "basis for a dialogue" among various stakeholders in the science education community (NRC, 2007, p. 214). Therefore it is necessary to examine the diversity in LP work in order to identify its core identity if LPs are to have enduring influence in science education.

In the last few years our understanding of LPs has become clearer and our definitions more precise. The science education research community generally agrees that every LP should include some of the same characteristics, even if these characteristics appear in different ways. A report from the Consortium for Policy Research in Education (CPRE) describes LPs as "hypothesized descriptions of the successively more sophisticated ways student thinking about an important domain of knowledge or practice develops…over an appropriate span of time" (Corcoran et al., 2009, p. 37). The CPRE report echoes key features of LPs previously described by the NRC (2007) and NAGB (2006). These ideas include the following: (1) LPs are hypotheses about learning in a given domain; (2) LPs include Upper and Lower Anchors, with the Upper Anchor grounded in societal goals for learning core knowledge and practices in science, and the Lower Anchor grounded in the ideas that students bring to the classroom; and (3) LPs describe ways students may develop more sophisticated ways of thinking in a domain, often with support of specific instructional strategies.

It is clear that the chapter authors in the Defining Strand rely, in part, on established definitions and norms for LP work although they do not discuss these definitions and norms explicitly. In each chapter, the authors begin by asking similar questions about LP construction: What are the big ideas, disciplinary knowledge, and practices that form the backbone of the LP? How do we identify entry points and intermediate steps in the LP? How many levels are there and at what grain size? Can we observe learning performances by students given how we have defined and operationalized our LP?

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 139–147.

The goals of the Defining Strand are to address the common challenges that arise in the process of defining learning progressions and to consider the implications of the decisions taken in this process. Many differences among LPs do not occur because researchers ask different questions; rather, the differences occur because LP researchers have made different choices about the LPs based on a certain discipline, audience, or philosophy of learning. Such choices lead these researchers to describe learning in particular ways that may or may not align with other researchers' descriptions. As a result, there is considerable discussion about LPs and their use among researchers and other stakeholders. For example, a LP definition that uses a particular set of instructional materials and learning conditions may limit the generalizability of this LP for large-scale purposes. For that reason, it is critical in the defining process to consider who will use the LP and for what purpose.

DECISIONS ON THE LP DOMAIN: CHOOSING THE BIG IDEAS

In defining a LP, researchers face the challenge of making the LP understandable not only to learning scientists, but also understandable to a broad, inevitably more diverse, audience. What the LP communicates is often the result of decisions made during the defining process. One important decision that occurs early in the defining process is the articulation of big ideas for the LP. Big ideas have broad explanatory power and can be used to shape a vision of science education in which curricula are designed around concepts central to the discipline. The authors of the defining strand offered different types of justifications for their choice of big ideas. Plummer (chapter 5) lists three criteria she used to choose a big idea for celestial motion. For Plummer, the big idea should (1) represent ways of knowing and understanding the universe, (2) describe explanatory models that can be learned beginning with a child's observations of the world, and (3) explain multiple, unified astronomical phenomena such that learning to explain an individual phenomenon helps learners increase knowledge of the overall big idea. Gunckel, Mohan, Covitt, and Anderson (chapter 4) propose big ideas that center on knowledge and practices associated with scientific, model-based views of socio-environmental systems. Unlike Plummer, their challenge was in defining LPs that span traditional scientific disciplines. For Schwarz, Reiser, Acher, Kenyon, and Fortus (chapter 6), choosing a big idea presents challenges because there is little consensus in the research community on the meaning of scientific modeling. Defining a LP for modeling practices required the researchers to choose a focus–in this case, models as a way to "embody some aspects of causal and often non-visible mechanisms or explanatory components of phenomena" (p. 104) rather than data modeling, as has been proposed by other researchers.

In identifying big ideas, LP researchers often push against the boundaries of the realities in schools; this is true especially in situations where a school does not teach the LP content despite the importance of the domain or practice to scientific literacy. For example, Plummer (chapter 5) notes that many school curricula do not include astronomy concepts although these concepts are recognized as central ideas in science. Schwarz et al. (chapter 6) note that modeling is not included in most school science curricula although modeling practices—as well as science practices in general—are important considerations in evaluations of scientific literacy (NRC, 2011).

How LP researchers select big ideas in science and the focus of school science curricula may differ significantly. This difference may cause tension in the science education community. In some cases, the topical organization of school science curricula may not connect explicitly to underlying explanatory models. For example, in the school year a teacher may cover the day/night cycle, phases of the moon, and the seasons but not the unifying principles of celestial motion that connect the three phenomena. Because of this omission, Plummer (chapter 5) and her research group developed a LP on celestial motion. They recognized that they could have developed LPs for each topical area (e.g., a LP for seasons, a LP for phases of the moon). However, they developed a LP for celestial motion because of its broad explanatory power and its potential to reduce the fragmentation of topics.

Thus one challenge for LP researchers when choosing a big idea is the extent to which they allow standards and school curricula to influence their definitions. Because LP research is a response to perceived problems with current curricula, standards, and assessment practices (Corcoran et al., 2009), it is natural that LP researchers take a somewhat skeptical view of these resources. Their perception is that much of the current science curriculum is fragmented, and, as a result, may lead to superficial coverage of science concepts (e.g., Kesidou & Roseman, 2002). Aligning a LP's domain to the current organization of standards and curricula may not provide the coherence LP researchers seek. Researchers must decide how to balance a new vision of the content and how it might be presented to students, while simultaneously recognizing how other participants in the science education community might respond to the new organization around big ideas. In short, LP researchers must consider the perceived usefulness and relevance of big ideas to other stakeholders while also considering how to bring coherence to science education. As the authors in the Defining Strand acknowledge, there are multiple ways to define the domain of their LPs. Therefore LP researchers should explicitly explain and justify the decisions they made in choosing the domain of the LP. Justifying choices about the domain of the LP is critical for communication both within the research community and across multiple groups with a stake in LP work.

DEFINING LP LEVELS

Levels are used in LPs to describe changes in student understanding in a domain. The use of levels in LP work resonates with earlier educational philosophies —e.g., developmental corridors (Brown & Campione, 1994) and zones of proximal development (Vygotsky, 1978). For example, Brown and colleagues first introduced bandwidths of competence (Brown & Reeve, 1987). Subsequently she described these bandwidths of competence as developmental trajectories, which are remarkably similar to LP researchers' descriptions of LP levels:

> One needs to understand a developmental trajectory that grows in stepping-stones toward mature thinking…beginning with knowledge of early precocity

> of children as they enter preschool…One can build on this early knowledge by extending and refining it and at the same time concentrating on suspected problems of interpretation (Brown, 1997, pp. 409–410).

Yet, as they work with LPs, LP researchers continue to wrestle with how to define and operationalize these levels. One of the most important tasks for LP researchers is capturing the complexity in student learning. It is extremely difficult to describe the development of learning over time, across many students, and across many age levels. In defining LPs, therefore, LP researchers make decisions about how to represent this complexity, a process that necessarily involves simplifications and trade-offs. In this section, we discuss three decisions LP researchers must make as they define and operationalize LP levels. We also discuss the influence of these decisions on how LP researchers capture student learning. These decisions relate to the grain size, composition, and research validation/refinement of LP levels.

Deciding on Grain Size

One decision that influences how researchers capture diversity in learning is the grain size of a LP. Some LPs represent discrete changes in conceptual networks while others represent major shifts in worldviews. Thus grain size can be thought of as the amount of content in a LP level or as the size of the shift between LP levels. Coarse-grained LPs have levels that focus mostly on salient themes and patterns in learning, while fine-grained LPs have levels that focus on more detailed descriptions of learning. Decisions about grain size affect how assessments are designed and how student performance is analyzed.

Unfortunately, there is no grain size that works with all LPs. Nor is there a grain size that speaks to all stakeholder groups. Schwarz et al. (chapter 6) made decisions on the basis of "what grain size is needed to capture change" (p. 103). The aspects of change that need to be captured will likely vary, depending on the purposes of LPs. For example, large-scale standards and assessments may require coarse-grained LPs. In contrast, teachers and curriculum developers may need fine-grained LPs that focus on details of learning in particular grades. LP researchers typically choose a grain size that captures critical elements of change in the LP. Every LP researcher must strike a balance between capturing just enough to describe key differences in learning but not so much that the LP is inaccessible to other audiences or that levels can not reliably distinguish student performances. Often, this balance is achieved only with several iterations of framework design and empirical testing that refine the levels.

One approach is to start with initial descriptions of levels and then use assessment results to revisit the decision about choice of grain size. Assessment results may indicate a difficulty in reliably classifying students into levels. In this case, levels of a fine-grained LP may be collapsed, resulting in a coarser-grained LP with fewer levels. On the other hand, LP researchers may find that existing levels do not capture important differences in student learning. When Schwarz et al. (chapter 6) encountered this difficulty, they added subdimensions that provided additional detail about student performances and more nuanced descriptions of changes in students' modeling practice. Adding this detail to the levels allowed them to analyze and describe student learning with greater accuracy.

Deciding What Constitutes a LP Level

One challenge researchers face when constructing LP levels concerns whether to include alternative conceptions (also called misconceptions or naïve conceptions) in the LP or to focus solely on scientifically correct ideas. As some LP researchers have found (e.g., Gunckel et al., this volume), misconceptions cannot be excluded from LPs because they are central to making sense of student knowledge and practices at different levels. Others argue that subsets of students' misconceptions serve as productive stepping stones and, thus should be included in the LP levels (e.g., Wiser, Smith, & Doubler, this volume). However, it may be politically challenging to include ideas that are not viewed as scientifically correct in LPs that are used for purposes such as developing standards (e.g., Foster & Wiser, this volume).

LPs are designed to describe students' knowledge and practices at different levels of learning sophistication. Because students at lower and intermediate levels may view the world quite differently than scientists do, alternative conceptions that seem scientifically illogical may make sense to students. Simply focusing on productive knowledge and practices may not be enough to describe the true nature of student learning at different levels Even though alternative conceptions are sometimes "dead ends" for learning, they are essential to making sense of students' knowledge and practices and for deciding how to approach instruction. Alternative conceptions at one level can be used to make progress toward another level. For example, Mohan, Chen, and Anderson (2009) describe an alternative conception called the gas-gas or the CO_2-O_2 cycle. This cycle can be viewed as a misconception about how plants and people exchange gases in the air. When left unchecked, the gas-gas cycle may become a barrier to understanding matter exchange between air and biomass. However, the gas-gas cycle, which also describes students' initial attempts at tracing matter, may be leveraged to increase students' knowledge of how matter transforms between systems.

It seems reasonable for LP researchers to focus on describing the most effective approach to instruction; thus alternative conceptions may play a lesser role as LP levels increase in sophistication. Plummer (chapter 5) makes this point. She describes lower LP levels as more consistent with children's alternative conceptions and upper levels as more consistent with scientifically correct ideas, while still noting the limitations of students' ideas in relation to the scientific views.

Decisions about whether to include ideas that are not fully scientifically correct may need to be made in consultation with other members of the science education community. It may be helpful for some audiences (such as teachers or curriculum developers) to have additional information about the ideas that their students are likely to hold. In contrast, standards developers—facing pressure to include only correct ideas in standards—may prefer levels with greater adherence to canonical ideas.

Clarifying and Validating Levels

Defining a LP typically requires more than one attempt. In fact, most LP researchers say their initial LPs are quite different from the ones they are currently working on. LPs are evolving frameworks that are continuously revised as researchers assess and model student understanding and as teachers and other practitioners begin to use the LP and associated materials. In this section, we consider three issues that LP researchers must grapple with as they work to clarify and validate LP levels: limitations of existing research, challenges in obtaining validity evidence, and the role of instruction.

Limitations of the existing research. The research on students' thinking and learning is a source of essential information for the work of defining a LP, particularly the lower LP levels. The work is hampered, however, when there is no published research for a particular topic or area. For example, Plummer (chapter 5) found that while some areas of astronomy have been investigated thoroughly, providing rich information on children's early understandings, there is a lack of research in other areas of astronomy. For the big idea of celestial motion, Plummer found that the research does not explain clearly how students progress from their initial ideas to the scientific explanation–in specific topics related to celestial motion (such as the apparent motion of the stars or the reason for the seasons) and between topics (such as the rich and complex scientific understanding of how the motion of objects in the solar system affects our observations from the earth). In addition, Plummer found that the research on celestial motion topics does not focus on how children learn to construct explanations that account for different frames of reference (earth-based observations and the actual motion of objects in space). She also found that research has not adequately explained the role of spatial knowledge and reasoning in how children learn about astronomy. To fill these gaps in the research, Plummer has begun using classroom-based research to gain insights into the patterns of learning for some of these topics; however, much work remains.

Challenges in obtaining validity evidence. At the start of the defining process, it may be easy for LP researchers to visualize the Upper Anchor, using guidance from standards, prior research, and years of experience. Yet when researchers search for evidence of this level in classrooms, they rarely are able to obtain sufficient data. For example, Gunckel et al. (chapter 4) found that only about 10 percent of students reason at the Upper Anchor. Given this low percentage, it is particularly difficult for researchers to test and refine this level. Similarly, Plummer (chapter 5) found that very few students reason about celestial motion phenomena at the Upper Anchor. Schwarz et al. (chapter 6) found little evidence of Upper Anchor modeling practices. To respond to the challenge posed by these results, which are not unique to these projects, the three research groups retained their Upper Anchor reasoning with little refinement but designed instructional materials to support students in achieving targeted understandings. However, it is still unclear if the support provided by the instructional materials will be sufficient, if the understanding described by the Upper Anchor is unrealistic, or if there are problems with student assessment.

The role of instruction. Researchers' philosophy on LP work and their approach to such work affect their opinions of science instruction and methods for validating levels. Some researchers view the Upper Anchor and the intermediate levels as examples that push beyond the level of learning students can achieve from status quo teaching. Many researchers think LPs can improve science instruction by reducing curricular fragmentation, by limiting the emphasis on facts, and by focusing on big ideas and explanatory models. The authors in the Defining Strand also point to limitations of the educational system in the United States. In particular, they focus on the limitations that affect their ability to validate ideas at the Upper Anchor. The lack of Upper Anchor knowledge and practice among students made clear that for this type of reasoning to be achieved by many students, instructional interventions are necessary. However, while researchers agree that status quo instruction is supporting Upper Anchor reasoning, questions regarding the role of instruction in LPs remain.

The validation of LPs requires specification about learning conditions necessary to achieve progress. Unless the assumption is status quo instruction, levels tend to represent ideal understanding, often achieved in ideal instructional conditions. Thus LPs validated in ideal learning conditions represent what we hope students *could do* given better instruction Most stakeholders in science find this assumption attractive since everyone supports the idea of increased learning achieved by efficient and effective instruction. The trade-off of defining a LP in this way, however, is the practical issue of recreating these ideal learning conditions in other contexts. If LP researchers show that students make progress under ideal learning conditions, how are we to replicate such learning conditions in other classrooms? The more requirements we set on using the LP in the classroom, the less likely it will be widely used

Another approach is to define and validate LP levels, in part, based on the reality of student learning in status quo instruction. When the empirical evidence reveals few or no students at the Upper Anchor, one reaction from researchers may be to align this level with the the most sophisticated learning observed. With this solution, enough evidence can be gathered to validate the Upper Anchor in a scientific domain of knowledge and practice. The approach is controversial, however, as many in science education argue that such an alignment sets low expectations, especially if more sophisticated understanding is achievable given better instructional support. The criticism of this approach is that the LP is defined by instructional reality rather than by what we hope could happen in classrooms.

CONCLUSION

Defining the LP domain and levels is part of the iterative design cycle of LP work. Researchers continuously refine their LPs through analysis of student performance, revision to assessments, and the use of instructional materials that support learning.

Classroom research and standards documents can help us construct hypothetical LPs. However, testing hypothetical LPs against actual student performance may pose various challenges that require researchers to refine the levels and shifts between levels. Thus LP researchers take a long-term perspective in using the iterative design cycle to advance their work. Yet it is this perspective that creates problems with their interactions with other education stakeholders. Standards and curriculum developers, for example, cannot afford to undergo year-after-year design cycles such as those used by LP researchers. Therefore, as LP researchers, we have to ask: When have we defined a LP sufficiently that it is ready for use? This is a critical question for LP researchers who hope that their LPs can be used to impact widespread change in science education.

LPs researchers trust that their LP frameworks and design products can be useful to many stakeholders in the science education community. The Using Strand of this book addresses the challenges LP researchers face in making those frameworks and products workable in that community. LP researchers must recognize that decisions made during the defining process (e.g., choice of domain and of grain size) have a significant effect on LPs. While there are no "right" choices, certain choices make the LP more or less useful. Explaining these choices advances the LP dialogue.

REFERENCES

Brown, A.L. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist, 52,* 399–413.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Brown, A. L., & Reeve, R.A. (1987). Bandwidths of competence: The role of supportive contexts in learning and development. In L. S. Liben (Ed.), *Development and learning: Conflict or congruence?* (pp. 173–216). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Kesidou, S., & Roseman, J.E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching, 39,* 522–549.

Mohan, L., Chen, J., & Anderson, C.W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46, 675–698.*

National Assessment Governing Board. (2008, September). *Science framework for the 2009 National Assessment of Educational Progress.* Retrieved fromhttp://nagb.org/publications/frameworks/science-09.pdf

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press.

National Science Foundation. (2009). *Research and Evaluation on Education in Science and Engineering program solicitation* (NSF 09–601). Retrieved from http://www.nsf.gov/pubs/2009/nsf09601/nsf09601.txt

National Science Foundation. (2010a). *Research and Evaluation on Education in Science and Engineering program solicitation* (NSF 10–586). Retrieved from http://www.nsf.gov/pubs/2010/nsf10586/nsf10586.txt

National Science Foundation. (2010b). *Discovery Research K-12* (NSF 10–610). Retrieved from http://www.nsf.gov/pubs/2010/nsf10610/nsf10610.txt

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. London, UK: Cambridge University Press.

ASSESSING LEARNING PROGRESSIONS

HUI JIN AND CHARLES W. ANDERSON

DEVELOPING ASSESSMENTS FOR A LEARNING PROGRESSION ON CARBON-TRANSFORMING PROCESSES IN SOCIO-ECOLOGICAL SYSTEMS

Learning progressions are descriptions of increasingly sophisticated ways of thinking about or understanding a topic (National Research Council [NRC], 2007). They provide promising frameworks for assessing students' understanding and learning. In our work to develop a learning progression for carbon-transforming processes, we involved participants from a wide age range (fourth grade through 11th grade) and from two countries (the United States [US] and China). We involved participants from a wide age range in order to develop a learning progression spanning naïve reasoning to sophisticated scientific reasoning. In addition, the study of how students progress under different cultural conditions will contribute to a better understanding of how students' learning is influenced by culture and schooling. The diversity of participants enabled us to collect extensive data for the development of the learning progression, but it also created special assessment challenges. In this chapter, we describe these assessment challenges and our responses.

Our research involves an inquiry process of drawing inferences about what students know and how they progress using evidence from students' performances. This process can be illustrated with the assessment triangle (Figure 1) that links three key elements: cognition, observation, and interpretation (NRC, 2001). Cognition refers to the models, theories, and beliefs about how students represent knowledge and develop competence in the subject domain; in the case of learning progression research, this is the learning progression framework. Observation includes the tasks or situations that allow researchers to elicit students' thinking. Interpretation comprises the methods and tools used to analyze the data and draw inferences, possibly leading to revision of the learning progression framework and assessment tasks.

When we began this research in 2004, our learning progression framework and initial assessments were based on our experience with and interpretation of assessment data from previous research (Anderson, Sheldon, & DuBay, 1990; Lee, Eichinger, Anderson, Berkheimer, & Blakeslee, 1993). In the process of eliciting students' understanding and interpreting students' responses to assessment questions, unexpected problems often emerged. We responded to these problems by revising or redesigning the assessments and the learning progression framework itself. Therefore, we adopted the approach of design-based research (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Collins,

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 151–181.

Joseph, & Bielaczyc, 2004; Edelson, 2002), which involves iterative cycles of assessment design and testing in which each testing cycle is an opportunity to collect data to inform subsequent assessment and learning progression framework design.

Our research goals include (1) developing a learning progression framework that describes students' increasingly sophisticated ways of reasoning about carbon-transforming processes and (2) developing associated assessments that are effective in eliciting a range of students' thinking about carbon-transforming processes. To reach these goals, our research was guided by the iterative assessment triangle (Figure 1), based on the work of the NRC (2001). The iterative assessment triangle represents the iterative cycles used to develop the learning progression framework and associated assessments.

As indicated in Figure 1, our research process has three phases: model of cognition (learning progression framework), observation, and interpretation.

- *Model of cognition* refers to the learning progression framework, which describes the progression of students' reasoning about carbon-transforming processes. It contains an Upper Anchor, the learning goal for high school graduates and, below that, qualitatively different achievement levels that reflect students' ideas. In our initial learning progression framework (Jin & Anderson, 2008; Mohan, Chen, & Anderson, 2009), the Upper Anchor was developed based on literature from environmental science and from national standards; hypotheses about lower achievement levels were based on our experience and reading of previous research. The learning progression framework is continuously revised and refined through the iterative cycles. We start each research cycle with the learning progression framework developed in the previous cycle. In each cycle, the learning progression framework guides our work in designing and revising assessments at the observation phase.

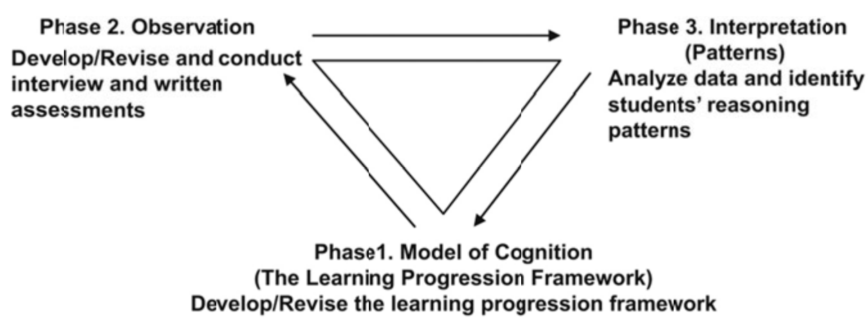


Figure 1. Research process: Iterative assessment triangle. Modified from the NRC's (2001) assessment triangle.

- *Observation* includes assessment design, assessment implementation, and data collection. We use both clinical interviews and written assessments to elicit students' ideas about carbon-transforming processes. The assessments are continuously revised and refined until they yield useful information for our understanding of students' ideas.
- In the *interpretation* phase, we develop coding rubrics to relate students' responses to the learning progression framework. Graduate students and post-doctoral fellows working on the research project apply the coding rubrics to interview and written data. Reliability checks for consistency of coding are performed during this process. The learning progression framework is continuously revised to resolve disagreements as well as other problems reported by the coders.

We conducted five research cycles during the past six years of the project. In this chapter we focus on the 2008–2009 cycle. We describe the preliminary model of cognition that we had at the beginning of the 2008–2009 cycle. Then we describe the assessment challenges we encountered and our responses to these challenges at the subsequent observation, interpretation, and model of cognition phases of the iterative assessment triangle.

PRELIMINARY MODEL OF COGNITION

An important goal of school science learning is to develop model-based reasoning—using models and theories as conceptual tools to analyze natural phenomena (NRC, 2007). After students graduate from high school, they will be responsible for making decisions about their personal lifestyles and about public policy. As consumers, voters, workers, and learners in society, their activities and decisions collectively impact environmental systems. In particular, as global climate change is becoming an increasingly serious issue, it is important for every citizen to understand how human energy consumption activities contribute to climate change.

Therefore, our learning goal for high school graduates is the use of scientific, model-based reasoning to explain how carbon-transforming processes contribute to global climate change. To elaborate on this goal, we reviewed relevant literature from environmental science (Long Term Ecological Research Network [LTER], 2007) as well as national standards documents (American Association for the Advancement of Science [AAAS], 1993; NRC, 1996). One important focus of environmental science is the supply-feedback chain between human society and environmental systems (LTER, 2007). National standards documents emphasize understanding both carbon-transforming processes and fundamental principles (i.e., matter conservation, energy conservation, and energy degradation). We incorporated these two ideas in the development of a Loop Diagram (Figure 2).

We study students' *accounts*—narratives that explain processes at multiple scales. The Loop Diagram, which is the Upper Anchor of our learning progression framework, represents scientific accounts that explain carbon-transforming processes at multiple scales, from atomic-molecular to global, with matter and energy conservation as constraints. It is constructed around three scientific elements—scale, matter, and energy. It also highlights two learning performances:

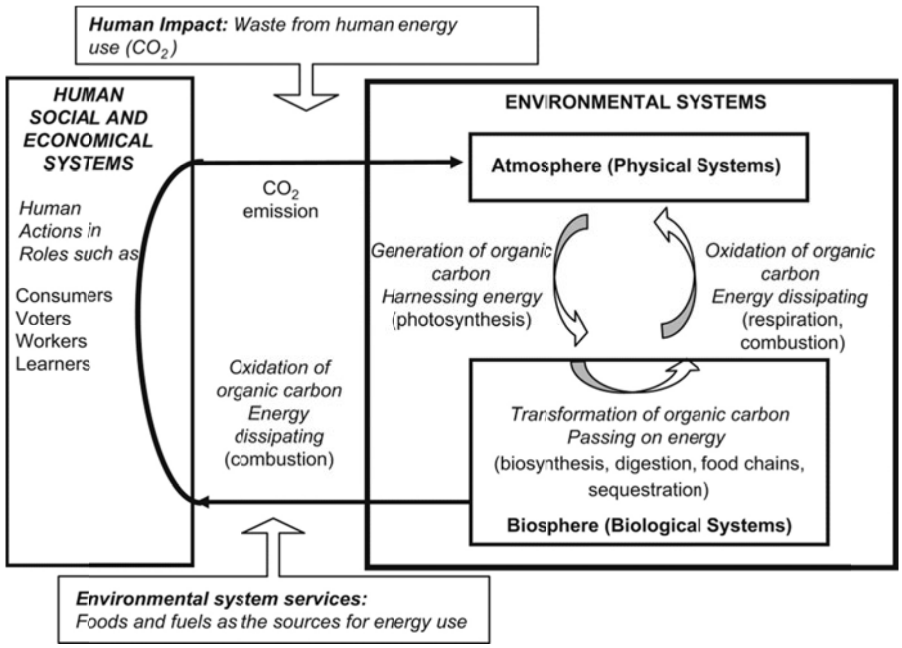


Figure 2. The Upper Anchor Loop Diagram. From "Developing a Multi-Year Learning Progression for Carbon Cycling in Socio-Ecological Systems," by L. Mohan, J. Chen, and C. W. Anderson, 2009, Journal for Research in Science Teaching, 46, p. 677. Copyright 2009 by Wiley Periodicals, Inc. Reproduced with permission of Wiley Periodicals, Inc. via Copyright Clearance Center.

- *Linking processes at multiple scales*
  Scientific models explain global fluxes of matter and energy as the cumulative effects of processes that take place at the atomic-molecular scale. There are three classes of carbon-transforming processes at the atomic-molecular scale: (a) organic carbon generation (photosynthesis), (b) organic carbon transformation (biosynthesis, digestion), and (c) organic carbon oxidation (cellular respiration, combustion). The cumulative effects of these atomic-molecular processes are the global carbon-transforming processes—carbon (matter) cycling and energy flow. Carbon recycles among the atmosphere, biosphere, and human socio-economic systems. That is, matter is transformed between the inorganic form (carbon dioxide) and the organic form (organic carbon-containing substances). Energy flows from sunlight to the biosphere and to the socio-economic systems, finally dissipating outside these systems as heat. These two global-scale processes of matter and energy explain humans' impact on environmental systems: people gain energy from foods and fuels from the biosphere; as we use energy, we emit carbon dioxide (and other greenhouse gases) into the atmosphere. Imbalance among these processes causes global climate change over time.

- *Constraining processes with matter and energy principles*
  The carbon-transforming processes are constrained by matter and energy principles—matter conservation, energy conservation, and energy degradation. In particular, two points should be noted. First, matter and energy are independently conserved in all physical and chemical changes. In other words, matter cannot be converted into energy and vice versa.[1] Second, whenever energy is transformed, heat is always released and cannot be recovered as usable energy.

The learning progression framework describes how students progress from their informal ways of reasoning towards the Upper Anchor. Each iteration of our learning progression framework describes students' progress using two parameters—*progress variables* and *levels of achievement* (Table 1). Progress variables are aspects of students' overall performance that differ for students at different levels of achievement. Students' learning performances relative to each progress variable can be ordered into different levels of achievement. Each revision in the iterative research cycles involves modifying one or both of these parameters.

We began the 2008–2009 cycle with some confidence in our general definitions of the levels of achievement. These levels are described in more detail by Gunckel, Mohan, Covitt, and Anderson (chapter 4) and in other papers (Jin & Anderson, 2008; Mohan, Chen, & Anderson, 2009). We found that more advanced students were able to produce accounts in terms of matter and energy, while many elementary and middle school students tended to rely on *force-dynamic reasoning*—a reasoning pattern identified by cognitive linguists. Research in linguistics and cognitive development indicates that people construct specific ways of reasoning as they learn their native languages. Cognitive linguists studying English grammar (Pinker, 2007; Talmy, 2000) and Chinese grammar (Dai, 2005; Lai & Chiang, 2003) suggest that both languages have implicit theories of cause and action: force-dynamic reasoning, which explains events in terms of actors, enablers, and results.

- Actors: Actors have internal goals and abilities/tendencies to take certain actions. Living actors such as plants and animals have internal self-serving goals and the ability to act toward those goals—to grow, maintain health, and move. Machines and flames also have the ability to act—to move or keep burning—but they need human actors to initiate the change—driving the car or igniting the flame. Dead plants and animals lose their ability to act and thus will change only by being acted on by actors or by "running down"—decaying.
- Enablers: Although actors have the ability to take certain actions, they need enablers to make changes happen. Each actor needs particular enablers. For example, people need air, water, and food to stay alive. Without them, people suffocate, dehydrate, or starve, and finally die. Similarly, plants need sunlight, water, soil, and air; flames need fuel, heat, and air; and so forth.
- Results: The actor uses enablers for certain actions or changes towards its natural tendency. The actions, or changes in general, cause certain results—the living or moving actor fulfills its goal or the dead actor deteriorates.

Scientific accounts share this general framework but with the meanings of each part substantially altered. Scientific accounts treat both actors and enablers as chemical entities, the composition and structure of which are explained in terms of matter and energy. The interactions between actors and enablers are not about actions and results, but about matter transformation and energy transformation at three scales—atomic-molecular, macroscopic, and global. In brief, scientific accounts are constructed around scientific elements—matter and energy at multiple scales. Students who have rich school science learning experiences may use matter and energy to construct accounts, but their accounts often indicate misconceptions. (This is Level 3 of our beginning learning progression framework, as shown in Table 1.)

The four levels of achievement are:

- Level 4 is defined as scientific accounts that are built upon model-based reasoning—tracing matter and energy within and across carbon-transforming processes at multiple scales. This model-based reasoning is illustrated in the Loop Diagram (Figure 2).
- Level 3 is defined as school science accounts that involve atoms, molecules, and energy forms but do not successfully conserve matter or energy.
- Level 2 is defined as force-dynamic accounts with hidden mechanisms that explain environmental events in terms of hidden processes and mechanisms but still focus on actors, enablers, and results.
- Level 1 is defined as macroscopic force-dynamic accounts that describe environmental events in terms of actors, enablers, and results.

We were less confident about the other parameter in our learning progression framework: progress variables that we used to structure our detailed descriptions of the different levels of achievement. Students' understanding is usually reflected in multiple dimensions of their learning performances, all of which can be used as progress variables. The Loop Diagram suggests two orthogonal types of progress variables:

- Carbon-transforming processes that generate (photosynthesis), transform (digestion and biosynthesis), and oxidize (cellular respiration and combustion) organic matter.
- Scientific elements—scale, matter, and energy.

We used these progress variables as the basis for the learning progression framework represented in Table 1. The learning progression framework organizes descriptions of each level of achievement first around the three carbon-transforming processes: organic carbon generation, organic carbon transformation, and organic carbon oxidation. Under each process, there are two subordinate progress variables: matter and energy. We found that students' reasoning about scale is implicit in their reasoning about matter and energy, so we did not make scale a separate progress variable. More detailed versions of Table 1, with individual learning performances for each level, can be found in other project papers (Jin & Anderson, 2008; Mohan, Chen, & Anderson, 2009).

Table 1. Learning Progression Framework at the Beginning of the 2008–2009 Research Cycle.

Levels of Achievement	*Progress Variables*					
	Organic Carbon Generation		*Organic Carbon Transformation*		*Organic Carbon Oxidation*	
	Matter	*Energy*	*Matter*	*Energy*	*Matter*	*Energy*
Level 4. Accounts based on model-based reasoning	Explain plant growth in terms of organic matter generation in photosynthesis	Explain plant growth in terms of energy transformation in photosynthesis	Explain human body growth in terms of organic matter transformation in digestion and biosynthesis	Explain human body growth in terms of energy transformation in digestion and biosynthesis	Explain burning, running, and decay in terms of organic matter oxidation in combustion and cellular respiration	Explain burning, running, and decay in terms of energy transformation in combustion and cellular respiration
Level 3. School science accounts	Explain plant growth in terms of changes involving glucose, sugar, or other familiar organic molecules May use matter-energy conversion for reasoning	Explain plant growth in terms of changes involving energy forms May use matter-energy conversion for reasoning	Explain human body growth in terms of changes involving glucose, sugar, or other familiar organic molecules May use matter-energy conversion for reasoning	Explain human body growth in terms of changes involving energy forms May use matter-energy conversion for reasoning	Explain burning, running, and decay in terms of changes involving oxygen and/or familiar organic molecules May use matter-energy conversion for reasoning	Explain burning, running, and decay in terms of changes involving energy forms May use matter-energy conversion for reasoning
Level 2. Force-dynamic accounts with hidden mechanisms	Explain plant growth in terms of hidden processes (e.g., making food)	Explain plant growth in terms of triggering event (e.g., energy powers plant growth)	Explain human body growth in terms of hidden processes (e.g., food is broken down; useful stuff is extracted out of food)	Explain human body growth in terms of trigger event (e.g., energy powers human growth)	Explain burning and running in terms of hidden processes (e.g., food becomes sweat). Explain decay in terms of hidden processes (e.g., wood changes into dirt)	Explain burning and running in terms of triggering events (e.g., energy powers flame; energy powers running). Explain decay as a process of losing energy or power
Level 1. Macroscopic force-dynamic accounts	Explain plant growth in terms of macroscopic force-dynamic reasoning (e.g., plants use enablers such as air, water, soil, and sunlight to grow bigger) Do not specify any invisible processes Do not recognize the role of matter or energy in plant growth		Explain human body growth in terms of macroscopic force-dynamic reasoning (e.g., humans use enablers such as air, foods, and water to grow bigger) Do not specify any invisible processes Do not recognize the role of matter or energy in human body growth		Explain burning, running, and decay in terms of macroscopic force-dynamic reasoning (e.g., flame needs match to support it; dead things decay when getting old) Do not specify any invisible processes Do not recognize the role of matter or energy in burning, running, or decay	

OVERVIEW OF THE 2008–2009 RESEARCH CYCLE

We began the 2008–2009 cycle with the learning progression framework described above. During the year, we went through each phase of the cycle:

- *Observation:* We conducted interviews and written tests in the US and China to elicit students' accounts of carbon-transforming processes. We conducted interviews and gave tests both before and after an instructional intervention (described in Gunckel et al., this volume) for the US students. We did not conduct an instructional intervention in China. In China, 24 students (8 students from each school level) participated in interviews and 300 students (100 students from each school level) responded to our written tests. These students are from urban and rural schools located in Southeast China. In the US, 24 students (8 students from each school level) participated in the pre-interviews and post-interviews; 527 students participated in the pre-tests (91 elementary school students, 214 middle school students, and 222 high school students); and 543 students participated in the post-tests (125 elementary students, 211 middle school students, and 207 high school students). These students are from suburban and rural schools in a Midwest state. Since this chapter focuses on assessment development, we note only that the intervention added useful variability to our US sample.

- *Interpretation:* We developed approaches to coding students' accounts that revealed what we saw as significant patterns associated with differences in students' proficiency and culture.
- *Model of cognition:* We revised the learning progression framework in light of what we had learned.

During this research cycle, two assessment challenges became increasingly crucial. The first is a challenge for the observation phase: Our participants came from a wide range of ages (from fourth to 11th grade) and from US and Chinese cultures. Assessment questions that make sense to one age and one culture may be understood quite differently—or not understood at all—by students of another age and from another culture. How do we develop assessments that are effective in eliciting accounts from all students that best represent their ways of reasoning? In particular, how can we ask elementary students about the processes represented in the Loop Diagram when the key scientific elements of the model—matter, energy, and scale—are invisible to them?

The second challenge comes at the interpretation and model of cognition phases: Students' accounts differ in many ways, but which differences are really important? How can we define levels of achievement and progress variables that provide valid and parsimonious descriptions of students' understanding of carbon-transforming processes at different ages and in different cultures?

In the next two sections we discuss these assessment challenges in greater depth and describe how we responded to them during the 2008–2009 research cycle.

RESPONDING TO THE ASSESSMENT CHALLENGE AT THE OBSERVATION PHASE

In this section we first elaborate on the nature of the challenge that we faced. We then describe our response to the challenge, as well as the issues we need to deal with in the future.

The Challenge: Eliciting Accounts from All Students

In our earlier research cycles, we used questions focusing on matter and energy at multiple scales to elicit students' accounts. We found that although some middle and high school students were able to understand questions that asked about matter and energy at different scales, younger students were often confused by the questions and therefore provided irrelevant or "I don't know" type responses. Below is an example. Our intent was to find out how well students understood the roles of matter transformation (especially from carbon dioxide and water to glucose and oxygen) and energy transformation (from sunlight to chemical potential energy) in plant growth.

Episode 1. Corn plants growing in sunlight

(An interview with a US seventh grader in our previous studies)

Interviewer: What are the materials you identified in this event [corn plants growing in sunlight]?

May: Water, soil, and sunlight.

Interviewer: How do they change?

May: They change by they… give their sources to the corns for them to grow, the water, soil, and sunlight. They give the corn water, and soil, and … that help the corn plants grow.

Interviewer: Does this event change the air?

May: I think it does change the air, because if there are more plants growing. It gives the air more… It refreshes the air. It makes the air smells like corn. And rain also refreshes the air. It washes everything away.

Interviewer: What are the things in the air that do not change?

May: The thing in the air that do not change is… (Silence) You mean with the corn plants grow or just in general?

Interviewer: When the corn plants are growing, what does not change in the air?

May: Oh. It does not change the color. It does not change the air by making it change color.

In the exchange above, the interviewer recognized that the student's account did not mention carbon dioxide or oxygen. In keeping with the interview protocol, the interviewer asked probing questions about the role of air: "Does this event change the air?" "What are the things in the air that do not change?" "When the corn plants are growing what does not change in the air?" However, the student did not recognize that air is required for the plant to grow. When asked to explain changes in air, the student focused on changes in the quality of the air (refresh the air) and observable properties of the air such as color and smell. She may not share the interviewer's assumptions that air is a mix of gases and that the matter of air changes because the tree uses it. Rather than enabling the student to elaborate on her own ideas about how and why corn plants grow, the interview protocol diverted both interviewer and student to a "dead end"—a series of questions about air that the student did not connect with the initial question about how plants grow.

Although we learned a lot about students' reasoning from our interviews and written assessments, our data also showed us that questions designed to elicit students' ideas about scale, matter, and energy did not effectively elicit force-dynamic accounts from younger students. Thus the assessment challenge we faced was how to collect data on younger students' understanding that we could connect to our Upper Anchor—the Loop Diagram (Figure 2). In particular, what questions can elicit accounts from younger students that provide evidence of their informal reasoning? How can we construct interview protocols and written assessments that are effective in eliciting accounts from all students?

Responding to the Challenge: Linking Processes and Alternate Forms of Questions

We responded to this assessment challenge in two ways. First, we organized our interviews and tests around *linking processes* that were familiar to students of all ages. Second, we developed *alternate forms of questions* for students from different age groups.

Linking processes. We could not organize our interviews and written assessments around the atomic-molecular and global carbon-transforming processes in the Loop Diagram (Figure 2), since they are invisible to many students, especially younger students. Therefore, our first step in the revision process was to organize the assessments around macroscopic processes that are familiar to all students. Figure 3 shows our revised version of the Loop Diagram. It shows the same relationships among organic carbon generation, transformation, and oxidation as Figure 2, but it is organized around familiar macroscopic processes.

In Figure 3, processes at three scales—atomic-molecular, macroscopic, and global—are linked. The atomic-molecular processes in the dashed boxes (photosynthesis, digestion and biosynthesis, cellular respiration, and combustion) explain the macro-processes in the grey boxes (plant growth, animal growth, weight loss, using electrical appliances, driving vehicles, flame burning, etc.). These processes are connected by matter transformation (straight arrows) and energy transformation (wavy arrows): in photosynthesis, organic carbon-containing substances are generated from carbon dioxide and water, and light energy transforms into chemical potential energy; organic carbon-containing substances transform, and chemical potential energy is passed on in biosynthesis and digestion; organic carbon-containing substances are oxidized into carbon dioxide and water, and chemical potential energy is released in cellular respiration and combustion. The atomic-molecular processes are embedded in two global scale processes: (1) carbon (matter) cycle—carbon moves from atmosphere to biosphere and human socio-economic systems and then back to atmosphere—and (2) energy flow—energy moves from light to biosphere and then to human socio-economic systems with heat dissipation.

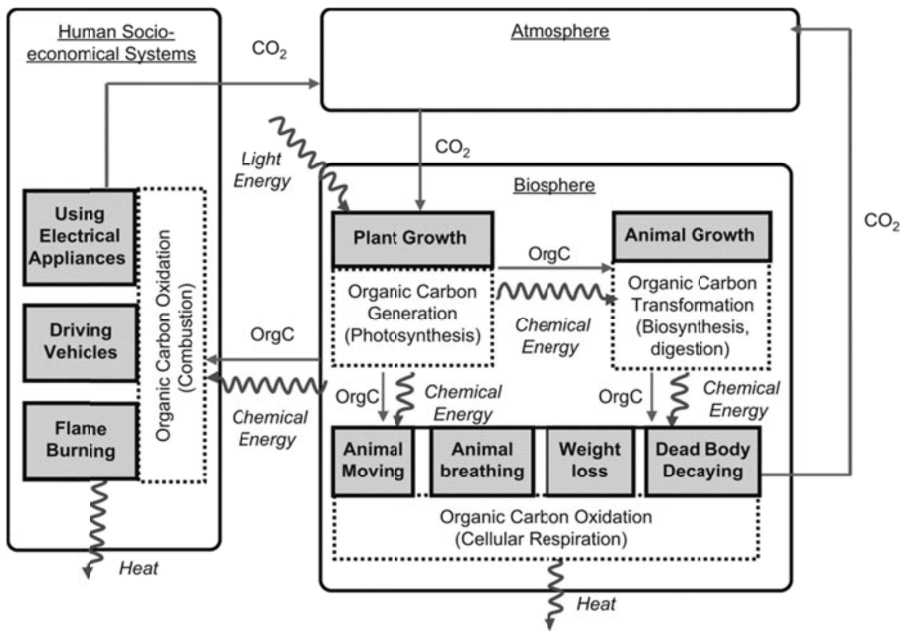


Figure 3. Linking processes in the Loop Diagram.

Students across grade levels have rich experience with the macro-processes. Students may not understand photosynthesis, but they know that plants grow and have special needs such as sunlight, water, and air. They may not understand combustion of fossil fuels as a chemical process, but they have experience observing their parents refill cars with gasoline and burn propane for barbecues. They may not understand cellular respiration, but they are likely to have experienced that running causes sweating and fatigue. Hence, we organized both interviews and written assessments at all levels around the same set of macro-processes that are familiar to all of the students in our research. By focusing on these *linking processes*, we were able to design assessments that elicited more detailed accounts from all students while enabling us to compare and contrast accounts at all levels of achievement.

Alternate forms of questions. While students from different age groups have rich experience with the macro-processes, their knowledge of processes at the atomic-molecular and global scales varies. Younger students learn very little about the atomic-molecular and global processes in school; however, middle and high school students should have some knowledge of these processes. To ensure that the assessments are effective in eliciting accounts from all students, we designed alternate forms of questions for students at different school levels. In particular, we developed branching-structured interviews and designed item pairs for the written assessments.

Branching-structured interviews. Our revised interview protocols have segments for each of the macro-processes in Figure 3. For each macro-process, we start with a set of general questions—questions that use everyday language to ask about familiar phenomena. These general questions can be understood by younger students and yet allow more advanced students to provide brief accounts about scale, matter, and energy. If students' responses to the general questions indicate some ideas about scale, matter, and energy, we ask follow-up questions that are designed to elicit more detailed higher-level accounts.

As explained earlier, lower-level students' accounts are constructed around force-dynamic reasoning, which contains three elements—actors, enablers, and results. Scientific accounts share this general framework but treat actors and enablers as matter and energy and explain the events in terms of matter transformation and energy transformation at multiple scales. Hence, we constructed our general questions around this shared framework to elicit students' ideas that may include elements of both scientific and force-dynamic reasoning. Take tree growth as an example. The major general questions are:

- What does the tree need in order to grow?
- You said that the tree needs sunlight/water/air/soil to grow. Then how does sunlight/water/air/soil help the tree grow?
- Where does sunlight/water/air/soil go when it is used by the tree?
- Do you think that sunlight/water/air/soil will change into other things inside the tree's body? Why?
- The tree gets heavier as it grows. How does that happen?

Students can interpret these questions either as questions about transformations of matter and energy or as questions about an actor (the tree) and its enablers. Thus these questions allow students to provide both scale-matter-energy and force-dynamic accounts. If the students' responses to the general questions indicate more sophisticated understanding, we ask a set of higher-level questions to elicit more detailed accounts about matter, energy, and scale. Some examples of higher-level questions about tree growth are:

- Do you think the tree's body structure is made from things outside of the tree? If yes, what are those things? How do these things change into the tree's body structure?
- If the student mentions glucose/starch/cellulose/carbohydrates, ask: Do you think the molecules you mentioned contain carbon atoms? If yes, where do the carbon atoms come from? Where are the molecules in the tree's body?
- You said that sunlight provides energy for the tree to grow. Where does that energy go when it is used by the tree? Do you think it is used up or becomes other things?
- If the student talks about CO_2-O_2 exchange, ask: You said that the tree needs carbon dioxide and breathes out oxygen. Where do the carbon atoms of CO_2 go?

These questions investigate how students link the macro-processes to atomic-molecular and global processes.

We use two examples to show how the branching-structured interview elicits both lower-level and higher-level accounts. Episode 2 is from an interview with a US fourth grader, Steve.

Episode 2. Tree Growth (general questions and responses)

(Pre-interview with a US fourth grader)

Interviewer: What does the tree need in order to grow?

Steve: Sun, water, soil, and that's it, I think.

Interviewer: You said that a tree needs sunlight, water. Do you think that these things help the tree to grow in the same way? In other words, are they alike or different?

Steve: The water helps it grow bigger and the sunlight, it needs light just like us to grow, and the soil, that's where it originally lived.

Interviewer: What happens to the sunlight inside of the tree?

Steve: (silence)

Interviewer: How about water? What happens to the water inside the tree?

Steve: It sucks into the roots and then it goes up, so it can make the leaves and the branches grow.

The interviewer's questions ask about the changes to water/sunlight/soil inside the tree. The accounts Steve provides in response to these questions are basically force-dynamic in nature. For example, his explanation of how water helps the tree grow is: "It sucks into the roots and then it goes up, so it can make the leaves and the branches grow." In other words, as long as the water goes into the tree's body, it makes the leaves and branches grow. This explanation indicates that Steve does not recognize the change of matter—water changing into part of the tree's body structure in photosynthesis. Instead, he relies on force-dynamic reasoning and treats water as an enabler that allows the tree to achieve its destined result—growth in this case—through an unspecified mechanism. As is typical of force-dynamic accounts, Steve's accounts are vague about internal mechanisms; although he is sure that the tree needs sunlight, he doesn't know what the tree does with this enabler. The evidence from this episode shows that Steve relies on macroscopic force-dynamic reasoning to explain the event of tree growth. His explanation of tree growth is about how the actor—the tree—uses enablers such as water, sunlight, and soil to accomplish its purpose of growing.

The branching-structured interview is also effective for eliciting higher-level accounts. Episode 3 is from an interview with a US eighth grader, Sue.

Episode 3. Tree Growth (general questions and responses)

(Pre-interview with a US eighth grader)

Interviewer: What does the tree need in order to grow?

Sue: Nutrients, water, sunlight, things to make it do photosynthesis.

Interviewer: So what do you mean by photosynthesis?

Sue: Like reproduce and get food and be able to produce carbon back or carbon or I mean oxygen. Sorry.

Interviewer: So you said the tree needs sunlight. So, how does the sunlight help the tree to grow?

Sue: The sunlight like gives it energy and things like nutrients and make it, so it grows.

Sue's responses to the general questions contain important elements of scientific accounts focusing on scale, matter, and energy as she talked about the process of photosynthesis and related it to carbon. However, her responses to these general questions do not provide enough information about her understanding of the atomic-molecular processes or matter and energy transformations.

Episode 4 shows how the interviewer used higher-level questions to elicit Sue's understanding of matter.

Episode 4. Tree Growth (follow-up questions about matter and responses)

(Pre-interview with a US eighth grader)

Sue: The carbon dioxide like makes it breathe. Like how we breathe in but they produce oxygen from the carbon dioxide.

Interviewer: How can carbon dioxide change into oxygen?

Sue: By the different like it's – it goes through like the system of like the tree or through the system of like a body.

Interviewer: So, if you compare carbon dioxide and oxygen, carbon dioxide has a carbon atom in it, right? Oxygen does not have that. So, how can't it have it?

Sue: Because like the things in carbon dioxide, it gets like – like during the process, it gets used as energy or used as different things to make the tree grow and to make it produce oxygen.

Interviewer: You mean the carbon atom of the carbon dioxide becomes the energy? Is that what you mean?

Sue: Yes. And carbon gets used for other things like carbon can go back into a different cycle like air. And then back into another cycle.

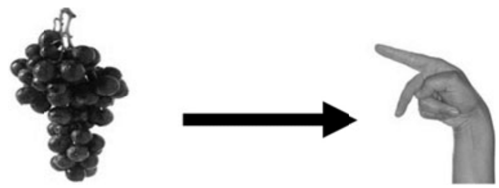
The interviewer's probes focus on a key difference between Level 3 and Level 4 accounts: Does Sue's account conserve matter, indicating that she recognizes chemical changes like photosynthesis cannot create or destroy carbon atoms (Level 4), or does she have less scientific understanding about how materials can change into other materials, or perhaps into energy (Level 3)? Sue's responses indicate that she could not account for the processes at an atomic-molecular scale. She suggests that the carbon in carbon dioxide "gets used as energy" and confirms this idea in response to a question from the interviewer. Although Sue attempted to use scale, matter, and energy to construct explanations of tree growth, the interview shows that she reasoned at an intermediate level (Level 3) rather than the scientific Upper Anchor.

Compared with the initial interview protocol, this branching-structured interview protocol is more successful in eliciting accounts from students with diverse science backgrounds. The general questions allow younger students to provide their informal accounts about macro-processes and also provide more advanced students with the opportunity to address scale, matter, and energy. By asking the general questions, the interviewer is able to find indicators of higher-level understanding and to decide whether it is necessary to ask higher-level follow-up questions to elicit more detailed accounts.

Item pairs in written assessments. During the earlier research cycles, we developed a set of open-ended items. We have continuously revised and refined these items using feedback from analysis of student responses. We have found it almost impossible, however, to design items that elicit good responses from both Level 1 and Level 4 students. Items worded to demand specific details about matter and energy transformations elicit guesses or "I don't know" from lower-level students; vaguely worded items elicit correct answers from upper-level students, but these correct answers lack the detail necessary to judge whether the responses are at Level 3 or Level 4.

Our solution to this challenge is to design item pairs. Each item pair contains two items that are about the same event or similar events but use different ways of asking questions in order to elicit both force-dynamic accounts and matter-energy-scale accounts. Some of these item pairs are open-ended items. Others are two-tier multiple-choice item pairs that require the student to choose and then explain.

For example, Figure 4 shows the "grape and finger movement" item. It is an open-ended item asking how a glucose molecule changes to help body movement. In the earlier research cycles, we asked this question to both middle and high school students. This item proved effective in diagnosing whether and how students' accounts conserve matter and energy in cellular respiration. Below are the responses from a US ninth grader. The student's account represents an attempt to conserve matter and energy; however, instead of conserving matter and energy separately, the account uses matter-energy conversion to explain how glucose helps the finger move (characteristic of Level 3).



The grape you eat can help you move your little finger.

a. Please describe how one glucose molecule from the grape provides energy to move your little finger. Tell as much as you can about any biological and chemical processes involved in this event.

The glucose molecule is converted to chemical energy in your body. Then your body uses that energy to make ATP, which is then used for cellular work, which allows you to move.

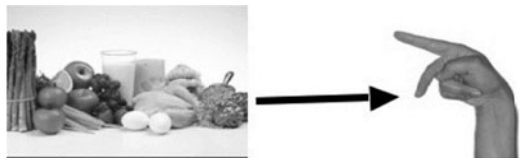
b. Do you think the SAME glucose molecule can also help you to maintain your body temperature, when it is used to provide energy to move your finger? Please explain your answer.

Yes, because in order to maintain your body temperature, your cells would need to work, and the cells get their working energy from ATP, which is converted by glucose.

Figure 4. Grape and finger movement item with example of a US ninth grade student's response (post-test).

Although the "grape and finger movement" item was effective in identifying and distinguishing between more sophisticated accounts, middle school students did not understand it. The major reason is that the two questions included in this item are posed at the atomic-molecular scale, which is usually invisible to younger students. When we used this item with middle school students, we received a lot of "I don't know" type answers. Some students even doubted the meaningfulness of asking this kind of question. For example, a student replied: "Dude, I'm only 14 and I didn't understand the ? [question]."

Hence, in the 2008–2009 research cycle we developed a corresponding elementary/middle school item—the "food and finger movement" item (Figure 5). It was revised from the "grape and finger movement" item. Instead of asking about a specific food substance—glucose in the grape—the item asks about food in general. It uses informal language that can be understood by younger students.



How do you think the foods you eat can help you move your little finger?

I think the foods help move my finger because it gives off energy that help you move and communicate. When someone starts to starve, there [sic] body gets very tired and weak. This happens because there [sic] body is not getting the nutrients it needs from the food that you eat.

Figure 5. Food and finger movement item with an example of a US fourth grader's response (pre-test).

In Figure 5, the student's response is constructed around the macroscopic actor (people), enablers (energy and/or nutrients from foods), and results (achievement of the goal of moving the finger). Thus it indicates force-dynamic reasoning. Although the student used the word energy in her explanation, she did not distinguish between energy, nutrients, and foods in general.

We found that we could also use a complementary strategy to develop higher-level versions of items that initially were not specific enough for upper-level students. Open-ended questions such as the one in Figure 6 use everyday language and do not require students to link or constrain processes. We found that such questions are not effective in identifying and distinguishing between accounts at higher levels.

In response to the item in Figure 6, the high school student provided a correct explanation of why plants need light to survive. However, since the question did not require the student to specify how light energy is transformed in the process of photosynthesis, the student did not provide any details. As a result, the student's account does not provide enough information for us to tell whether energy is being conserved in explaining photosynthesis. Her response could be either at Level 3—incorrect description of energy transformation—or Level 4—correct description of energy transformation.

In order to obtain more detailed accounts, we revised this item into a two-tier multiple-choice item pair. The first tier is a multiple-choice question while the second tier requires students to justify their choices. For the first tier, the options are characteristic accounts developed based on students' responses to the open-ended question that we had used previously. The elementary/middle school item contains options that represent lower-level accounts, while options in the high school item represent higher-level accounts about scale, matter, and energy. Our data indicate that these two-tier multiple-choice item pairs are more effective than the corresponding initial version in diagnosing and distinguishing between accounts at different levels. Figures 7 and 8 show how the revised versions of the "light for plants" item assess and distinguish between accounts at different levels.

Do you think plants need light to survive? Circle one: Yes No

If your answer is "yes," please explain why plants need light AND where the light energy goes after it is used by plants. If your answer is "no," please explain why plants can live without light.

Yes, because without light they can't perform photosynthesis and make food. With light energy they make food.

Figure 6. Light for plants item (initial version) with an example of a US ninth grader's response (pre-test).

Do you think plants need light to live? Please choose the best answer(s) from the list below.

a. Not all plants need light to live.

b. Light warms the plants.

c. Without light, plants will die in darkness.

d. Light helps plants to be healthy.

e. Light helps plants to make food.

f. Light helps plants breathe.

Please explain why you think these are the best two answers.

Choice: b. c.

Explanation: A plant needs light to live because when it is dark it's colder and they will get to [sic] cold and die.

Figure 7. Light for plants item (elementary/middle school version) with an example of a US fourth grader's response (pre-test).

In the item in Figure 7, the options of the first tier represent two levels of reasoning. Options a, b, c, d, and f are macroscopic force-dynamic accounts (Level 1). Choice a does not recognize that all plants need sunlight. Choices b, c, d, and f explain why plants need sunlight in terms of perceptions, including warmth, darkness, health, and breathing. These accounts do not mention any invisible processes. The student chose b and c. Both his choices and explanations indicate macroscopic force-dynamic reasoning. Option e is a more sophisticated account. It links the macro-process to the invisible process of "making food" and therefore indicates Level 2 reasoning. Although option e is more advanced than the other options, it does not address details about scale, matter, and energy. So it is not effective in identifying the level of more sophisticated accounts. Therefore, a high school version of the item pair (Figure 8) was developed.

Sunlight helps plants to grow. Where does light energy go when it is used by plants? Please choose the ONE answer that you think is best.

a. The light energy is converted into glucose of the plants.

b. The light energy is converted into ATP in the plants.

c. The light energy is used up to power the process of photosynthesis.

d. The light energy becomes chemical bond energy.

e. The light energy does not go into the plants' body.

Please explain why you think that the answer you chose is better than the others. (If you think some of the other answers are also partially right, please explain that, too.)

Choice: a.

Explanation: Because the plants take the light energy and convert into glucose. After that, glucose units combine to make starches that the plant can use to function. Starches are fatal [sic] for plant survival.

Figure 8. Light for plants item (high school version) with an example of a US 10th grader's response (post-test).

The item in Figure 8 contains options describing how energy and matter change in the atomic-molecular process of photosynthesis. Options a and b use matter-energy conversion for reasoning. Option c treats light energy as the power that triggers the process of photosynthesis; this is correct, but the energy is not used up as this option suggests. Options a, b, and c represent the common misconceptions identified from previous research cycles. They are at Level 3. Option d is the scientific account that successfully traces energy in photosynthesis (Level 4). Option e does not recognize light energy as being related to any hidden process involved in tree growth. Students who choose this option reason at levels lower than Level 3, although this version of the item does not effectively distinguish between Level 1 and Level 2 accounts.

In the example response, both the student's choice and justification indicate an attempt to trace energy, but it is not clear that she distinguishes between chemical potential energy and matter that has chemical potential energy. Instead of conserving matter and energy separately, her account explains the event in terms of matter-energy conversion—light energy is converted into molecules (glucose and starches). Therefore, her account is at Level 3.

We have not finished revising our interviews and written assessments, however. The general strategies of constructing branching-structured interviews and item pairs are difficult to execute in practice, and they leave many aspects of students' accounts insufficiently explored, such as the basis for their beliefs and the connections that they make between accounts of different processes. We continue to examine and revise our assessments, based on both the quality of students' responses and the statistical indicators of item quality.

RESPONDING TO THE ASSESSMENT CHALLENGE AT THE INTERPRETATION AND MODEL OF COGNITION PHASES

In this section, we describe the assessment challenge we encountered at the interpretation and model of cognition phases and our responses to that challenge. We also discuss the persistent issues that need to be addressed.

The Challenge: Describing and Comparing the Development of Students' Accounts in Meaningful Ways

Currently, most empirical studies of learning progressions are conducted in one country. In our research, we used a learning progression framework to compare students' understandings under different cultural and educational conditions. We believe that this investigation will allow us to better understand how culture impacts students' learning. In the 2008–2009 research cycle, we involved students from two countries (the US and China). These students use different languages for reasoning, have different science backgrounds, and are exposed to different educational approaches. Although the interview and written assessments effectively elicited accounts from both US and Chinese students, we encountered an assessment challenge as we were interpreting the data and revising the learning progression framework to include both US and Chinese students' accounts.

In our earlier research cycles, we constructed the learning progression framework (Table 1) around progress variables based upon scientific processes (organic carbon generation, organic carbon transformation, and organic carbon oxidation) and scientific elements (matter and energy). Based on this framework, we developed detailed rubrics to code students' responses to written assessments. This process of coding led to both conceptual and empirical difficulties.

Conceptually, the descriptions of Level 1 and 2 reasoning focused on what was not there (scientific concepts of matter and energy) rather than what was there (force-dynamic accounts of actors, enablers, and results). In other words, the progress variables—matter and energy—do not capture students' ways of informal reasoning at the lower levels. They are not valid progress variables that allow us to identify and describe younger students' characteristic ways of reasoning.

There were also empirical questions about the usefulness of coding matter and energy separately. The US written assessment data show that the correlation between students' achievement on the matter and energy progress variables was .96, indicating that our separate codes for matter and energy were largely redundant (Choi, Lee, & Draney, 2009; Mohan, Chen, Baek, Choi, & Lee, 2009). In other words, the matter and energy columns of Table 1 do not really describe separate progress variables.

At the same time, our US-China written assessments showed that rubrics based on the learning progression levels seemed to work better for US students than for Chinese students. For example, step thresholds—the level of proficiency at which a student has a 50% chance of being coded at one level of the learning progression framework versus the level above—were generally consistent across items for US students. This was much less true for Chinese students. Many Chinese students received higher-level codes—Levels 3 and 4—on some items but not on others. Thus the factors that made a response easy or difficult for Chinese students were different from those we found in the coding of learning-progression-based levels in US responses.

There were no strong correlations between item difficulty and specific processes or the matter and energy progress variables (Chen, Anderson, & Jin, 2009). That is, the Chinese students did not perform consistently better or worse than US students on matter items or energy items, or photosynthesis items or combustion items. The pattern of different difficulties seemed to be associated with characteristic(s) of the items that were not included in our 2008 progress variables.

Our US-China interview study (Jin, Zhan, & Anderson, 2009) suggested a possible explanation for the problems we encountered when analyzing the Chinese written data. In the interview study, we found that although some Chinese students were able to name scientific terms when explaining the events, they relied on relatively lower-level reasoning in their accounts, as illustrated in Episode 5. This type of performance was apparent only in the Chinese interviews.

Episode 5. Tree Growth

(Interview with a Chinese seventh grader)

Interviewer: You said that the tree inhales carbon dioxide and produces oxygen. Could you explain how carbon dioxide changes into oxygen?

Peng: Water.

Interviewer: How can water help the carbon dioxide to change into oxygen?

Peng: Chemical reaction.

Interviewer: Could you explain what this chemical reaction is?

Peng: Probably water plus carbon dioxide and become $C_6H_{12}O_6$. I don't know.

Interviewer: You mean sugar?

Peng: I don't know. How can the tree have sugar?

Interviewer: Do you think the tree contains sugar?

Peng: I don't think so.

Interviewer: Let's see this picture. The tree grew from a small plant into a big tree. Its mass increased a lot. Do you agree?

Peng: Yes.

Interviewer: So, where did the increased mass come from?

Peng: Water.

Interviewer: Is there anything else?

Peng: And nutrients from soil.

In Peng's first three responses, she appears to provide a sophisticated (Level 4) chemical explanation that traces matter through photosynthesis, but then in the remainder of her responses, she reverts to what seems to be a much less sophisticated (Level 2) explanation. How can we capture this kind of performance in our interpretations of students' accounts?

One hypothesis is that the scientific words the students used might influence the coders' decisions. In our previous study (Chen et al., 2009), Chinese students were rated at higher levels for some items because these items were designed in ways that cued students to use scientific vocabulary. However, our 2008 framework, and codes based on that framework, did not distinguish between scientific vocabulary and other aspects of students' accounts. As a result, many Chinese responses were rated at higher levels, not because the students were able to reason at higher levels, but because they were able to recite scientific words. This led us to think about alternative progress variables that might be more effective in understanding and comparing US and Chinese students' accounts.

Responding to the Challenge: Explaining and Naming as Progress Variables

Rather than treating the highly correlated scientific elements of accounts—matter and energy—as progress variables, we began to explore progress variables focused on performance elements of accounts, which we labeled *explaining* and *naming*.

The explaining progress variable is about the nature of the accounts—the specific reasoning that students use to explain why and how the macro-processes happen. The explanations are always constructed around different types of causal reasoning. As described above, younger students tend to rely on force-dynamic reasoning that explains the macro-processes in terms of actor, enabler, and result. More advanced students begin to pay attention to scale, matter, and energy and to explain macro-processes in terms of changes of matter and energy in invisible processes. Thus we combined the separate matter and energy progress variables into a single explaining progress variable. At higher levels, the explaining progress variable includes students' ideas about both matter and energy.

We used the 2008–2009 interview data to construct a new progress variable that we called naming. This progress variable focuses specifically on the vocabulary students use—from words that describe actors and enablers in informal terms to more formal scientific names for substances, forms of energy, and carbon-transforming processes. The revised version of the learning progression framework using explaining and naming progress variables is represented in Table 2.

The levels of the explaining progress variable are quite similar to the levels of the initial learning progression framework described above. They describe a line of development from force-dynamic to scientific reasoning as students learn to trace matter and energy through processes at multiple scales. The naming progress variable describes students' use of specific words and/or phrases in their accounts. Logically, accounts at different explaining levels build upon different sets of words. For example, accounts at explaining Level 1 are constructed using words about actors, enablers, and results, while accounts at explaining Level 3 build upon words about molecules and energy forms. Based on this idea, we first developed four groups of words that were aligned with the four explaining levels. However, some words may be more familiar to students than other words in the same group simply because they are commonly used in everyday life. Hence, we added two adjusted levels—Level 1.5 (easier hidden mechanism words) and Level 2.5 (easier scientific words). Due to cultural differences, the US and Chinese versions of the naming levels are slightly different. One example relates to the word, "combustion." In English, "combustion" is the scientific term used to refer to the chemical change. In everyday life, people use "burning" to refer to the same process. In Chinese, there is only one word "燃烧," which is used in both everyday life and science. Hence, we put combustion at Level 3 in the US version and put 燃烧 at Level 2.5 in the Chinese version.

Although the naming and explaining levels are aligned in a logical way, students may construct accounts that indicate different naming and explaining levels. We found that some students were able to recite higher-level words but relied on lower-level reasoning to make accounts; there were also students who adopted higher-level reasoning to make accounts but lacked the necessary words to explain the specific processes.

Table 2. Revised Learning Progression Framework.

<table>
<tr><th colspan="2">Explaining Progress Variable</th><th colspan="2">Naming Progress Variable</th></tr>
<tr><td>Level 4. Linking processes with matter and energy as constraints</td><td>Explain macro-processes by reasoning across scales: link carbon-transforming processes at atomic-molecular, macroscopic, and global scales with matter and energy as constraints</td><td>Level 4. Scientific statements</td><td>MATTER: scientifically appropriate names for both reactants and products; both gases and solids/liquids named as material reactants or products
ENERGY: all forms of energy involved in the chemical change; heat as byproduct</td></tr>
<tr><td rowspan="2">Level 3. Changes of molecules and energy forms with unsuccessful constraints</td><td rowspan="2">Explain macro-processes in terms of change of molecules and/or energy forms at atomic-molecular or global scale but do not successfully conserve matter/energy</td><td>Level 3. Scientific words for organic molecules, energy forms, and chemical change</td><td>MATTER (organic molecules): glucose, $C_6H_{12}O_6$, monosaccharide, glycogen, lipid, ATP, ADP, carbohydrate, hydrocarbon, octane
ENERGY (bonds, energy forms): C-C bond, C-H bond, light energy, kinetic energy (US version), electrical energy, chemical energy, heat energy
PROCESS (chemical reaction): cellular respiration (US version), combustion (US version), oxidation, light reaction, dark reaction</td></tr>
<tr><td>Level 2.5. Easier scientific words with mixed meanings</td><td>MATTER: Fat, sugar, starch, organic matter, carbon, molecule, atom
ENERGY: stored energy, motion energy, 动能(motion/ kinetic energy in Chinese version)
PROCESS: photosynthesis, decomposition/decomposer, chemical reaction/change, 燃烧 (combustion/burning in Chinese version), 呼吸作用 (respiration/breathing in Chinese version)
OTHERS: chloroplast</td></tr>
<tr><td rowspan="2">Level 2. Force-dynamic accounts with hidden mechanisms</td><td rowspan="2">Explain macro-processes in terms of unobservable mechanisms or hidden actors (e.g., decomposer), but the focus is on enablers, actors, and results rather than changes involving matter or energy.</td><td>Level 2. Hidden mechanism words</td><td>MATTER: carbon dioxide, oxygen, nutrients, gas (as in gas, liquid, and solid)
ENERGY: calories, electricity
PROCESS: digestion, digest, digestive system, break down
OTHERS: decomposer (e.g., bacteria, fungi, micro organisms), cell, power plants</td></tr>
<tr><td>Level 1.5. Easier hidden mechanism words</td><td>ACTOR: organs (e.g., lung, stomach, heart), machine parts (e.g., engine, cylinder, piston), material
ENABLER: fuels (e.g., gasoline, diesel, oil, coal, petroleum), heat</td></tr>
<tr><td>Level 1. Macroscopic force-dynamic accounts</td><td>Explain macro-processes in terms of the action-result chain: the actor uses enablers to accomplish its goals; the interactions between the actor and its enablers are like macroscopic physical pushes-and-pulls that do not involve any change of matter/energy</td><td>Level 1. Words about actors, enablers, and results</td><td>ACTOR: body parts (e.g., leaves, roots, leg)
ENABLER: water, air, sunlight, food (e.g., food, milk, bread), bugs, wind, lighter, etc.
RESULT: strong, healthy, grow, run, warm, etc.</td></tr>
</table>

Using the revised learning progression framework, we could identify Episode 5 as an example in which the student's naming level is ahead of her explaining level: Peng stated that carbon dioxide changed into oxygen through a chemical reaction in which water plus carbon dioxide becomes $C_6H_{12}O_6$. She named both reactants (carbon dioxide and water) and products (oxygen and $C_6H_{12}O_6$) of photosynthesis. Hence this account is at naming Level 4. However, Peng's responses to the follow-up questions indicate that she used the scientific words without giving them scientific meanings. Although she was able to describe the chemical reaction of photosynthesis correctly, she still claimed that the increased mass of the tree came from water and nutrients from soil, indicating that she did not connect photosynthesis with tree growth. Peng's accounts indicate force-dynamic reasoning with hidden mechanisms—water and nutrients from soil somehow change into the tree's body structure, and the carbon dioxide the tree breathes in somehow becomes oxygen. Hence, this account is at explaining Level 2.

As shown in Episode 6, there were also a few students who were able to use more sophisticated reasoning to explain the macro-processes but lacked the necessary knowledge about the specific contexts such as terms for specific molecules, processes, and so on.

Episode 6. Flame Burning

(Post-interview with a US eighth grader)

Interviewer: Can you tell me about what is happening inside the [candle] flame as it burns?

Eric: Not specifically, all I know is that it is a chemical reaction and change and that's about all I know for sure, as to what's happening inside the flame itself.

Interviewer: Does this process require energy, the process of burning?

Eric: Yes it does, because it needs energy to perform the chemical changes and it takes the energy that is in the wick and uses that for energy a, to help take more energy out, and b, to send energy out in the form of heat and light.

Interviewer: The melting candle loses weight as they burn, how does this happen?

Eric: The wax of the candle will melt and then often it will pour over the side and spread onto the table or whatever it's sitting on, or else it will slowly evaporate into the air.

Interviewer: You said it slowly evaporates into the air, what form is that?

Eric: I guess it would be wax vapor or something like that, and it basically the molecules of the wax spread apart and far enough from the heat, that because of the heat they become a gas and float into the air.

Interviewer: Are there chemical changes that are happening that the wax to what floats in the air, what is that that floats in the air from the wax?

Eric: It would be whatever chemicals the wax is made of, I am not sure what it is, the molecules of those chemicals will be transferred to the air.

Interviewer: You said that this process requires energy, what are the energy sources?

Eric: The energy source would be directly the wick, which got it from whatever the wick was made of, and it uses that stored energy for the energy of burning.

Interviewer: Do you think energy is released from this burning?

Eric: Yes.

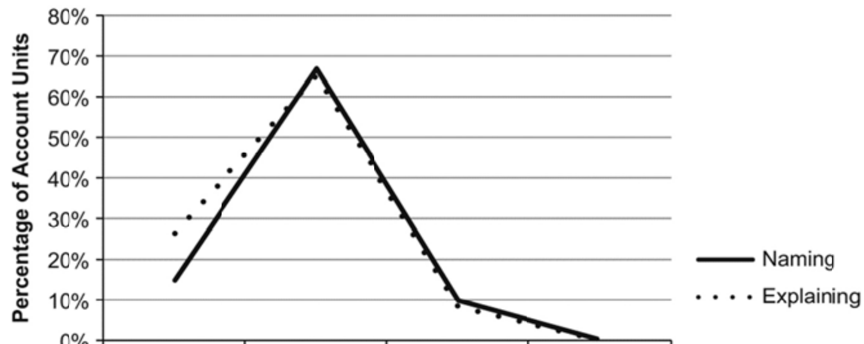
Interviewer: How is it released?

Eric: I am not sure, I believe it is just… the energy of it is changed from the stored energy into light energy or heat energy.

Eric identified a chemical reaction, although he was not sure exactly what the chemical reaction was: "Not specifically, all I know is that it is a chemical reaction and change and that's about all I know for sure, as to what's happening inside the flame itself." He made the common mistake of thinking that the wick, rather than the wax, was burning. Given this assumption, however, he was still able to construct an account that conserved both matter (the wax evaporates but is still present in the air) and energy (the stored energy of the wick is converted to heat and light). Eric's account is at explaining Level 4 because it conserves matter and energy separately. His account is at the relatively lower naming Level 2.5 because the most sophisticated terms in his accounts are Level 2.5 terms—stored energy, chemical reaction, and molecule. Eric did not mention any specific molecules, nor did he identify the chemical reaction as combustion.

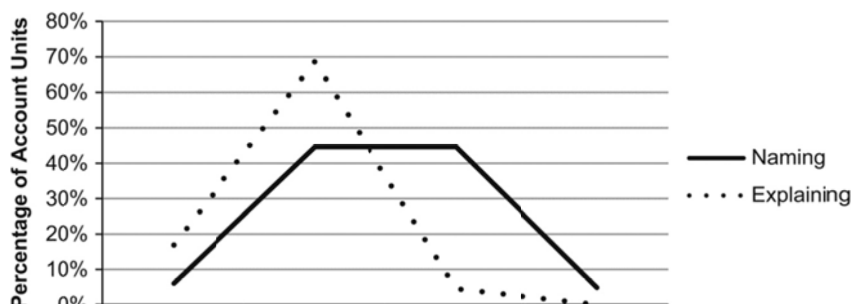
After we developed the final learning progression framework, we used it as a guide to code students' interview responses. Our interview questions are structured around eight macro-processes. We developed naming-explaining coding rubrics for each macro-process based on the learning progression framework (Table 2). Student interviews were divided into eight account units (one for each macro-process), which were analyzed using the rubrics. We generated graphs that show the distribution of students' account units along the naming progress variable and the explaining progress variable. Figures 9 and 10 show the distribution of account units at different levels for the US interviews and the Chinese interviews.[2]

Except at the lower end of the learning progression, the naming-explaining distribution graphs indicate that both groups show higher levels for naming than for explaining, but the difference is much greater for the Chinese students. This indicates that although Chinese students used scientific terms, they sometimes did not understand the scientific meanings of these words and still relied on lower-level reasoning to make accounts. The naming and explaining performances show two different patterns of achievement for US students and Chinese students. This helps explain why the coding rubrics we developed based on US student performances were less effective with the Chinese student data. Chinese students were sometimes coded at higher levels of achievement in previous studies, not because they reasoned at higher levels, but because they used vocabulary words that were reliably associated with high-level reasoning by US students.



	Level 1	Level 2	Level 3	Level 4
Naming	14.90%	66.90%	9.90%	0
Explaining	26.40%	65.20%	8.30%	0

Figure 9. Distribution graphs for US pre-interviews (16 middle and high school students).



	Level 1	Level 2	Level 3	Level 4
Naming	6%	44.60%	44.60%	4.80%
Explaining	16.90%	68.70%	4.50%	0

Figure 10. Distribution graphs for Chinese interviews (16 middle and high school students).

Our analyses and revisions continue. We have not conducted systemic studies of the relationship between the interview data and written assessment data. In addition, the explaining progress variable describes students' performances in a very general way. We need to think about whether and how to make the progress variable include more about the science content that we are interested in—carbon-transforming processes.

CONCLUSION

In this chapter we described one research cycle using an iterative process that involves three phases: observation, interpretation, and a cognitive model or learning progression framework (Figure 1). To develop a learning progression framework that captures major reasoning patterns of students and to better understand the nature of their learning, we involved participants from a wide age range and from two countries. The diversity of participants allowed us to collect rich data, but it also created special assessment challenges. Our responses to the challenges are grounded in our data and in cognitive and linguistic theories about how students understand the world. To effectively elicit accounts from all students, we re-designed the interview protocol and written assessment items at the observation phase. At the interpretation and model building phases, our work focused on identifying progress variables—naming and explaining—that accurately describe students' learning performances and capture important differences between US accounts and Chinese accounts.

Our research is ongoing. As we continue with cycles of observation, interpretation, and model building, we continue to encounter assessment challenges. We conclude this chapter by noting five challenges that we are presently working on.

1. *Designing interview protocols.* Our interviewers—teachers and graduate students—are often relatively inexperienced and may not be native speakers of English. Thus we need to develop protocols and training methods that provide specific interview questions and interviewer support for "on the spot" decisions about follow-up questions to further explore students' thinking. This is especially challenging when students give unexpected answers. Our interview data indicate that interviewers sometimes failed to probe incomplete responses, skipped questions incorrectly, or gave inappropriate clues to students.

2. *Investigating students' accounts of large-scale systems.* Our decision to focus on macroscopic linking processes means that we have not investigated students' understanding of large-scale systems. There are two important aspects of understanding large-scale systems: classification of macro-processes and connections among the macro-processes. In scientific accounts, the macro-processes are classified and connected in terms of matter transformation and energy transformation. However, students may use informal ways to classify and connect the macro-processes, which often indicate their specific ways of reasoning. We developed some interview questions and written assessment items about classification and connections to explore these informal ways of reasoning, but most of the questions were not sufficiently effective in eliciting students' ideas. Time is also a concern. In most of these interviews, the interviewers spent most of their time exploring students' ideas about the macro-processes and did not have enough time to ask about the classification of and connections between macro-processes.

3. *Investigating students' arguments.* Both interviews and written assessments fail to explore students' reasons for their claims or how they defend them. We did not investigate students' argumentation skills, although argumentation is an important component of environmental literacy. We are interested in exploring how students use data to defend their claims about each of the linking processes.

4. *Cross-analysis of interview and assessment data.* As noted above, we have not conducted systemic analyses between the interview data and written assessment

data. Clinical interviews allowed us to identify important patterns in students' informal ways of reasoning. However, due to the small interview sample size, findings cannot be used for statistical generalizations. Compared with interviews, written assessment items are less effective in eliciting detailed accounts. This is especially true for younger students whose responses can be very short and vague due to their limited writing abilities. Therefore, we need to use interviews to validate our interpretations of the relatively brief and incomplete accounts that students give on written assessments. We used naming and explaining as progress variables to analyze interview data, but students' written responses usually contain evidence of only one progress variable. In such situations, how do we conduct the cross-analyses? One possible solution is to use item clusters. An item cluster contains items that ask different questions about the same macro-process. We expect that data from a cluster of items rather than a single item would provide evidence of both naming and explaining.

5. *Investigating the development of students' ideas about matter and energy.* By focusing on general causal reasoning patterns, the explaining progress variable allowed us to avoid dealing with inconsistency between the two lower levels that are about force-dynamic reasoning and the two higher levels that focus on matter and energy. Although younger students do not use matter and energy to account for the macro-processes, they develop intuitive ideas based on their everyday experiences, and some of those ideas are powerful precursors of matter and energy. Therefore, we need to identify progress variables that capture the common facets of both scientific performances of matter and energy and students' lower-level performances that are related to matter, energy, or their precursors. Such progress variables are both science-based and performance-based.

These are the challenges we are working on in our current research cycle. We look forward to tackling these, as well as additional challenges that will arise as we continue to develop a learning progression on carbon-transforming processes.

ACKNOWLEDGEMENTS

This research is supported by grants from the National Science Foundation: Learning Progression on Carbon-Transforming Processes in Socio-Ecological Systems (NSF 0815993); Targeted Partnership: Culturally Relevant Ecology, Learning Progressions and Environmental Literacy (NSF-0832173); and CCE: A Learning Progression-Based System for Promoting Understanding of Carbon-Transforming Processes (DRL 1020187). Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This chapter is based upon the work of the Environmental Literacy Project Carbon Group (Michigan State University) and BEAR Center (University of California, Berkeley). We would like to thank Jinnie Choi, Karen Draney, Mark Wilson, Lindsey Mohan, Jing Chen, Li Zhan, Kennedy Onyancha, Hamin Baek, Yong-Sang Lee, and Jonathon Schramm for their invaluable contributions.

NOTES

[1] Modern physical theories hold that this statement is not strictly true. We believe, however, that students need to learn to conserve matter and energy independently before addressing matter-energy conversions in nuclear reactions.

[2] In our final coding results, account units coded as naming Level 1.5 were included in the category of naming Level 1 and account units coded as naming Level 2.5 were included in the category of naming Level 2.

REFERENCES

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

Anderson, C. W., Sheldon, T. H., & DuBay, J. (1990). The effects of instruction on college non-majors' conceptions of respiration and photosynthesis. *Journal of Research in Science Teaching, 27*, 761–776.

Chen, J., Anderson, C. W., & Jin, X. (2009, April). *American and Chinese secondary students' written accounts of carbon cycling in socio-ecological systems*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Choi, J., Lee, Y., & Draney, K. L. (2009, April). *Empirical validation of a learning progression*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 13*, 15–42.

Dai, J.-h. E. (2005). *Conceptualizations and cognitive relativism on result in Mandarin Chinese: A case study of Mandarin Chinese bǎ construction using a cognitive and centering approach* (Doctoral dissertation, Louisiana State University, Baton Rouge, LA). Retrieved from http://etd.lsu.edu/docs/available/etd-04152005–112438/unrestricted/Dai_dis.pdf

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *Journal of the Learning Sciences, 11*, 105–121.

Jin, H., & Anderson, C. W. (2008, April). *Developing a long-term learning progression for energy in socio-ecological systems*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Jin, H., Zhan, L., & Anderson, C. W. (2009, April). *A cross-cultural study: Comparing learning progressions for carbon-transforming processes of American and Chinese student*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

Lai, H.-l., & Chiang, S.-m. (2003). Intrapsychological force-dynamic interaction: Verbs of refraining graining in Hakka. *Taiwan Journal of Linguistics, 1*(2), 35–64.

Lee, O., Eichinger, D., Anderson, C. W., Berkheimer, G. D., & Blakeslee, T. D. (1993). Changing middle school students' conceptions of matter and molecules. *Journal of Research in Science Teaching, 30,* 249–270.

Long Term Ecological Research Network. (2007). *Integrative science for society and the environment: A plan for research, education, and cyberinfrastructure in the U.S.* Retrieved from http://www.Lternet.edu/decadalplan/

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

Mohan, L., Chen, J., Baek, H., Anderson, C. W., Choi, J., & Lee, Y. (2009, April). *Validation of a multi-year carbon cycle learning progression: A closer look at progress variables and processes*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Garden Grove, CA.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. New York, NY: Penguin Group.

Talmy, L. (2000). *Toward a cognitive semantics*. Cambridge, MA: MIT Press.

AMELIA WENK GOTWALS, NANCY BUTLER SONGER AND LEA BULLARD

ASSESSING STUDENTS' PROGRESSING ABILITIES TO CONSTRUCT SCIENTIFIC EXPLANATIONS

Learning science does not just consist of memorizing a body of facts; rather, the development of scientific knowledge is a dynamic endeavor that represents a complex interplay of content and scientific practices (National Research Council [NRC], 2007). While policy documents and standards of the 1990s listed content standards and practice standards (often called inquiry standards) separately (e.g., NRC, 1996), more recent standards and policy documents emphasize that science knowledge is a fusion of content and practices (NRC, 2011). Despite this shift, policy documents do not provide clear guidelines on how to guide students towards knowledge that is a fusion of content and practices (Songer, Kelcey, & Gotwals, 2009). Various pedagogical methods, such as curricular scaffolds, have been shown to support student learning of fused content and practices (McNeill, Lizotte, Krajcik, & Marx, 2006; Songer et al., 2009). However, these same research studies also suggest that details of the support structures, including the best means for both scaffolds and fades, require greater investigation.

This chapter describes the project, *BioKIDS: Kids' Inquiry of Diverse Species*, in which we developed and implemented an assessment system in order to better understand how students develop knowledge that fuses content and practices. The discussion of our assessment system focuses on the assessment triangle (NRC, 2001), which consists of three vertices: (1) *Cognition,* which refers to the knowledge to be assessed, including the learning theory that explains how students may learn this knowledge; (2) *Observation,* which refers to the type of task that would best elicit performances that demonstrate understanding of this knowledge; and (3) *Interpretation*, which refers to a method of interpreting the performances gathered from the task. Each vertex of the triangle must not only make sense on its own, but must also connect to the other vertices in clear and meaningful ways.

To address the cognition vertex of the assessment triangle, we briefly describe the content and practices targeted in our project and the context in which we have implemented our curricula and assessments. For the observation vertex, we describe how we designed assessment tasks to gather evidence of the ways that students develop fused content and practices over time. In this section we address challenges that we have faced based on our design decisions – specifically, the benefits and possible limitations of incorporating scaffolds into our assessment

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 183–210.

tasks. For the interpretation vertex, we explain how we designed coding rubrics to analyze student responses to our assessment tasks. We discuss the challenges of interpreting student responses with respect to a learning progression, especially in terms of the possible range of performances and types of errors that students may exhibit as they develop more sophisticated knowledge (Gotwals & Songer, 2010; Songer et al., 2009; Songer & Gotwals, 2012).

COGNITION VERTEX: FUSED CONTENT AND PRACTICE KNOWLEDGE FOCUSED ON BIODIVERSITY

The first vertex of the assessment triangle is cognition, which refers to the learning theory that drives the assessment as well as the kinds of knowledge, skills, and abilities valued by this learning theory (NRC, 2001). We utilize a learning progression framework as the foundational template for the development of curricular, assessment, and professional development products. Our definition of learning progressions builds on others' definitions that emphasize the increasingly sophisticated ways of thinking about a particular topic (e.g., NRC, 2007), while also specifying the representation of both content and practices. In particular, we define learning progressions as follows:

> Learning progressions take a stance about both the nature and the sequence of content and inquiry reasoning skills [practices] that students should develop over multiple curricular units and years. Learning progressions are successively more sophisticated ways of thinking about a topic that can be used as templates for the development of curricular and assessment products… The learning progression can only be evaluated indirectly, through the evaluation of the curricular products, professional development modules, and assessment instruments that are constructed from the learning progressions template. (Songer et al., 2009, p. 612).

In the BioKIDS project, our learning progression includes a content progression and a practice progression. Both our content progression and our practice progression consist of a series of ideas or levels that are sequenced to encompass the life science content fostered in our curricular programs over a three-year period (fourth through sixth grades).

Our content progression includes three strands (classification, ecology, and biodiversity) that build over time and intersect at key junctures (Table 1). Each content idea builds on previous ideas, such that more complex content ideas are at the top of the progression while more basic ideas are at the bottom of the progression. For example, in fourth grade, the content progression sequences increasingly use more complex content ideas in both classification and ecology. The ability to use these content ideas in tandem to explain phenomena provides the foundation of the fifth grade portion of the content progression where more complex classification ideas and biodiversity ideas are developed. The full version of the content progression includes more elaborated content statements for each idea as well as ideas between the lower and upper ideas in each cell. (For a full description of the content progression, see Songer et al., 2009).

Table 1. Representation of BioKIDS Content Progression.

	Classification Strand	*Ecology Strand*	*Biodiversity Strand*
6th Grade		Complex Ecological Idea: A change in one species can affect different members of the food web… . . . Middle Ecological Idea: Plants and animals of a habitat can be connected in a food chain	Complex Biodiversity Idea: Humans and other factors affect biodiversity… . . . Middle Biodiversity Idea: Biodiversity differs in different areas…
5th Grade	Complex Classification Idea: Patterns of shared characteristics reveal the evolutionary history… . . . Middle Classification Idea: Organisms are grouped based on their structures…		Middle Biodiversity Idea: An area has a high biodiversity if it has both high richness and abundance . . . Basic Biodiversity Idea: A habitat is a place that provides food, water, shelter…
4th Grade	Middle Classification Idea: Organisms have different features that allow them to survive . . . Basic Classification Idea: There are observable features of living things	Middle Ecological Idea: Only a small fraction of energy at one level … moves to the next level . . . Basic Ecological Idea: Every organism needs energy to live…	

Note. For more details, see Songer et al. (2009).

Our practice progression focuses on the development of evidence-based explanations. Evidence-based explanations are part of the core of scientific practice and are considered an essential skill for scientifically literate citizens (NRC, 2007). In addition, we selected this practice because research demonstrates that students who engage in the development of evidence-based explanations can significantly improve their understanding of scientific concepts, their understanding of appropriate use of evidence, and their ability to provide coherent and logical arguments (Bell & Linn, 2000; Scardemalia & Bereiter, 1991; Toth, Suthers, & Lesgold, 2002; White & Frederiksen, 1998). Finally, an examination of the quality and coherence of students' explanations can serve as a valuable way to assess their understanding (Metz, 1991).

Similar to others, we have adopted a modified version of Toulmin's (1958) model of argumentation to support teachers and students in creating scientific explanations (Bell & Linn, 2000; Driver, Newton, & Osborne, 2000; Erudan, Simon, & Osborne, 2004; Lee, 2003; McNeill et al., 2006; Sandoval, 2003). In our work, we emphasize three essential aspects of evidence-based explanations: (1) articulation of causal claims; (2) use of appropriate and sufficient evidence to support these claims; and (3) use of reasoning that draws on scientific principles to explicitly link the evidence to the claim. Claims are assertions or conclusions in response to a scientific question (in our project, claims are either given to students by the teacher or curriculum or students create their own claims). Evidence consists of scientific data (either collected by students or given to students by the teacher or curriculum) used to support students' claims. Data used as evidence must be appropriate and sufficient. Appropriate data is relevant to the question or problem and supports the claim. There is sufficient evidence when enough relevant data is used to convince someone of the accuracy of the claim (McNeill & Krajcik, 2007). Finally, reasoning is a justification that utilizes salient scientific principles to show why the data counts as evidence in support of the claim.

Scaffolding

Despite research on the value of students' work with evidence-based explanations, our research, and that of others, testifies to the fact that guiding students towards an understanding of appropriate evidence and the development of sound explanations is not a straightforward task. For example, students often struggle with articulating clear claims based on the data they have. They often do not fully understand what counts as evidence (Sadler, 2004) or how to incorporate appropriate evidence (Lee & Songer, 2003; Sandoval, 2003) and sufficient evidence (Sandoval & Millwood, 2005) in their explanations. Middle school students have particular difficulty with the reasoning component of explanations (Lizotte, Harris, McNeill, Marx, & Krajcik, 2003; Gotwals, 2006). Germann and Aram (1996) found that students had a hard time presenting evidence in a convincing way; in other words, students did not provide reasoning to indicate why the evidence was appropriate. Often students make claims but do not back up the claims with evidence or reasoning (Jimenez-Aleizandre, Rodriguez, & Duschl, 2000).

Despite the problems that students face in building explanations, studies have shown that when students work through repeated exposures within parallel examples that guide or scaffold their construction of evidence-based explanations, they make significant gains in content knowledge and increase their ability to provide clear and coherent explanations (Lee, 2003; McNeill et al., 2006; Songer et al., 2009). Educational scaffolds have also been found to help younger students work with complex scientific information and participate in scientific inquiry activities (Metz, 2000).

Our practice progression builds from our earlier work and that of others in the design of scaffolds. Table 2 presents our practice progression focusing on evidence-based explanations. There are four levels with two conditions at each level – one condition states what students are able to do with scaffolding (indicated with an "s") and the other condition states what students can do without scaffolding.

Table 2. Practice Progression for Evidence-Based Explanations.

Level 4 Student constructs a complete evidence-based explanation (without scaffolding)
Level 4s Student constructs a complete evidence-basedexplanation (with scaffolding)
Level 3 Student makes a claim and backs it up with sufficient and appropriate evidence but does not use reasoning to tie the two together (without scaffolding)
Level 3s Student makes a claim and backs it up with sufficient and appropriate evidence but does not use reasoning to tie the two together (with scaffolding)
Level 2 Student makes a claim and backs it up with appropriate but insufficient (partial) evidence (without scaffolding)
Level 2s Student makes a claim and backs it up with appropriate but insufficient (partial) evidence (with scaffolding)
Level 1 Student makes a claim (without scaffolding)
Level 1s Student makes a claim (with scaffolding)

Note. For a full description, please see Songer et al. (2009)

The fusion of content and practice occurs through learning objectives that are associated with each curricular activity and each assessment task. Figure 1 presents a sample from our sixth grade program that includes two content ideas, one practice idea, and the learning objective that fuses them in the context of curricular and assessment activities (in this case, the Detroit River ecosystem).

Biokids Project

The BioKIDS project used the content and practice progressions described above as templates to create a coherent set of curricula, assessments, and professional development products. We worked with fourth, fifth, and sixth grade teachers and students in a large Midwestern urban district. The curricular activities have scaffolds that support students in fusing content knowledge in ecology with the practice of evidence-based explanations. As students develop more sophisticated knowledge, the scaffolds fade to allow students to tackle more of the practice on their own. In addition, the curricula are educative (Davis & Krajcik, 2005), giving teachers information about alternative student ideas, tips for introducing and teaching evidence-based explanations and suggestions for using the written scaffolds in the student workbooks. Teachers implemented the curricula and administered the assessments after participating in a weeklong summer meeting and monthly professional development meetings. The research team collected multiple forms of data, including written assessments (both embedded and summative), think-aloud interviews as students completed the assessment tasks, interviews with teachers and students, and classroom observations. Next we describe how we created our assessment tasks to collect information about student learning in the BioKIDS project.

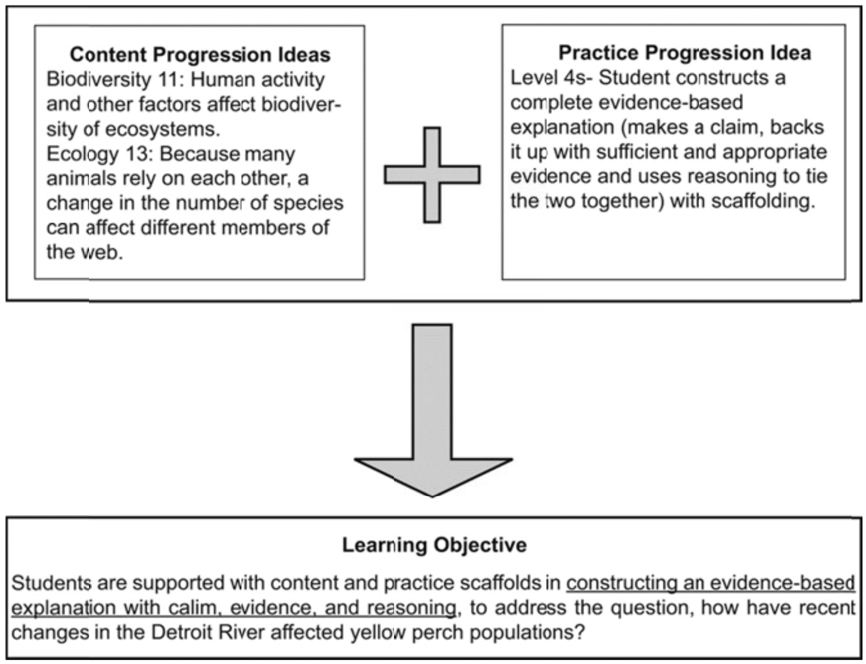


Figure 1. Fusing content and practice ideas to form a learning objective.

OBSERVATION VERTEX: ASSESSMENT TASK STRUCTURES

The second vertex of the assessment triangle is the observation vertex, which describes the "kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills" (NRC, 2001, p. 47). Assessment tasks allow us to gather evidence of what students know and can do. All assessments, explicitly or not, are based on a model of how students learn and on methods for designing tasks to elicit evidence of students' abilities (Mislevy, Steinberg, & Almond, 2003). In order to design assessments matched to our learning progression framework, we need tasks that allow students at all levels of the content and practice progressions to demonstrate their understandings. *Inquiry and the National Science Education Standards* state that "it is a mistake to think that all instruction and assessments should aim for the higher level of outcome" and that when the student fails at higher order problems it needs to be known whether they are "lacking specific skills or the knowledge needed for success" (NRC, 2000, p. 78). This stance is in line with the idea of using learning progressions as a way to design opportunities for students to give us evidence at many levels as they progress towards fused content and practices. When students are not fully successful at higher order problems, assessments based on this stance allow us to see where their knowledge begins to fall apart. Therefore we designed a suite of tasks that were mapped to one or more content ideas from the content progression and to one idea from the practice progression. These assessment tasks allowed us to gather evidence of how students use content ideas to formulate evidence-based explanations associated with a range of ecology and biodiversity topics.

Task Design Decisions – What Types of Tasks Can Be Used to Elicit a Range of Student Knowledge?

In order to gather information about our complete sample of students and the ways they develop fused science knowledge as they engage with our curricular units, we needed a suite of assessment tasks matched to students at a variety of levels on our content and practice progressions. Given the range of ability levels in a classroom and the vast range of prior experiences that students bring to the classroom, providing well-matched tasks for all students may not be possible. However, using tasks that target different levels of evidence-based explanations about focal content provides students at multiple levels with opportunities to demonstrate what they know and can do.

We know that students tend to struggle with formulating evidence-based explanations (Lee & Songer, 2003; Sadler, 2004; Sandoval, 2003). We also know that scaffolding students in creating the key components of explanations (claim, evidence, and reasoning) allows them to create better and more coherent explanations (Lee & Songer, 2003). Thus, to tackle the challenge of providing students at multiple levels along the practice progression with assessment tasks well matched to their levels; we drew on our previous work with Detroit 6th grade students. In this work, we used written scaffolds, similar in structure to those in our curriculum that provide students with different levels of explanation tasks (Gotwals & Songer, 2010).

Our assessments included four task levels that varied as to the amount and type of scaffolding. Both content scaffolds (that provide content hints about what to include for the claim, evidence, and reasoning) and practice scaffolds (that provide prompts and hints about the three components of evidence-based explanations) were included. In items at the Minimal level, students are given evidence and are asked to choose a claim that matches the evidence (e.g., Figure 2); in Intermediate I items,

TYPE OF ORGANISM	WHAT THEY EAT
small fish	algae
large fish	small fish
heron	small fish, large fish, frogs, insects
frog	insects

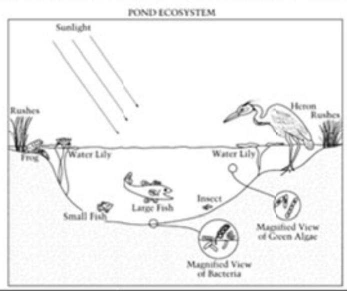
Look at the table above. Which organisms might compete for food ?
(circle one)

A. Heron and large fish
B. Insects and large fish
C. Small fish and heron
D. Frog and small fish

Figure 2. Minimal task – students use data to choose a claim.

students construct an explanation with structural practice scaffolds and content scaffolds (e.g., Figure 3); in Intermediate II items, students construct an explanation with only structural practice scaffolds (e.g., Figure 4); and in Complex items, students construct an explanation with no scaffolds (e.g., Figure 5). The different levels of scaffolding offer students a range of opportunities to demonstrate what they are able to do with and without support.

Use the picture and the table of eating relationships below to help you answer question 3.



TYPE OF ORGANISM	WHAT THEY USE FOR ENERGY
small fish	water lily
large fish	small fish, water lily
heron	small fish, large fish, insects

3. **Write a scientific explanation for the following question.**

<u>Scientific Question</u>: Is the large fish a producer or a consumer?

Make a CLAIM:
Write a sentence that answers the scientific question.

Hint: Think about how producers and consumers get energy.

Give your REASONING:
Write the scientific concept or definition that you thought about to make your claim.

Hint: Think about the definition of the scientific term you used.

Give your EVIDENCE:
Look at your data and find two pieces of evidence that help answer the scientific question.

1.

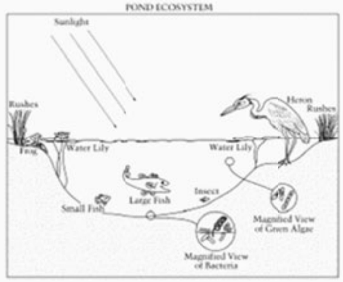
2.

Hint: Think about where the large fish gets its energy.

Figure 3. Intermediate I task with both content and practice scaffolding.

Use the picture and the table of eating relationships below to help you answer question 4.

POND ECOSYSTEM



TYPE OF ORGANISM	WHAT THEY USE FOR ENERGY
small fish	water lily
large fish	small fish, water lily
heron	small fish, large fish, insects

4. **Write a scientific explanation for the following question.**
<u>Scientific question</u>: Is the heron an herbivore, omnivore, carnivore, or decomposer?

Make a CLAIM:
Write a sentence that answers the scientific question.

Give your REASONING:
Write the scientific concept or definition that you thought about to make your claim.

Give your EVIDENCE:
Look at your data and find two pieces of evidence that help answer the scientific question.
1.

2.

Figure 4. Intermediate II task with practice scaffolds.

David is in Mr. Leon's science class. For homework, David has to make a table of all the living things he sees at the playground and what each of the living things eats. Below is the table that David made.

Name of living thing	What it eats	Number seen
Snake	Frogs, mice	3
Grass and seeds	--	Too much to count
Grasshopper	Leaves, grass	14
Frog	Grasshoppers	10

Give a scientific explanation for the following question.

Scientific question: Which 2 living things on David's list are predators?

Figure 5. Complex task with no scaffolds.

Methodology

In this chapter we present findings from a study of 615 responses by fourth, fifth, and sixth grade students to our written assessment tasks. We used six tasks at each scaffolding level to create three tests. The tests had either 12 or 13 tasks, with 2 or 3 tasks at each of the four scaffolding levels. There were 2 or 3 linking tasks between the assessments to allow for calibration of all tasks on a single scale. Students' responses to the assessments were analyzed using basic descriptive statistics as well as item response theory (IRT; Hambelton, Swaminathan, & Rogers, 1991) to determine the relative difficulty of our assessment tasks and students' ability to provide salient information about a range of students.

In addition, we drew on information from think-aloud interviews with 20 students with a range of abilities (as identified by their teachers). The interviews had two parts, a concurrent think-aloud section and a cognitive interview section. Common think-aloud procedures were used in order to examine students' thought processes as they worked on the assessment tasks (Ericsson & Simon, 1993). After students completed the assessment, the interviewer reviewed the assessments with the students by asking them to clarify responses on tasks, to explain how they reasoned about a task, and/or to talk about perceptions of the tasks, for example, which tasks were difficult and which were easy. Following standard procedures (DeBarger, Quellmalz, Fried, & Fujii, 2006; Ericsson & Simon, 1993), interviews were transcribed and segmented. Then the segments were coded for evidence of content and the practice of evidence-based explanation as well as for indications of which aspects of the tasks most influenced how students responded. We analyzed the coded data for the presence of the focal content ideas and for the types and level of evidence-based explanations that students used when interacting with the tasks. We also noted how students interacted with the structural components of the tasks to allow us to make inferences about the effectiveness of our scaffolds.

Typical Progression of Students' Explanations

In our written assessments, students tend to provide more information about their thinking for tasks with scaffolding. In addition, when we compare the IRT difficulty parameters of tasks with different levels of scaffolding, items with the most scaffolding tend to be the easiest for students (Gotwals & Songer, 2010). Figure 6 shows one student's very different responses to two similar items with different levels of scaffolding. [Note that in the figures with student responses, the students' answers are retyped in the right column for easier reading.] The first part of Figure 6 shows a student's response to an Intermediate II item with practice scaffolds while the second part shows his response to a complex item that has no scaffolds. While the tasks are testing different content ideas,[1] they are based on the same scenario and data. Thus we know that the student is able to read the table to gather evidence about what the heron eats. It is possible that the student does not know the definition of predator or prey and that is why he does not expand on his response in the complex task. However, his correct claim (that large fish are both predators and prey) indicates that while he understands the content, he does not provide evidence or reasoning when not prompted to do so. For some students at higher levels of the learning progression, the scaffolds did not seem to influence how they responded (i.e., they included all information whether prompted to do so or not). Similarly, for some students at lower levels, the scaffolds did not help them respond to questions. However, for students in the middle, such as the student whose responses are shown in Figure 6, the scaffolds provide opportunities to elicit a more complete picture of what they know and can do as they fuse content with the scientific practice of building evidence-based explanations.

In addition, we found that the scaffolds direct students to attend to each aspect of evidence-based explanation. In the analysis of the think-aloud interviews, we discovered that almost every student read the scaffolds while responding to the items, indicating some attention to the scaffolds. For example, Figure 7 presents one student's response to a scaffolded item. Initially, the student just gave an explanation (bold text) that includes a claim and one piece of evidence. However, after reading the scaffolding (underlined text), she included a second piece of evidence and attempted to provide reasoning to link her evidence to the claim. Had there been no scaffolding, we would not have known that this student knew the number of body segments of these two animals or that it was an appropriate piece of evidence.

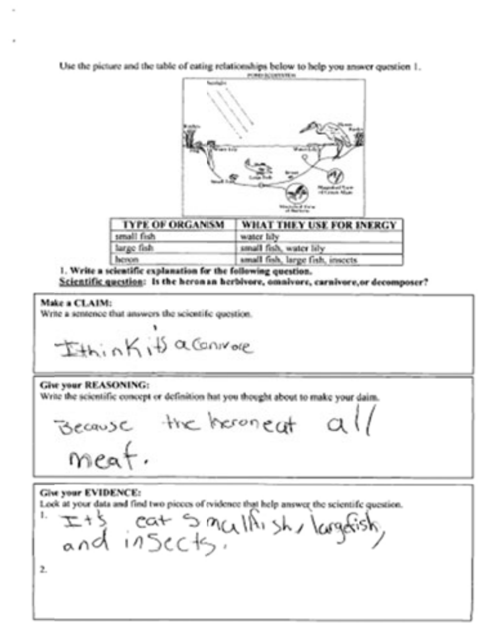
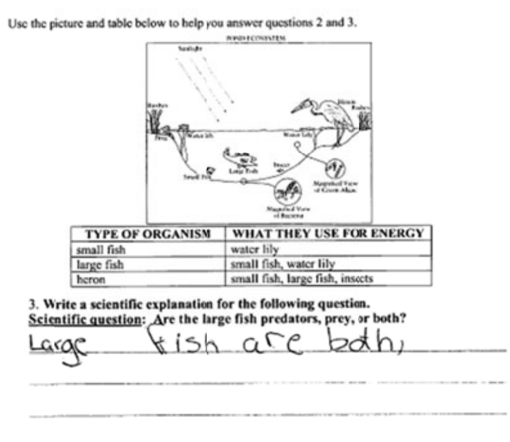
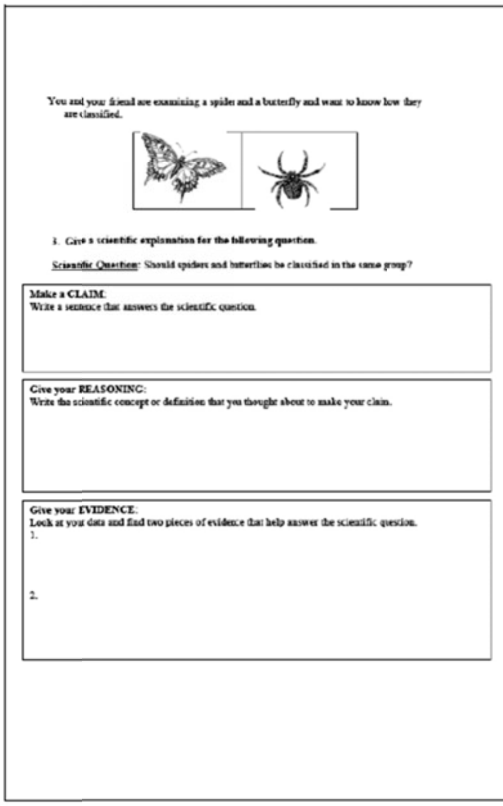
 Use the picture and the table of eating relationships below to help you answer question 1. TYPE OF ORGANISM / WHAT THEY USE FOR ENERGY: small fish – water lily; large fish – small fish, water lily; heron – small fish, large fish, insects 1. Write a scientific explanation for the following question. Scientific question: Is the heron an herbivore, omnivore, carnivore, or decomposer? Make a CLAIM: Write a sentence that answers the scientific question. I think it's a Conivore Give your REASONING: Write the scientific concept or definition that you thought about to make your claim. Because the heron eat all meat. Give your EVIDENCE: Look at your data and find two pieces of evidence that help answer the scientific question. 1. It's eat smallfish, largefish, and insects. 2.	**Scientific Question**: Is the heron an herbivore, omnivore, carnivore, or decomposer? **Response:** **Claim**: I think it's carnivore **Reasoning:** Because the heron eat all meat **Evidence:** It's eat small fish, large fish, and insects
 Use the picture and table below to help you answer questions 2 and 3. TYPE OF ORGANISM / WHAT THEY USE FOR ENERGY: small fish – water lily; large fish – small fish, water lily; heron – small fish, large fish, insects 3. Write a scientific explanation for the following question. Scientific question: Are the large fish predators, prey, or both? Large fish are both,	**Scientific Question:** Are the large fish predators, prey or both? **Response:** Large fish are both,

Figure 6. Student DTS09550111's response to tasks with and without structural practice scaffolds.



Student Think-Aloud Response:
… Give a scientific explanation for the following question. Scientific question: should spiders and butterflies be classified in the same group? Um spiders and butterflies should or should not be classified in the same group, mmhmm, **I think that they should not because butterflies, these, they fly and spiders usually don't fly, I think that's my explanation**. And the reasoning, write the scientific concept or definition that you thought about to make your claim. My evidence supports my claim because it because it's explaining why the spiders and butterflies should not be in the same group or not be classified in the same thing. My evidence, find two pieces of evidence to support the claim, they should not be classified in the same group because spiders do not fly and butterflies do and because oh and I think because butterflies have three body sections and spiders only have two because they are part of arachnids.

Figure 7. Student DTF08070107's think-aloud response illustrating the use of explanation practice scaffolding (bold text is the initial explanation; underlined text indicates where the student reads the written scaffolds).

Challenges with Using Scaffolding in Assessments

Based on our analyses, it is clear that the scaffolds provide students with supported environments in which to demonstrate their fusion of the content and practices of creating evidence-based explanations. However, we have also discovered that students do not always find that the scaffolding makes assessment tasks easier. For some students, the scaffolding may actually introduce more complexity or difficulty. In fact, in a few instances, IRT analyses indicate that the complex items, those without scaffolding, have a lower difficulty level than intermediate items with scaffolding (Gotwals & Songer, 2010). While this result may be due to the content of the items or to students' below grade level reading comprehension, it also may indicate that, for some students, the scaffolding has no impact or even makes the items more difficult to respond to.

After the think-alouds, we interviewed students about how they used the scaffolds to respond to the items. In one interview, a student indicated that while he noted the scaffolds, he didn't pay much attention to them:

> **Interviewer**: So, in the boxes here, we have these directions…for claim, it says, "write a sentence that answers the scientific question", for reasoning, it says, "write the scientific concept or definition you thought about to make your claim." Did those directions help you, or did you pay attention to them?
>
> **Student DTF08400109**: No, I didn't really pay attention to them, they didn't really help me.

In examining this student's think-aloud responses, it seems he used the scaffolds in some cases but not in others. When he read the scaffolds, he was more likely to provide evidence and some form of reasoning to back up his claim. When he did not use the scaffolds, he tended to give a claim and either one piece of evidence or no evidence. These responses suggest that while he may not have thought that he needed to use the scaffolds, he did not fully understand how to create a complete evidence-based explanation. When he read the scaffolds, his responses improved.

This was not a general pattern in our interviews since only this student said he did not pay much attention to the scaffolds. However, the scaffolds were built into the assessments in order to help students who may not have understood what they needed to include in an explanation (claim, evidence, and reasoning) and what each component entails. The scaffolds did not give this student the intended guidance. He provides a possible reason in his response about which items he felt were most difficult: the scaffolded items are more difficult than the unscaffolded tasks, and the scaffolded tasks are for "advanced people" whereas the unscaffolded tasks are for people "who can't really give you all that stuff".

> **Student DTF08400109:** I feel like this [scaffolded task] is for the advanced, for the advanced people who could give, who could give, three things just like that… [referring to the three boxes] yeah, [question] number one [an unscaffolded task] would be for people who can't, who can't really give you all that stuff, but they could tell you a claim…because this, and because that….without the pressure of giving a reason and giving evidence…actually**,** I felt like they was kinda hard…because sometimes when you make a claim, you don't really have a good reason why you made it, you just say, "hey, this is pretty good, I'm gonna try this out and it sounds like a good answer, so why not give it a try, and then, it says, give us a reason,' and you just sit there like, ok, now I gotta make up a reason for this.

There are many possible explanations why this student (and possibly others) did not interact with some of the scaffolded assessment tasks as we designed them. The scaffolded tasks look more complicated since they have boxes and additional instructions or prompts. The complex tasks, which just ask a question and do not provide additional prompts, may seem more familiar to students and thus may not be as threatening as unscaffolded tasks. The scaffolded tasks may actually intimidate students and thus cause them to either ignore the prompts and hints or to shy away from even engaging in them. The additional reading may also intimidate students or pose a problem for students whose reading skills are not at grade level.

These assessment items are closely tied to our curricular units that have a parallel scaffolding structure. Large parts of the units help students use the scaffolds to build evidence-based explanations. Then the scaffolds slowly fade to give students opportunities to build explanations on their own. These assessment tasks may not be a good fit for a large-scale assessment such as the National Assessment of Educational Progress (NAEP; Alonzo, Neidorf, & Anderson, this volume) as they are mapped to a particular curricular program. However, if we are interested in gathering evidence of how students develop their reasoning as a result of specific instruction or when guided in particular ways, then these tasks may provide us with evidence of how our curricular units support students' increasing sophistication in developing evidence-based explanations. In the following section, we describe how we used our learning progression to interpret students' responses to our assessment items and the challenges that we faced in doing so.

INTERPRETATION VERTEX: CODING STUDENTS' RESPONSES

The third vertex of the assessment triangle is interpretation: "all the methods and tools used to reason from fallible observations" (NRC, 2001). Every assessment must be designed with an interpretation model in mind. Part of the coherence of our assessment system comes from developing assessments and interpreting students' responses to the assessments with the same learning progression framework.

While not necessarily unique to learning progression-based assessment, interpreting students' responses to tasks can be difficult. This is especially true for open-ended tasks that are particularly useful for unpacking how well students have begun to fuse content and practices. Students bring a range of prior knowledge to any task. By their very nature, open-ended tasks are subject to a wide variety of interpretations by students and may elicit reasonable answers that have little to do with the knowledge that the test developer intended to assess. Tasks that involve a fusion of content and practices invite multiple ways for students to bring these ideas together; some of these ways may be difficult to classify. In assessments mapped to a learning progression, accurately interpreting and coding students' partial understandings or middle knowledge – "the varieties of not-quite-successful attempts at complex scientific ideas that students manifest on the path towards sophisticated understanding" (Songer et al., 2009, p. 629) – are particularly important. This is because we are interested not just in whether students are right or wrong but also in the intricacies of how they develop more sophisticated knowledge. In order to systematically interpret students' responses to the scientific explanation assessment tasks, we developed a coding scheme based on the levels of our practice progression. This approach is similar to approaches used in other learning progression research (e.g., Alonzo & Steedle, 2009; Mohan, Chen, & Anderson, 2009).

Our coding rubrics are mapped directly to the practice progression. We used the information from the content progression to determine what would count as the correct claim, appropriate and sufficient evidence, and accurate reasoning for each task. The use of both the content progression and the practice progression provides a clear link to the cognition vertex of the assessment triangle. This approach also helps us deal with a main challenge of learning progression-based assessments: gathering evidence of students' knowledge at multiple levels along a learning progression. With our coding rubrics, we are able to interpret students' responses in terms of their competence in using content to formulate evidence-based explanations. Table 3 illustrates the generalized coding rubric for our open-ended items. For each item, we customized the rubric to the particular content. The nature of the intersection between the content and the practice progressions determines the types and amount of content that should be infused into students' claims, as well as what counts as appropriate and sufficient evidence and adequate reasoning. A successful (Level 4) explanation is highly dependent on the content, the data available, and the scientific question posed. Perhaps even more important, we wanted to document the middle levels (Level 2 and Level 3) as this type of middle knowledge is often messy since students may be able to demonstrate a certain level of content or practice in one situation but not in another. Therefore it may not be obvious what these responses look like (Gotwals & Songer, 2010).

Table 3. General Scoring Rubric for Writing Evidence-Based Explanations, Informed by Practice Progression.

Level 4 Student constructs a complete evidence-basedexplanation (with an accurate claim, appropriate and sufficient evidence, and reasoning).
Level 3 Student makes an accurate claim and backs it up with appropriate and sufficient evidence but does not use reasoning to tie the two together
Level 2 Student makes an accurate claim but does not back it up with evidence or reasoning claim and backs it up with insufficient or partially inappropriate evidence
Level 1 Student makes an accurate claim but does not back it up with evidence or reasoning

The specific task rubric identifies how to code students' responses based on the fusion of the content ideas from the content progression and the generalized scoring rubric in Table 3 that is taken from the practice progression. Figure 8 illustrates the coding rubric for the assessment task in Figure 3.

Explanation: **Level 4 (4)**: Contains all parts of explanation (claim, 2 pieces of evidence, reasoning) **Level 3 (3)**: Contains correct claim & at least 2 pieces of appropriate evidence (can contain additional incorrect evidence) OR correct claim, 1 piece correct evidence + reasoning **Level 2 (2):** Contains correct claim and at least1 piece appropriate evidence (can contain additional incorrect evidence) **Level 1 (1):** Contains correct claim, but no evidence or incorrect evidence **Level 0 (0):** Incorrect claim ——————————— **Claim** Correct: The large fish is a **consumer** Incorrect: The large fish is a producer or other claim **Evidence** Possible evidence (must be specific about what large fish eats or get energy from): • The large fish eats (or gets energy from) small fish • The large fish eats (or gets energy from) water lily • The large fish does not use sunlight for energy **Reasoning** <u>Explicit written statement</u> that ties evidence to claim with a reasoning statement i.e.: "Consumers get their energy from eating other organisms" OR "Consumers eat other organisms for food"	**Sample Student Responses:** **Claims:** "Consumer"; "It is a consumer"; "The large fish is a consumer"; "The big fish is a consumer" **Partial Evidence (1 piece of appropriate – can include additional inappropriate evidence):** "it eats the little fish"; "one is that large fish eat small fish and they don't produce"; "Large fish are consumers because they consume small fish when they eat them"; "Large fish eat small fish and nothing eats them" **Appropriate and Sufficient Evidence:** "The large fish gets energy from small fish and from water lilies"; "The large fish eat plants and animals but it doesn't get any energy straight from the sun"; "The large fish eats water lily and small fish" **Reasoning statements:** "A consumer is something that eats something else to get its energy"; "An animal is called a consumer when it has to eat for energy"; "Consumers are living things that eat other living things and get their energy that way"

Figure 8. Coding rubric for the item in Figure 3.

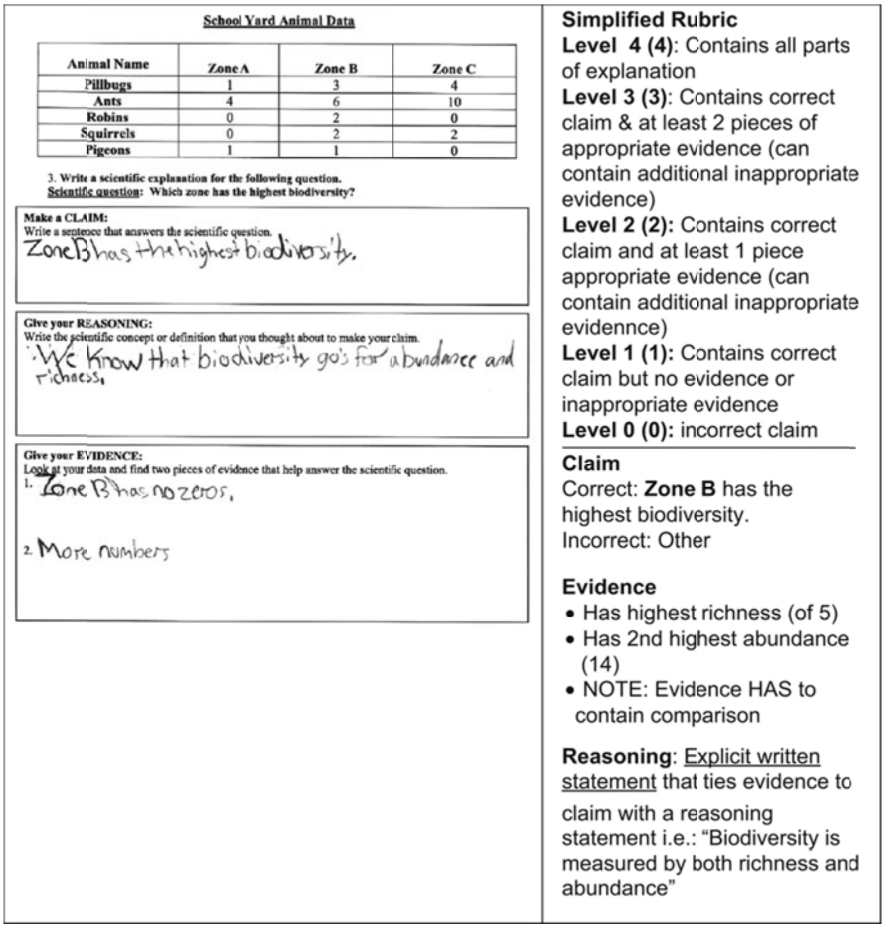
Challenges in Interpreting Students' Responses with the Learning Progression Rubric

The combination of multiple levels of tasks and four levels of coding provides a range of students with the opportunity to provide us evidence of their strengths and weaknesses as mapped to our content and practice progressions. On-going analyses of students' responses to the assessment tasks support our hypothesized coding levels in that lower ability students (identified as such by their teachers) tend to give claims but do not back them up with appropriate or sufficient evidence or reasoning. The middle level students tend to provide claims with evidence, but the evidence is often insufficient to fully back up their claims. Only the highest ability students tend to give responses that include all parts of a scientific explanation: claim, evidence, and reasoning (Gotwals & Songer, 2010). However, while we have been able to discern strong trends in student responses that correspond to our learning progression framework, not all student responses fit neatly into our coding scheme or are easy to interpret. Therefore there are challenges in dealing with these discrepant or vague student responses.

Interpreting responses with vague evidence. One difficulty with interpreting and coding students' responses arises when a level of practice is implied, but the response does not explicitly state the key components of a scientific explanation. The student response in Figure 9 illustrates this difficulty. The item asks students to use a table of animal observations to explain which zone in the schoolyard has the highest biodiversity. A Level 4 respnse to this item would include a claim that Zone B has the highest biodiversity; evidence that Zone B has the highest richness (number of different kinds of animals) and second highest abundance (total number of animals); and reasoning that richness and abundance both play a role in determining the biodiversity of a given area although, because biodiversity is a measure of the variety of organisms in an area, the richness variable in this item is more significant in determining total biodiversity.

The student whose work is presented in Figure 9 has the correct claim: Zone B has the highest biodiversity. In addition, the student approaches the correct reasoning: "We know that biodiversity go's [sic] for abundance and richness." However, this student is not explicit about how he interpreted the data in the table. The evidence "Zone B has no zeros" could mean that Zone B has more types of animals (higher richness) than the other zones (that do have zeros). The evidence "More numbers" could refer either to a higher richness or to the fact that the numbers add to more (abundance). While these statements suggest the evidence that we are looking for in this question, the evidence is not explicit enough to be considered appropriate and sufficient. This response is also not specific enough to code even at Level 2 since in the coding rubric for this item, the evidence has to be a specific comparison of abundance and richness amounts to other zones. We can infer from the student's response that he was able to use the data when formulating the claim, but he is not explicit enough in how he used the data. Specifically, not all data are evidence – data only become evidence when they are interpreted with respect to the scientific question and the theory underlying the explanation (Sandoval, 2003).

We do not have enough information about how the student draws upon the data as evidence for his claim. His response could indicate at least two possible situations. The first situation is that there is a third dimension underlying how students respond to this task – the practice of interpreting data. It is possible, and even likely, that students must be able to find patterns in the data presented in the table before they can formulate an evidence-based explanation. This is a dimension that is not captured in our scoring rubrics and perhaps needs to be considered for all tasks that require interpretation of tables or graphs. However, it seems unlikely that this student was unable to interpret the data and relate it to the ideas of abundance and richness since his claim and reasoning seem to indicate otherwise. Therefore, the second possible situation is that this student's response illustrates a type of middle knowledge that is not fully captured in our coding rubric. It is possible that this student did not know what counts as appropriate and sufficient evidence or the reason for being explicit about the use of evidence. Thus this type of response provides some evidence of the type of middle knowledge that students may have in learning how to use evidence. However, when students are not explicit, we face a challenge in interpreting their responses with respect to our coding rubric.



School Yard Animal Data

Animal Name	Zone A	Zone B	Zone C
Pillbugs	1	3	4
Ants	4	6	10
Robins	0	2	0
Squirrels	0	2	2
Pigeons	1	1	0

3. Write a scientific explanation for the following question.
Scientific question: Which zone has the highest biodiversity?

Make a CLAIM:
Write a sentence that answers the scientific question.
ZoneB has the highest biodiversity.

Give your REASONING:
Write the scientific concept or definition that you thought about to make your claim.
We know that biodiversity go's for abundance and richness.

Give your EVIDENCE:
Look at your data and find two pieces of evidence that help answer the scientific question.
1. Zone B has no zeros.
2. More numbers

Simplified Rubric
Level 4 (4): Contains all parts of explanation
Level 3 (3): Contains correct claim & at least 2 pieces of appropriate evidence (can contain additional inappropriate evidence)
Level 2 (2): Contains correct claim and at least 1 piece appropriate evidence (can contain additional inappropriate evidennce)
Level 1 (1): Contains correct claim but no evidence or inappropriate evidence
Level 0 (0): incorrect claim

Claim
Correct: **Zone B** has the highest biodiversity.
Incorrect: Other

Evidence
- Has highest richness (of 5)
- Has 2nd highest abundance (14)
- NOTE: Evidence HAS to contain comparison

Reasoning: Explicit written statement that ties evidence to claim with a reasoning statement i.e.: "Biodiversity is measured by both richness and abundance"

Figure 9. Student DTF09450322's response with vague evidence.

Responses that include reasoning but no evidence. The response in Figure 9 is also evidence of another type of challenge in interpreting students' responses with our hierarchical coding rubric: some students give responses that include a reasoning statement to back up their claim (a statement of the appropriate scientific principle or phenomenon) but do not include specific evidence. In contrast, our hierarchical coding scheme assumes that students will progress from providing a claim to backing it up with evidence and will include both reasoning and evidence only at the highest level. The student response in Figure 10 provides another example. This complex item asks whether, according to the data chart, large fish are predators, prey, or both. A Level 4 response would include the claim that large fish are both predators and prey, evidence that large fish eat small fish and are eaten by the heron, and reasoning that defines a predator as an animal that eats another animal and prey as an animal that is eaten.

The student response in Figure 10 has a correct claim ("They are both predators and prey"), as well as a reasoning statement that defines predators and prey ("They go after there [sic] food and others go after them"). While this reasoning statement is not as specific or scientific as we would hope for, it gives us some evidence that this student understands what predators and prey are. However, the student does not include specific evidence from the table. Rather, he hints at the evidence when he says that he looked at the table ("I looked at the table and it came to me."). While we can infer what the student saw in the table, he does not include the specific evidence of what the large fish eats and what eats the large fish. The student may know what evidence is appropriate and sufficient, but, since he felt that it was implied in his answer, he did not need to be explicit. However, a characteristic of good supporting evidence is that it uses specific data to back up the claim. Just stating that he used the table does not show that this student really understands what counts as evidence or the reason for being explicit about the evidence.

The responses in Figures 9 and 10 give us interesting information about students' middle knowledge. If we look across time, the responses can provide evidence about the pathways that students take in developing evidence-based explanations. Yet the explanations still pose a challenge when interpreting the responses. Specifically, our rubric, which is based on both the literature and our past research, indicates that students tend to include evidence in their responses before they include reasoning. When responses include reasoning without evidence, they do not fit neatly into our coding scheme. While we are still determining the best ways to deal with such responses, we are currently considering that these responses contain a claim and partial support where partial support can include any type of rationale that supports the claim (evidence or reasoning). In order to accurately capture these types of responses, we are coding the responses as Level 2 with an extra code that indicates that students have included reasoning without evidence. We plan to re-analyze the students' responses, looking for patterns in the items (whether certain items tend to have more responses with reasoning and no evidence) and patterns among the students (whether certain students tend to provide responses with reasoning and no evidence across items).

Interpreting responses that interchange evidence and reasoning. Another challenge that we have faced in our coding concerns students' responses to the scaffolded assessment tasks. Specifically, some students were able to provide multiple pieces of a scientific explanation but interchanged the evidence and the reasoning pieces in spite of the scaffolding. While this does not, necessarily, point to confusion about the purpose of the explanation as a whole, it does point to difficulty in differentiating the identity and purpose of the components of the explanation. A student at the upper level of our practice progression would be able to make this distinction.

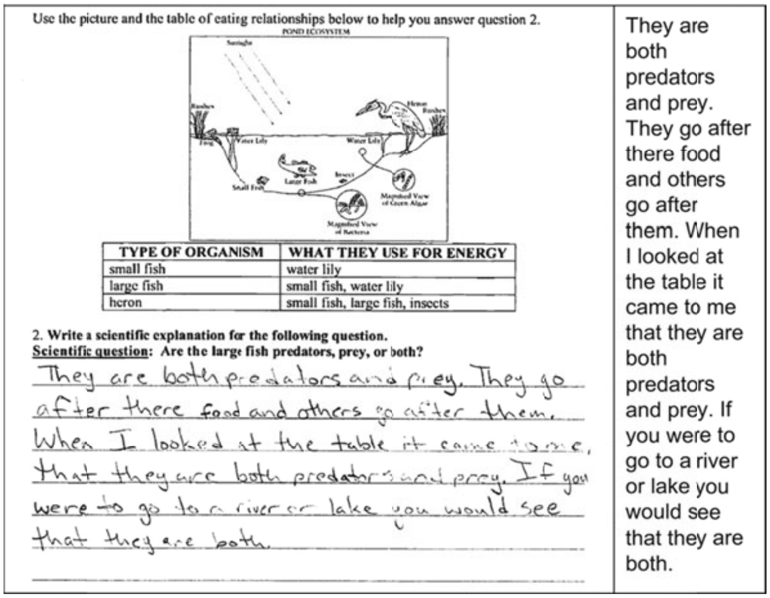


Figure 10. Student DTF09400118's response that includes reasoning but no evidence.

Figure 11 contains a student's response to a scaffolded explanation item that asks whether a heron is an herbivore, carnivore, omnivore, or decomposer. A Level 4 response to this item would provide a claim stating that the heron is a carnivore, evidence that the heron eats small fish, large fish, and insects, and reasoning that carnivores eat other animals or consumers.

In this example, the student correctly claims that the heron is a carnivore and begins to give reasoning for her claim ("it eat's [sic] any kind of small animal"). However, she includes the evidence and reasoning in the wrong boxes.[2] Specifically, when asked for reasoning, she provides evidence from the table that the heron eats insects and small animals. In the box for evidence, the student repeats part of the reasoning that she included in the claim box ("it likes eating small animals") and adds, "the heron is big. It can eat a lot." This reasoning indicates a common misconception that larger animals eat smaller animals, and predators must be larger than their prey (Gallegos, Jerezano, & Flores, 1994).[3] While "it likes eating small animals" and "it eat's [sic] any kind of small animal" are not complete statements of scientific reasoning, they provide evidence that the student has the beginnings of the content idea of what a carnivore is. In addition, this student clearly is beginning to reason from evidence. However, the types of support that she provides, evidence and reasoning, are not differentiated and appear in the wrong boxes. This transposing of evidence and reasoning may indicate that this student is not yet able to link her understandings of how to support her claim with the specific terminology used in the scaffolds. This response may also indicate that the student does not understand the identifying characteristics or purpose of these different explanation components.

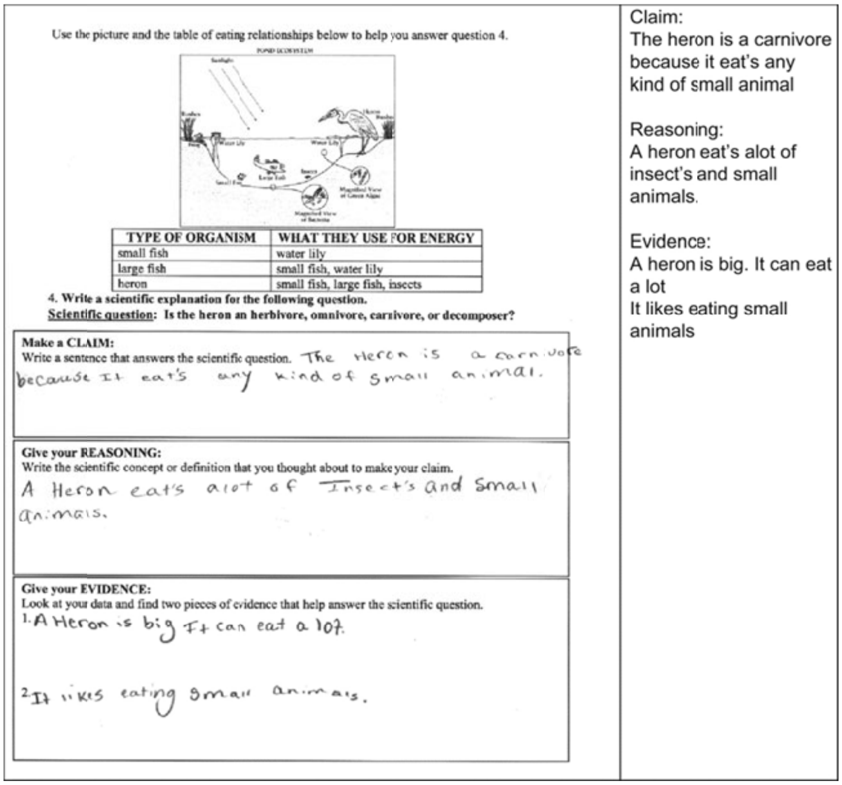


Figure 11. Student DTF09400111's response with evidence and reasoning transposed.

Similar to the way in which we deal with students who provide reasoning, but not evidence, we are now including an extra code for scaffolded items. This code indicates when students have included evidence and reasoning in the wrong places. We plan to re-analyze the students' responses, looking for patterns in both the items (whether certain items tend to have transposed reasoning and evidence) and patterns among students (whether certain students tend to provide responses with transposed reasoning and evidence across items). This analysis will provide us with evidence about the types of middle knowledge that some students have as they begin to fuse content and practices in their scientific explanations. In addition, the analysis will allow us to determine whether our practice progression needs to be refined to include information about students' understandings of the components of evidence-based explanations.

Our content and practice scaffolds allow us to explore the intricacies of how students understand the different components of explanations. For example, the student response in Figure 11 illustrates that some students may be beginning to reason from evidence but might not yet know the meaning of the specific terms (claim, evidence, and reasoning). Therefore, students may provide aspects of their explanation in the wrong box. While some may argue that it does not matter whether students place information in the correct boxes as long as they have sufficiently supported their claim, we agree with the National Research Council (1996, 2007) that understanding what constitutes evidence, recognizing the components of scientific explanations, and knowing how the pieces fit together into a coherent whole are all essential aspects of developing scientific literacy. Students who place components of a scientific explanation in the wrong boxes are not given a lower code in our coding rubric. However, their responses are tagged during coding and later are systematically examined to help us to think about the nature and validity of our practice progression.

Learning from the Challenges with Interpreting Students' Responses

Understanding the ways that students utilize scaffolds as they formulate evidence-based explanations helps us characterize students' middle knowledge and learn what they can do with and without support. However, this middle knowledge is particularly difficult to tease out of students' responses because a characteristic of middle knowledge is that it is messy. Students are often inconsistent or vague in expressing their developing understandings (Gotwals & Songer, 2010). In addition, when students are not clear about each component of their scientific explanation, about what counts as evidence, or about the need to be explicit in stating all components of their explanation, interpreting their responses can pose challenges for placing these responses on our practice progression-based coding rubric. The student responses discussed above pose these challenges.

While we found some responses that did not follow the pattern reflected in the practice progression and some responses that were difficult to interpret based on the evidence that students provided, the majority of the responses followed this sequence and posed none of the challenges described above. In the IRT models, since both the student and the item fit statistics that are within the acceptable range, there is validity evidence for our articulation of the cognition vertex –the fused content and practices learning progression framework. Given the diversity of learners, it is likely that there will always be responses that do not fit neatly into a given progression. As Briggs (chapter 15) writes, "misfit is our friend," meaning that we can learn as much or more from responses that do not fit into our posited framework or model as from responses that do. This conclusion applies not only to statistical modeling but also to more qualitative analyses or interpretations of individual student responses. Even when the quantitative models adequately capture the patterns in students' responses, exploring the few "outliers" can also provide us with information about how students learn.

However, questions remain. What do we do with responses like these? And do responses like these invalidate our whole practice progression? In response to the first question, as explained above, our current method is to introduce an additional code for these responses based on the type of discrepancy and then to analyze them at both the student and the item level. This analysis is currently underway. The findings from this analysis will help us both to iteratively refine our content and practice progressions and to answer the second question about whether these responses invalidate our practice progression. Since the majority of student responses fit into our practice progression, we do not anticipate that our whole practice progression will be invalidated. However, these non-conforming responses are particularly important as we restructure our content and practice progressions. They will also play a large role in subsequent revisions of the assessments and curriculum.

IMPLICATIONS

As our work and that of others indicate, implementing quality assessments mapped to learning progressions is a challenging new area in science assessment design and evaluation. The work in this chapter has illustrated our project's approach to three challenges researchers who create learning progressions-based assessments may face: developing assessments based on a conception of science knowledge that includes fused content and practices; developing tasks that allow a range of students an opportunity to illustrate what they know and can do; and interpreting responses from students with middle knowledge.

Developing Tasks for Fused Content and Practices

The push to include the fusion of content and practices in the conceptualization of how students should learn about (and are assessed in) science (e.g., NRC, 2007) has forced many learning progressions researchers to think about how to develop cognitive models that combine how students develop what they "know" with what they "can do" with this knowledge. Developing tasks that require students to fuse multiple dimensions of content and practices can pose challenges.

Our project has conceptualized the cognition vertex of the assessment triangle as having two dimensions – a content progression and practice progression – that we integrate within each assessment task, curricular activity, and coding rubric. Our approach, however, offers just one way of addressing this challenge. We cannot, feasibly, design or implement an exhaustive set of items where each content idea is assessed at each level of the practice progression. Thus, in picking and choosing the content ideas to fuse at different level(s) of practice, we limit the types of information we can gather. In other projects, such as the environmental literacy project, the design of their assessments uses scenarios that require students to reason with content. Student responses are then interpreted using a single (though multifaceted) learning progression for how students reason [practice] with content (e.g., Jin & Anderson, this volume). For example, the environmental literacy's learning progression includes four levels with increasingly sophisticated ways of explaining phenomena. While the nature of the explanation may vary across contexts, at a given level, students are expected to provide explanations with similar characteristics (such as use of force-dynamic reasoning). While there is no one right way of approaching design decisions about fusing content and practices into learning progression assessments, it is important to make clear the purpose of the assessments (in our case, to examine learning based on our curricular units) and to be clear about the tensions and trade-offs in these decisions.

Providing Scaffolds in Assessment

Learning progressions outline pathways through which we can support students in moving past what they can do on their own towards more sophisticated knowledge. Learning progression assessments, then, must be able to elicit responses from students at multiple levels and provide evidence to locate students on this pathway. Our design decision to embed scaffolds in both our practice progression and in our assessment tasks allows a range of more and less guided opportunities for students to demonstrate their fused content and practice knowledge. The incorporation of these scaffolds into our assessments allows us to tease out where students may understand the content but may not be able to create a coherent evidence-based explanation, or where students may understand the components of an explanation but do not have the content knowledge to fully support their claims using appropriate and sufficient evidence or reasoning. In addition, the practice scaffolds serve as a resource for unpacking and characterizing the nature of students' middle knowledge so that we may better understand the intricacies of how students develop more sophisticated scientific explanations.

However, specifying levels from each progression in the creation of tasks can limit the levels of performance that students can demonstrate. For example, in a scaffolded task, students who may not need the scaffolds are unable to demonstrate their fused content and practices knowledge in an unsupported environment. In addition, the scaffolds can sometimes create unexpected difficulties for students, perhaps because of below-grade-level reading comprehension or because of the complex structure and appearance of the task. Thus, while there are benefits from using scaffolds that give us insight into students' middle knowledge, there are associated trade-offs since we may not get a fully accurate picture of everything that students are able and unable to do.

Interpreting Middle Knowledge

An important dimension of our work on learning progression-based assessments is characterizing the nature of students' developing understandings in the form of the range of middle knowledge that they express. However, when assessing fused content and practices knowledge, there is a significant possibility of multiple interpretations of students' responses – both in terms of identifying all the ways that students might express aspects of fused middle knowledge and in attributing any incorrect ideas to either content misconceptions or to less than fully-formed practices.

We have used our content and practice progressions to develop rubrics to interpret and code students' responses to our assessment items. While not unique to learning progression-based assessment, when assessing multiple dimensions of students' understandings (such as content and practices) there are often difficulties in interpreting and inferring students' knowledge from their written (and verbal) responses. This is especially true when student responses imply an understanding but the thinking behind these responses is unclear. We have highlighted the messy middle knowledge associated with our practice of building evidence-based explanations; however, our rubrics do not include content-based misconceptions. While it may be possible to do develop rubrics that can gather evidence of the students' messy middle knowledge in fused content and practice, placing students at a single level (or giving them a single code) for both aspects of middle knowledge is challenging.

CONCLUSIONS

Our assessment approach was developed in response to our assertion that if assessments only examine students' knowledge of content, then inquiry is devalued by both teachers and students and is less likely to occur in the classroom (Schafer, 2002). Through a focus on assessments designed to characterize the range and nature of middle knowledge associated with building evidence-based explanations, our approach can contribute information not available from learning progression assessments that only focus, for example, on either content progressions or practice progressions.

This chapter has primarily examined students' responses that have challenged our thinking about the nature of their middle knowledge. While the good fit of the statistical models that we used to analyze students' responses to our assessment items provided validity evidence that our fused progression adequately represents low, middle, and high level responses, some students did not fall neatly into our hypothesized patterns. While it may be tempting to avoid looking any deeper after finding that the statistical models adequately fit the data, it is important to look at the different ways that students express their knowledge and the extent to which these ways are captured in our learning progression framework. The students' responses provide us insight into the messiness of developing fused science knowledge. We are currently working to refine our learning progression framework to capture these different types of middle knowledge. However, as we realize that students' middle knowledge is often messy, this work may be difficult.

ACKNOWLEDGEMENTS

This research was funded by the National Science Foundation under Grant ESI-0628151 to the Deep Think project at the University of Michigan and at Michigan State University. The opinions expressed herein are those of the authors and not necessarily those of the NSF.

NOTES

1 Both tasks are based on 4th grade ecology ideas. The first is E3: Most animals use particular kinds of organisms for food. Some general groups are herbivores, carnivores, omnivores, and decomposers. The second is E4: An animal that eats another organism is a predator, and the organism that it eats is its prey…

[2] The boxes are scaffolds to remind students to include each component of the explanation – claim, evidence, and reasoning.

[3] Our coding rubric does not specifically categorize content misconceptions into separate categories or levels; rather, it focuses mainly on students' use of content in the formulation of evidence-based explanations. However, our curricular units (Songer, 2006) identify common misconceptions about each activity and also identify methods teachers can use to elicit and respond to students' ideas.

REFERENCES

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education, 22*, 797–817.

Davis, E.A.,& Krajcik, J.S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher, 34*(3), 3–14.

DeBarger, A. H., Quellmalz, E., Fried, R., & Fujii, R. (2006, April). *Examining the validities of science inquiry assessments with cognitive analyses.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*, 287–312.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Erudan, S., Simon, S., & Osborne, J. (2004). TAPing into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*, 915–933.

Gallegos, L., Jerezano, M. E., & Flores, F. (1994). Preconceptions and relations used by children in the construction of food chains. *Journal of Research in Science Teaching, 31*, 259–272.

Germann, P. J., & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching, 33*, 773–798.

Gotwals, A. W. (2006). *The nature of students' science knowledge base: Using assessment to paint a picture* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3237965)

Gotwals, A.W., & Songer, N.B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education*, *94*, 259–281.

Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London, UK: Sage Publications.

Jimenez-Aleizandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education, 84*, 757–792.

Lee, H.-S. (2003). *Scaffolding elementary students' authentic inquiry through a written science curriculum* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3096136)

Lee, H.-S., and Songer, N.B. (2003). Making authentic science accessible to students. *International Journal of Science Education, 25*, 1–26.

Lizotte, D. J., Harris, C., McNeill, K., Marx, R., & Krajcik, J. (2003, April). *Usable assessments aligned with curriculum materials: Measuring explanation as a scientific way of thinking.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

McNeill, K., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 233–265). New York, NY: Taylor & Francis.

McNeill, K., Lizotte, D., J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences, 15*, 153–191.

Metz, K. E. (1991). Development of explanation: Incremental and fundamental change in children's physics knowledge. *Journal of Research in Science Teaching, 28*, 785–797.

Metz, K. E. (2000). Young children's inquiry in biology: Building the knowledge bases to empower independent inquiry. In J. Minstrell & E. H. van Zee (Eds.), *Inquiry into inquiry learning and teaching in science* (pp. 371–404). Washington, DC: American Association for the Advancement of Science.

Mislevy, R. J., Steinberg, L.S., & Almond, R.G. (2003).On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education*

Standards: A guide for teaching and learning. Washington, DC: National Academy Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

National Research Council. (2011). A *framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching, 41*, 513–536.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences, 12*, 5–51.

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*, 23–55.

Scardemalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *Journal of the Learning Sciences, 1*, 37–68.

Schafer, W. (2002, August). *Describing assessment for teaching and learning.* Paper presented at the Optimizing State and Classroom Tests: Implications of Cognitive Research for Assessment of Higher Order Reasoning in Subject-Matter Domains conference, University of Maryland, College Park.

Songer, N.B. (2006). BioKIDS: An animated conversation on the development of curricular activity structures for inquiry science. In R. K. Sawyer, (Ed.) *The Cambridge handbook of the learning sciences* (pp. 355–369). New York, NY: Cambridge University Press.

Songer, N. B., Kelcey, B., & Gotwals, A.W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning in biodiversity. *Journal of Research in Science Teaching*, *46*, 610–631.

Songer, N.B. & Gotwals, A.W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal for Research in Science Teaching, 49*, 141-165.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). Mapping to know: The effects of representation on guidance and reflective assessment on science inquiry. *Science Education, 86*, 264–286.

Toulmin, S. (1958). *The uses of argument*. New York, NY: Cambridge University Press.

White, B., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*, 3–118.

ALICIA C. ALONZO, TERESA NEIDORF AND CHARLES W. ANDERSON

USING LEARNING PROGRESSIONS TO INFORM LARGE-SCALE ASSESSMENT

In 2006, the National Research Council (NRC) recommended that states use science learning progressions (LPs) to align curriculum, instruction, and assessment around big ideas. The State of Massachusetts is currently exploring the use of LPs to revise its Science and Technology/Engineering standards (Foster & Wiser, this volume). However, to date, neither Massachusetts nor any other US state has incorporated science LPs into its assessment system. At the national level, the new science assessment framework and specifications for the 2009 National Assessment of Educational Progress (NAEP) call for the use of LPs (National Assessment Governing Board [NAGB], 2007, 2008). The framework states that LP research can be used to "inform the generation of related performance expectations across grades" (NAGB, 2008, p. 90) and to develop assessment items. Both the NRC and NAGB were seeking ways to incorporate findings and methods from research on students' science learning into large-scale assessment systems; they both pointed to LPs as a promising vehicle for accomplishing this goal.

While the NAEP specifications document (NAGB, 2007) includes some examples of LPs and related items, the framework (NAGB, 2008) provides little guidance as to how LPs might actually be included in the assessment or reports of assessment results. Indeed, the framework includes several caveats about LPs that may influence the extent to which they can be used appropriately in large-scale assessment contexts.[1] Although LP-based items are not part of the 2009 NAEP operational assessment, their inclusion in the science framework has raised interest in the potential role of LPs in large-scale assessments.

There is wide variation in how those in the research and policy communities define "learning progressions" as well as in what might be considered "large-scale assessments." The 2006 NRC report defined LPs as "descriptions of the successively more sophisticated ways of thinking about an idea that follow one another as students learn" (p. 48) but provided little guidance as to the form these frameworks should take. Both this report and the NAEP specifications (NAGB, 2007) include LP examples that vary in scope (from one representing 13 years of instruction to one representing learning in a single curriculum unit). Thus for our definition, we refer to a 2007 NRC report that expanded the earlier definition of LPs to specify that they cover "a broad span of time (e.g., 6 to 8 years)" (p. 219). While LPs with a much smaller scope may be needed for other applications (such as teachers' formative assessment practices; Alonzo, 2009), for the types of

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 211–240.

large-scale assessments that are the focus of this chapter (see below), a broad scope provides an appropriate level of detail about expected changes in student thinking. In addition, consistent with many of the examples in the NRC and NAGB documents, we view LPs as more than orderings of content to be mastered. We embrace the idea of LPs as descriptions of students' thinking that necessarily include correct and complete aspects of students' knowledge as well as developing—albeit not completely accurate—aspects. In addition to variation in scope and in what is considered to progress in a LP, current work on LPs in science reflects variation in what ought to be included—whether it is just the framework (the description of student thinking) or the framework plus associated assessments and/or instructional materials (Corcoran, Mosher, & Rogat, 2009). Therefore, in the rest of this chapter, we use the phrase "learning progression framework" (abbreviated as "LP framework") to indicate descriptions of students' thinking (without associated resources). This is consistent with the way the NRC and NAGB documents use the phrase "learning progression." For simplicity, we use "LP-based" to label items, assessments, and assessment approaches that are based upon LP frameworks and "LP levels" to refer to levels of a LP framework.

Given the recommendations in national policy documents (e.g., NRC, 2006; NAGB, 2007, 2008), our focus in this chapter is on large-scale summative assessments at the state, national, and international levels that provide information about student achievement in the broad domain of science. This includes assessments such as NAEP and TIMSS[2] as well as state accountability assessments for grades K-8; it excludes subject-specific assessments—such as the SAT®, the GRE®, Advanced Placement (AP®) exams, state or local end-of-course assessments, and benchmark or interim assessments. This focus is certainly not intended to reflect a judgment on the utility of LP frameworks for these other assessments. Indeed, we see much potential in the use of LP frameworks to inform subject-specific and benchmark assessments. In both applications, a LP framework may provide coherence among curriculum, instruction, and formative and summative assessment. However, consideration of these applications is beyond the scope of this chapter, in which we use "large-scale assessment" to refer only to the types of assessments listed above.

Clearly, the assessment landscape addressed in this chapter includes significant variation. Most fundamentally, national/international assessments such as NAEP and TIMSS characterize populations of students, while state accountability assessments describe achievement of individual students (and schools through the aggregation of individual scores). As a result, assessments such as NAEP and TIMSS assess only a small fraction of the students in the target population (with different examinees answering different questions), while state accountability assessments are administered to all students (often with a common test). Despite this difference, all of these assessments are designed to report student achievement in the broad domain of science (often subdivided into areas such as earth, physical, and life science) rather in a specific area of science (such as force and motion or genetics). In addition, these assessments seek responses from students who have experienced a wide range of different learning environments, both in-school and out-of-school. While there are important differences among the assessments included in this chapter in terms of details of format, development, administration, and reporting, the underlying criteria for evaluating items to be included in these assessments are similar. We believe that these commonalities permit meaningful discussion of issues that must be considered by anyone interested in using LP frameworks to inform large-scale assessment development.

As explored in this chapter, LP frameworks represent a fundamentally different way of conceptualizing and reporting student achievement. LP-based items may be used to align large-scale assessments with the best current research in the field. In particular, (a) assessments may be grounded in research on student learning, (b) assessments and reports of assessment results may be organized around a smaller number of strands or big ideas that show continuity across grade levels and are assessed in greater depth, and (c) student learning may be described in terms of qualitatively changing knowledge and practice rather than mastery of a checklist of standards or benchmarks.

LP frameworks may push large-scale assessment systems towards a new vision of student achievement. Eventually, LP frameworks may result in entirely new types of assessments; however, in this chapter we focus on how LP-based items may be incorporated—in a limited way, at first—into existing large-scale assessment systems. There are two very practical reasons for this focus. First, we are more likely to make small adjustments to current systems than to create an entirely new assessment system. Since significant infrastructure exists for current assessment systems, it seems unwise not to capitalize upon this infrastructure for assessing large numbers of students. Second, currently it is not possible to create an assessment including only LP-based items (due in large part to the limited availability of validated LP frameworks covering the entire science curriculum).

Incorporation of items based on well-researched—but not yet fully validated—LP frameworks into large-scale assessments (particularly as part of pilot/field testing) allows us to collect data from a representative sample of the target population. Researchers often do not have access to such representative samples; thus data collected as part of large-scale assessments may be especially vital for refining the LP framework and associated items to ensure that they represent the target population rather than the (often narrower) population participating in research studies. Therefore, we see great potential in using LP frameworks to inform existing large-scale assessments—both for pushing existing assessments towards greater alignment with current research and for laying the groundwork for new types of assessments that may be envisioned in the future.

However, it is our contention that a LP-based approach to student assessment is not necessarily compatible with current large-scale assessment systems. This is an argument we develop in the chapter as the basis for considering the challenges associated with incorporating LP-based items into the current large-scale context. In particular, we address a central question: How can LP-based items fit into a large-scale testing apparatus designed to make different claims about student achievement? Consideration of this question necessarily involves an acknowledgement of the requirements and constraints imposed by large-scale assessment systems as well as an exploration of the ways current practices can be modified in order to (a) heed calls for the use of LPs in large-scale assessments and (b) take advantage of the opportunities LP frameworks may offer to assess science content in greater depth. We are writing this chapter at a time when important changes in large-scale assessment systems are "on the horizon"; these changes may significantly impact the compatibility between current large-scale assessment systems and LP-based items.[3] In this chapter, we focus on current realities of the large-scale assessment context. We are hopeful that future developments will address some of the challenges we discuss in the chapter.

Because the NAEP assessment framework is the first to recognize the potential role of LPs and because of our familiarity with this assessment system, we use NAEP as a context for exploring challenges, noting contrasts with other types of large-scale assessment where relevant. We begin by discussing differences between the purpose of a LP-based approach and those of current large-scale assessment systems. We follow this discussion with a detailed consideration of contrasting practices related to (a) item development, (b) pilot/field testing, (c) item analysis and evaluation, (d) design of operational assessments, and (e) scoring and reporting. We conclude the chapter with recommendations for both researchers and assessment developers.

CONTRASTS BETWEEN A LP-BASED APPROACH AND CURRENT LARGE-SCALE ASSESSMENT SYSTEMS

As described above, LP frameworks have the potential to transform large-scale assessment of student achievement in science. However, realization of this potential depends critically upon the alignment between LP-based approaches and large-scale assessment systems, in terms of both purposes and specific features of the testing apparatus.

This section explores these contrasts, starting with the different purposes underlying the LP-based approach and current large-scale assessment systems. This discussion is followed by consideration of differences in assessment practices. The former discussion represents policy considerations: whether large-scale assessment systems should incorporate the different purpose (and vision of student achievement) associated with a LP-based approach. The latter discussion represents design considerations: the adjustments needed in current practices for implementation a LP-based approach in a large-scale assessment context.

Contrasting Purposes: Policy Considerations

Current large-scale assessment systems. Large-scale assessments typically make claims about whether students (or groups of students) have reached particular benchmarks for their grade level. These benchmarks often describe students' mastery of broad areas of the curriculum, such as "science." For example, all NAEP assessments define three levels of achievement—basic, proficient, and advanced:

> **Basic** denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade. **Proficient** represents solid

> academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter. **Advanced** signifies superior performance (NAGB, 2008, p. 10).

Although examples are provided to illustrate what students at each achievement level for a particular grade level "should know and be able to do in terms of the science content and practices identified in the [NAEP] framework" (NAGB, 2008, p. 127), NAEP reports focus primarily on broad descriptions of student achievement. State and local assessments define levels of achievement in similarly broad ways. Indeed, the No Child Left Behind (NCLB) Act of 2001, which influences much current state assessment policy, requires that states' academic achievement standards describe three levels of achievement—two high levels (proficient and advanced) and one lower level (basic). For both NAEP and state assessments, these levels are designed to indicate the extent to which students (or groups of students) have mastered a set of desired knowledge and skills (defined by the assessment framework) for a particular grade level.

Despite this general commonality of purpose, large-scale assessments differ in the particular types of claims they make. The NAEP assessments make claims about groups of US students at three grade levels (four, eight, and 12)—both overall and for key reporting subgroups. Results are reported as average scale scores and percentages of students who reach each NAEP achievement level. NAEP releases a national-level "report card" and also produces reports based on state- and urban district-level results. In addition to providing a "snapshot" of students' achievement in a particular year, the NAEP assessment system reports trends in student achievement over time.

In contrast, state assessments make claims about individual students. Although these results may be aggregated at the school level as the percentage of students achieving a level of proficient or higher, this aggregation is based upon scores for individual students. States produce reports for various stakeholders, including students, teachers, parents, schools, districts, and the public. While some states currently use growth models to track the performance of individual students from year to year, all states report the percentages of students who demonstrate achievement at each performance level (overall and by subgroups) to meet NCLB requirements. Although this reporting is similar to NAEP's reporting of the percentages of students at each achievement level, the specific trend reporting and item release requirements for NAEP and state assessments may be quite different.

LP-based approach. At a superficial level, LP frameworks may seem similar to standards documents or assessment frameworks in that they specify performances associated with student knowledge and practice at different levels of sophistication. Both standards-based assessments and LP-based assessments seek to measure and report students' increasing scientific competence, but the claims they make are different. Standards-based assessments measure and report the acquisition of scientific knowledge, describing competence in terms of the standards that have and have not been mastered. In contrast, LP-based assessments make claims about where the performances of students (or groups of students) fit in a succession of increasingly sophisticated ideas and practices. In contrast to typical achievement levels (as described above), LP levels articulate the ideas and practices characteristic of students at a given level of sophistication. Importantly, these descriptions do not include only scientifically correct ideas; at lower levels they also identify non-canonical ways of thinking that may be useful for learners in terms of their present ideas and practices and may be productive, leading to more sophisticated understandings. Just as "pioneer species" in an ecosystem can create the conditions in which other organisms can flourish, non-canonical ideas and practices can represent steps toward mature scientific understanding. Thus LP-based assessments provide evidence of succession in a *conceptual ecology* (Posner, Strike, Hewson, & Gertzog, 1982; Toulmin, 1972[4]). Rather than considering the acquisition of scientific knowledge to be indicated solely by an increase in the number of items that students are expected to answer correctly, this view focuses on the evolution of students' conceptual ecologies from lower to higher levels of sophistication.

Ideally, standards-based assessment frameworks reflect coherence from grade level to grade level, with expected knowledge and practices increasing in sophistication with increasing grade level. However, because these assessments report on students' mastery of the standards at a given grade level, there is typically little consideration of how performance at that grade level relates to expectations for the grades above and below. In contrast, LP-based assessments place student performances within a larger framework, typically one spanning multiple grade levels. Although LP-based assessments may still specify a desirable level of performance at a given grade level, the goal of such assessments is not simply to report whether students have achieved that level. Instead, claims from LP-based assessments characterize students' knowledge and practices, even if they are above or below expectations for a given grade level. As discussed below, LP-based assessments may require opportunities for students to express a greater range of knowledge and practices than standards-based assessments do since the latter focus on a narrower range of competencies. In contrast to a standards-based assessment, which administers an item tapping a given standard only at the associated grade level, a LP-based assessment might administer the same (or a similar) item to students at multiple grade levels.

Necessarily, LP level descriptions—and the corresponding claims about students' knowledge and practices—are related to specific areas of the science curriculum. These claims contrast with claims by large-scale assessment systems that span much broader content domains. For example, student achievement on the NAEP science assessment is reported in terms of science overall as well as for the broad content areas of life, physical, and earth science (Grigg, Lauko, & Brockway, 2006). In contrast, LP-based assessments yield claims about students' knowledge and practices for much narrower domains, which may cut across traditional content areas. For example, drawing upon examples in this volume, LP-based assessments may provide information about students' levels of understanding of celestial motion (Plummer, this volume) or carbon-transforming processes (Jin & Anderson, this volume).

Contrasting Practices: Design Considerations[5]

The use of LP frameworks to inform large-scale assessment systems requires some level of agreement about the purpose of the LP-based approach; as described above, this is a policy consideration. However, this "buy-in" is only the first step in incorporating LP-based items into a large-scale assessment system. The purpose behind existing large-scale assessment systems influences every aspect of their design and use. As such, the incorporation of LP-based items would necessitate some adjustments to current practices. These possible adjustments are discussed in the next five subsections: (a) item development, (b) pilot/field testing, (c) item analysis and evaluation, (d) design of operational assessments, and (e) scoring and reporting.

Item development. In considering the use of frameworks to inform item development, it seems important to distinguish between content progressions—which use research evidence to sequence particular science content and practices—and LPs—which attempt to describe the way students' thinking develops over time (within and across grades). While current standards and assessment systems use content progressions to establish expectations for students at different grade levels,[6] the LP-based approach to item development requires attention to students' conceptual ecologies—the ideas characteristic of students at a given level of sophistication (including ideas that are fully correct as well as those that are still being refined).

LP-based item development is complicated by the nature of LP frameworks themselves. Typical content frameworks that guide the development of large-scale assessments are relatively stable. For example, the earlier NAEP framework (NAGB, 2004) was used for a ten-year period (1996–2005). As policy documents, these frameworks are revised in response to changing expectations for what students should know and be able to do. While revisions may be based upon research findings, such findings are not their primary influence. Many other considerations—such as typical curricula for the target population—also come into play. In contrast, LP frameworks are hypotheses about the way student thinking develops (Corcoran et al., 2009; NAGB, 2007). As such, they are subject to evaluation and revision in light of additional evidence. Thus the LP frameworks that guide item development must be considered to be provisional, with revisions expected in response to data collected as the items are administered. Importantly, because large-scale administrations (of both pilot/field and operational assessments) offer opportunities to collect data from a much larger, more representative sample of students than is typically available to researchers, it is expected that LP frameworks and associated items will change in response to the data collected in large-scale assessment systems.

Item development for typical large-scale assessments is a linear process in which frameworks are written and then used to inform item development. In contrast, the development of LP frameworks and associated assessment items is an iterative process, consistent with the "assessment triangle" described in *Knowing What Students Know* (NRC, 2001). Although the highest level of a LP framework can be defined according to standards-based expectations, the lower levels must be constructed empirically, based upon actual student performance data (NRC, 2007). Thus the LP frameworks and associated assessments co-evolve, with students' responses to assessment items used to both validate and revise the frameworks.

Consistent with the overall purpose of traditional large-scale assessments, a large item pool is developed to cover the assessment framework, with each independent item designed to measure a specific assessment objective in order to allow evaluators to make judgments about students' mastery of particular science knowledge and/or practices. In contrast, a LP-based approach requires coherent sets of items that provide evidence of students' thinking about a specific concept. This evidence may be used to determine students' LP levels. This determination involves evaluating the level of sophistication of students' thinking rather than assigning a score based solely on the accuracy and completeness of their answers. While an individual item (or small set of items) may be used to determine whether student thinking reflects a particular level of the LP framework, this information is most useful when one is reasonably confident that a student's thinking is consistent with a subset of the levels of the LP framework. Without a priori information about students' general location on the LP framework, items that elicit responses at a range of levels are more consistent with the LP-based approach. This is because—rather than checking for particular understandings expected at a given grade level—the LP-based approach characterizes students' thinking even if it is not consistent with expectations for their grade level. However, items that elicit responses at a range of levels are challenging to write. In order to obtain an accurate evaluation of the sophistication of students' thinking, the item prompt and options for multiple-choice (MC) items or instructions for constructed-response (CR) items must make sense to students who have a wide range of understandings (possibly at different grade levels) and must elicit their highest level response. Challenges also arise in differentiating between responses at two adjacent levels for a given item because adjacent levels may differ on aspects of student thinking not tapped by a particular item (see the example item in Figure 2).

LP-based items may look similar to more traditional items, but there are important differences in how LP-based items are designed and scored. Given the purpose of LP-based items, the scoring must yield information with respect to the levels of the LP framework rather than simply indicate whether the response is correct or incorrect—or (for polytomous items) the extent to which required components of a complete and correct answer are included.

In the following subsections, we give descriptions and examples of some common LP-based item formats. We link our examples to a LP framework (Table 1) that describes students' explanations of phenomena involving physical change. This LP framework is derived from the carbon LP framework (see Jin & Anderson, this volume). We have expanded and slightly revised the carbon LP framework to address matter-transforming processes involving physical change. Thus neither the LP framework nor the associated items have been empirically validated.[7]

Table 1. Learning Progression Framework for Explanations of Physical Change Phenomena.

General Level Description	*Characteristics of Explanations*
Level 4: Linking phenomena at different scales with conservation of matter	• Macroscopic phenomena are explained in terms of accurate changes at the atomic-molecular scale. – Explanations for phase changes provided in terms of atoms or molecules (e.g., organization and/or movement at the atomic-molecular level). • Conservation of matter constrains explanations at both macroscopic and atomic-molecular scales and for all phases of matter. – Chemical identity is preserved at the molecular level (e.g., the molecules involved in ice, water, and steam are the same).
Level 3: Linking phenomena at different scales with unsuccessful constraints	• Macroscopic phenomena are explained in terms of inaccurate changes at the atomic-molecular scale. – Chemical identities may be altered during phase change (e.g., air molecules may be thought to change into water molecules during condensation). – Macroscopic properties are attributed to atoms and molecules (e.g., during thermal expansion, the metal molecules expand; during freezing, molecules of water change into molecules of ice). • Conservation of matter constrains explanations in the solid and liquid phases but not the gas phase. – Mass is conserved for transformations involving solids and liquids. – Gas molecules may be viewed as having less mass than molecules in the liquid or solid phase; therefore, substances gain or lose mass during phase transformations involving the gas phase.
Level 2: Force-dynamic accounts with hidden mechanisms	• Readily observable phenomena are linked to hidden mechanisms with limited conservation of matter constraints. – Matter is considered to have mass even if it is not readily perceptible (e.g., gases and small amounts of liquids and solids have mass); however, readily perceptible matter is thought to have more mass (e.g., a pile of fine sand is thought to have less mass than the rock that was ground to create the sand). – In physical change involving a single phase, type of matter is conserved (e.g., fine sand can be seen as tiny pieces of rock). – In phase changes, one substance may be turned into another (e.g., water is turned into ice during freezing; air is turned into water during condensation).
Level 1: Macroscopic force-dynamic accounts	• Focus on perceptions of readily observable phenomena without conservation of matter constraints. – Gases and small amounts of liquid and solid substances are not considered to have mass. – Ice, water, and steam are viewed as three different substances that may appear or disappear without concern for conservation of matter (e.g., when liquids evaporate, they are thought to disappear).

Ordered multiple-choice (OMC) items. The OMC item format was proposed specifically to assess student learning with respect to hierarchically ordered descriptions of understanding—such as LPs (Briggs, Alonzo, Schwab, & Wilson, 2006). As illustrated in the example in Figure 1, these items look like typical MC items. However, instead of one correct and three incorrect options, all of the options in an OMC item are mapped directly to levels of the LP framework, such that all options contribute information about students' LP levels. Thus, instead of a dichotomous judgment about students' mastery of item content, each response option in an OMC item is assigned a score that coincides with the LP level it is intended to reflect. Rather than choosing the "correct" or "best" answer, the intent is that students will select the option that is most consistent with their thinking about the topic. Because a given LP level may include a variety of different ways of thinking about the targeted content, having more than one option at the same level (such as in the example in Figure 1) makes it more likely that students will find an option corresponding to their thinking about the item content. In traditional MC items, the response options typically represent fairly distinct—and sometimes mutually exclusive—ideas. In contrast, there may be only subtle differences between the options in an OMC item since the items attempt to differentiate between students' ideas that, while similar, differ in sophistication.

After taking a hot shower, Steven noticed that the bathroom mirrors were covered with droplets of water. Where did the water come from?

a) The hot air in the bathroom turned into water. (Level 2)
b) Molecules of air became molecules of water. (Level 3)
c) Water appeared from the hotness of the shower. (Level 1)
d) Water in the air collected on the mirror. (Level 4)
e) Water sprayed onto the mirror from the shower. (Level 1)

Figure 1. Sample OMC item linked to the learning progression framework in Table 1.

The item in Figure 1 illustrates a common challenge in developing OMC items for LP-based assessments: writing options that are appropriate for students at varying levels of the LP framework. In the "physical change" LP framework, there is a sharp distinction between Level 2 and Level 3 in terms of students' use of the words "atoms" and "molecules" to explain macroscopic phenomena. However, the use of the word "molecule" may either confuse students at lower levels of the progression or cue them to select an option with a "science-y" word that does not truly reflect their understanding of the phenomenon. In this item, the highest level response does not include the word "molecule" (or the word "condensation") in order to avoid cuing the highest level response with scientific language. However, this response is still consistent with Level 4, at which students are expected to understand that water (even when invisible to the naked eye) must be traced throughout the scenario without a change in chemical identity.

Multiple true-false (MTF) items. Another selected-response item format, MTF, may be useful for LP-based assessment. In these items, students are presented with a list of objects or conditions (e.g., physical materials, organisms, states of matter, celestial bodies) and are asked to determine whether each meets a given criterion. Alternatively, students are presented with a set of related statements and are asked to select all statements that apply to a given phenomenon. Patterns in students' responses can be linked to particular levels of the LP framework. These items are scored based on the LP level indicated by the response pattern rather than by the number of statements correctly identified by the student. An example of a MTF item is shown in Figure 2. In the scoring guide for this item, both Level 4 and Level 3 responses demonstrate conservation of mass/matter. Level 4 is distinguished from Level 3 by the recognition that macroscopic differences between ice and water are not due to changes in the water molecules.

The item in Figure 2 illustrates another common challenge: the difficulty of writing items that allow differentiation among all levels of the LP framework. In this case, it is not possible to differentiate between students at Level 1 and Level 2 since students at both levels are likely to think the mass of a liquid is different from that of a solid.

ITEM: Place an "X" in each row of the table below to indicate properties that either change or do not change when ice melts to form water.

Property	Changes	Does NOT change
Mass		
Volume		
Number of molecules		
Size of molecules		
Type of molecules		
Hardness of molecules		
Distance between molecules		

SCORING GUIDE:

LP Level	Response
4	• Changes: Includes VOLUME and DISTANCE BETWEEN MOLECULES • Does NOT Change: Includes MASS, NUMBER OF MOLECULES, SIZE OF MOLECULES, TYPE OF MOLECULES, and HARDNESS OF MOLECULES
3	• Changes: Includes one or more of the following: o SIZE OF MOLECULE o TYPE OF MOLECULE o HARDNESS OF MOLECULE • Does NOT Change: Includes MASS and NUMBER OF MOLECULES
1 or 2	• Changes: Includes MASS

Figure 2. Sample MTF item (with associated scoring guide) linked to the learning progression framework in Table 1.

Constructed-response (CR) items. Typical CR items often attempt to elicit the mental models that students use to explain science phenomena; such items may also be used to assess student understanding with respect to a LP framework. However, while scoring guides for typical CR items evaluate the accuracy of students' responses and the extent to which they reflect mastery of the targeted content, a LP-based approach requires scoring guides that are linked to levels of the LP framework and that reflect the nature of thinking revealed by the responses. Unlike CR items in typical science assessments, in which specific scoring guides are developed to score individual items, LP-based assessments require more holistic scoring guides that are based on the underlying LP framework. LP-based scoring guides link to general LP framework levels, show how to apply these levels to the particular item, and provide specific examples that clearly illustrate the nature of responses at each level. Experts in the particular LP framework should be involved in developing the scoring materials.

An example of a CR item is shown in Figure 3. This item is a CR version of the OMC example in Figure 1. The response options in the OMC version should reflect the common types of responses expected for the open-ended version of the item. Therefore, students' responses to CR items can be used to develop OMC items.[8]

ITEM: After taking a hot shower, Steven noticed that the bathroom mirrors were covered with droplets of water. Where did the water come from?

SCORING GUIDE:

LP Level	Response
4	Response indicates that water droplets on the mirror originated as water vapor in the air.
3	Response indicates that molecules of another substance were turned into water (e.g., air molecules turned into water molecules);
2	Response indicates that another substance was turned into water (e.g., the air in the bathroom turned into water)
1	Response indicates that water appears due to some macroscopic feature of the situation (e.g., hotness of the bathroom or coldness of the mirror) AND/OR response indicates some macroscopic mechanism for the appearance of the water (e.g., water sprayed directly onto the mirror)

Figure 3. Sample CR item (with associated scoring guide) linked to the LP framework in Table 1. For both Level 3 and Level 4, the specific use of the word "molecules" is not required. To distinguish between Level 2 and Level 3, students at Level 3 must demonstrate some awareness of chemical identity below the macroscopic level—for example, they may refer to "particles" or "atoms" instead of molecules.

Choose-explain (C-E) items. In this type of item, students are presented with a set of options (such as in a MC format) and are then asked to explain their choice. Thus these items are more constrained than CR items but more open-ended than OMC or MTF items. While the options in C-E items may be directed at specific LP levels, as with OMC items, scoring of C-E items is based on both the students' choices and their explanations. The explanation provides further evidence for determining the appropriate LP level. Thus we may have more confidence in assigning a LP level based upon a C-E item than we do with an OMC or MTF item; however, like CR items, C-E items cannot be automatically scored. Nonetheless, the additional information they provide may argue for their inclusion in operational assessments. This item format is also useful for developing OMC items and for refining levels of the of LP framework.

An example of a C-E item is shown in Figure 4. Explanations at both Level 4 and Level 3 demonstrate conservation of mass. Since the LP framework does not address knowledge of the difference in density between liquid water and ice, the scoring guide does not distinguish between the correct choice (E) and the alternate choice (B), which both indicate mass is constant. However, both choices are included because students may hold views consistent with either (E) or (B); to the extent possible, items should allow students to select an answer that matches their thinking about the item content. This additional information may also be useful if students' responses to LP-based items are included as part of a more general "science achievement" scale (see below for more discussion of this issue).

ITEM: Which statement below best describes what happens when ice melts to form water?

a) Mass increases; volume increases.
b) Mass is constant; volume increases.
c) Mass decreases; volume is constant.
d) Mass decreases; volume decreases.
e) Mass is constant; volume decreases.

Use your understanding of the melting process to explain your answer.

SCORING GUIDE:

LP Level	Response
4	B or E; explanation indicates that changes occur to the arrangement of molecules of water (but that they are still the same molecules).
3	B or E; explanation indicates that the number and/or mass of molecules remains constant but that the chemical identity changes from ice molecules to water molecules.
2	A, C, or D; explanation indicates that ice turns into water.
1	A, C, or D; explanation indicates that water has greater/less mass than ice, without indicating a transformation.

Figure 4. Sample C-E item (with associated scoring guide) linked to the learning progression framework in Figure 1.

Item development summary. Accurate assessment of LP levels may require the use of multiple item formats. Large-scale assessments that include CR items may be more conducive to the inclusion of LP-based items than those that are based only on MC items. NAEP includes a range of item formats, with approximately half of the assessment time devoted to CR items. In contrast, some state assessments include only MC items.

Science content is another consideration in the incorporation of LP-based items into large-scale assessments. Items included in large-scale assessments should align with the frameworks that guide assessment development. For example, the NAEP framework specifies the content to be assessed at grades four, eight, and 12. As illustrated in the 2009 framework document (NAGB, 2008), the content of a given LP framework may be present in the assessment framework at all three grade levels, and a set of LP-based items may be administered at multiple grades. Typically, however, the NAEP assessment uses items to determine whether students have mastered the content specified at their grade level. In contrast, the LP-based approach assumes that students at a given grade level may hold ideas at a range of LP levels. Thus LP-based items may cover science content that is more advanced and/or more basic than the content included in the assessment framework. In addition, although a topic may be included in the framework at a particular grade level, a LP-based item assessing that topic may be included in the assessments for multiple grade levels. (As discussed below, this has implications for the way the results of LP-based items are reported.)

Pilot/field testing. Pilot/field testing is used to collect data from a representative sample of the target population. The purpose of pilot/field testing is to evaluate the quality of items before inclusion in an operational assessment. Separate assessment administrations (as for NAEP) or sets of items embedded in operational assessments (as for many state assessments) may be used for pilot/field testing. Due to the preliminary nature of most available LP frameworks as well as the iterative process required to develop these frameworks and associated assessment items, LP-based items would likely require multiple pilot/field testing phases. This testing significantly increases the length of time required for the development process and thus has serious implications for the timing of the pilot/field testing relative to the operational assessment. For example, NAEP and TIMSS currently conduct science assessments approximately once every four years and include a pilot/field test phase prior to each operational assessment. The inclusion of LP-based items would increase the amount of pilot/field testing currently required and thus may mean more time is required from the beginning of item development until the inclusion of items in an operational assessment. This may be less of an issue for state assessments that are conducted every year and that may already include new item pilot/field testing as part of each operational assessment.

In addition, the collection of pilot/field test data must be carefully structured in order to yield information that can be used to refine the LP framework and to evaluate its associated items. An important consideration is the extent to which understandings at a particular level of the LP framework "hang together" across different items and contexts. For example, to investigate Level 3 of the LP framework in Table 1, we would be interested in whether students use conservation of matter as a constraint in multiple contexts involving liquids and solids (e.g., thermal expansion and freezing) and whether students who attribute macroscopic properties to atoms and molecules also have difficulty using conservation of matter as a constraint on changes involving the gaseous phase. These questions cannot be answered by distributing LP-based items across different test booklets. Instead, some students in the pilot/field test sample must answer a set of LP-based items (e.g., items tapping different contexts—such as thermal expansion, freezing, and evaporation—and different aspects of the LP framework—such as the identity of atoms/molecules and the properties of atoms/molecules) so that their performance can be examined across the set of items. This requirement may require significantly different pilot/field test designs than are currently used in large-scale assessment systems. For example, rather than distributing LP-based items across a set of pilot/field test booklets, with only 1–2 LP-based items per booklet, a LP-specific pilot/field test booklet may be required as part of the item development process.

The inclusion of LP-based item sets raises an issue about item-level sample sizes that relates to the nature of item analyses needed. When only item-level descriptive statistics are needed to evaluate the quality of items, pilot/field tests can use much smaller sample sizes than operational assessments. However, the more complex analyses required for the evaluation of LP-based item sets may warrant larger pilot/field test samples. Estimates of required sample sizes depend upon clear statistical criteria for evaluation of LP-based item sets. As discussed below, these criteria are still being developed.

Item analysis and evaluation. Both developers of large-scale assessment systems and LP researchers have developed ways to evaluate the results of field/pilot tests in order to select items for inclusion in operational assessments. The criteria for evaluating items for inclusion in typical large-scale assessments are well-established. However, for the reasons discussed below, these criteria may be inappropriate or inadequate for evaluating LP-based items. At the same time, LP researchers have developed criteria for evaluating the quality of items for assessing students with respect to LP frameworks; however, these criteria may differ from those used to evaluate items for large-scale assessments. Criteria for both traditional large-scale and LP-based items influence the methods—and timeline—for considering the quality of LP-based items and their potential for inclusion in large-scale assessments. A challenge in incorporating LP-based items into large-scale assessment systems is to conceptualize "quality" and the psychometric techniques that ensure items will be of sufficient quality, from both the large-scale and LP perspectives.

In addition, any evaluation of LP-based items necessarily includes concurrent evaluation of the LP framework itself. Field/pilot tests of LP-based items provide evidence about both the items and the LP framework. Anderson (2008) identified three criteria for the validity of LP frameworks: conceptual coherence, compatibility with current research, and empirical validation. The third criterion relies in large part upon the observation and interpretation of student performances: responses to LP-based items (and other assessment techniques such as interviews). In particular, a valid LP framework should "describe actual observed performances by real students" (Anderson, 2008, p. 4), and students should demonstrate consistent performance (e.g., provide responses at the same—or similar—level(s) of a LP framework) across different items or modes of assessment. LP frameworks must be constantly evaluated in terms of these criteria, in light of available evidence (including that from the administration of items as part of large-scale assessment systems).

In typical large-scale assessment development, items at the pilot/field test stage are evaluated individually for their technical properties (e.g., item difficulty and item discrimination). These criteria are based on the requirements of scaling methods (the techniques used to transform the raw score—number of items answered correctly—to a score on the assessment). While LP-based items may be evaluated using the same criteria, they may not function as well as typical large-scale assessment items according to these criteria.

We next present descriptive item statistics that are typically used to select items for inclusion in operational large-scale assessments. We also address how LP-based items might be evaluated according to similar criteria.

Item difficulty. Item difficulty is one of the most important and commonly used statistics in evaluating pilot/field test items for traditional large-scale assessments. One measure of item difficulty is the item mean score, which for dichotomous items is the percentage of students who respond correctly. For polytomous items, item difficulty is the average score expressed in terms of the proportion of maximum points on the item. Thus item mean scores range from 0 to 1. For large-scale assessments, extremely easy or extremely difficult items (item mean scores <0.1 or >0.9) may not be selected for the operational assessment due to scaling requirements. While some items from the lower and upper ends of the scale are included, the majority of items fall in the middle range (item mean scores between 0.3 and 0.7). Thus, even for criterion-referenced assessments, this metric entails a norm-referenced interpretation.

Item mean scores may be calculated for LP-based items; however, the interpretation of these statistics for LP-based items is substantially different from that for traditional polytomous items. Because LP-based items are not designed to reflect "correctness," but rather the level of students' understanding, item mean scores must be interpreted as the mean LP level. There are at least two ways of thinking about this statistic. First, viewing the LP framework as a description of a continuous variable (the sophistication of students' knowledge and practices) in terms of discrete levels, a mean level of 2.6 indicates a level of sophistication between Level 2 and Level 3. Second, viewing the LP levels as representing milestones in students' thinking, each with a characteristic conceptual ecology, a mean level of 2.6 indicates a greater probability of holding ideas consistent with Level 3 as compared to Level 2. For a LP framework with 4 levels, it is inappropriate to interpret the mean level of 2.6 as indicating 65% correct.

While the desired difficulty range of items on a traditional assessment may be established in advance, the expected LP level is less clear. Judgments about this level are likely to be influenced both by empirical data and by matches to content expected at different grade levels (i.e., standards or assessment frameworks). Consideration of mean item scores for large-scale assessments ensures that the test will not be too easy or too difficult for the vast majority of examinees and thus that the assessment allows them to demonstrate their understanding of science. Extending this idea to LP-based items means asking whether an item will elicit evidence of thinking from students at, above, and below the LP framework level expected for a particular grade level. For example, does an item about the chemical identity of substances involved in a phase change (Level 3 and Level 4 of the LP framework in Table 1) make sense to a student who does not yet view conservation of matter as a constraint (Level 1)?

Item discrimination. Item discrimination is a measure of an item's ability to discriminate between higher- and lower-performing students based upon correlations between the item score and an overall measure of performance (i.e., total test score). Item discrimination indices range from -1 to +1, and negative indices are clear indicators of a problematic item. Items with low but positive item discrimination may also be flagged in large-scale assessments; the criteria vary depending on the nature of the item pool and the range of item difficulties. While some items with lower discrimination are included in operational assessments to ensure framework coverage, ideally items will have discrimination indices at or above 0.5. Because of this requirement, MC items selected for inclusion in operational assessments typically include one clearly correct response and a set of clearly incorrect responses. The structure of the LP-based OMC item format may lead to inherently lower item discrimination since options are written to correspond to levels of the LP framework, and distinctions between the ways that students at adjacent levels are expected to think about a phenomenon may be subtle. Thus these items may not correlate well with a total score for the overall assessment since students who perform well on traditional MC items may be attracted to lower level options in OMC items that reflect typical (although not completely correct) student thinking.

Measures of the consistency of students' responses—the extent to which they select responses at about the same LP level across different items—may be more appropriate for evaluating sets of LP-based items. Cronbach's alpha and/or polytomous item response theory (IRT) models may be used to evaluate this consistency; however, as noted by Briggs and Alonzo (chapter 13), these techniques may not account for important features of OMC items. Thus additional psychometric techniques may be required to fully evaluate this aspect of LP-based item quality.

Frequency distributions. In typical large-scale assessment systems, the percentages of students who choose particular responses (for MC items) or who respond at particular score levels (for CR items) are evaluated for any unusual response patterns. Items may be flagged if there are incorrect options for MC items or score levels for CR items with very low percentages of students. Mean overall test scores for students in each response category may also be used to identify MC items with problematic options or CR items for which the mean overall test scores do not follow the expected pattern across increasing item score levels. Because LP-based items aim to provide students with the opportunity to respond at different LP levels and because we do not necessarily expect students at a given grade level to be distributed across all levels, low percentages of students may choose OMC options or respond to a CR item at a given level of the LP framework. This is particularly true if the same item is administered to students at different grade levels. For example, suppose that an OMC item linked to a four-level LP framework is administered to students in grades four and eight. We might expect students in grade four to respond primarily at Levels 1–3, with very low percentages selecting the Level 4 option. We might expect students in grade eight to respond primarily at Levels 2–4, with very low percentages selecting the Level 1 option. Rather than indicating there is a problem with the item, these results may provide an accurate reflection of student thinking at the two grade levels. In order to determine if results such as these indicate a poor item or reflect the distribution of LP levels across the target population, one needs to examine response patterns across a set of LP-based items and across grades for cross-grade items.

To evaluate an individual LP-based item, additional information about the distribution of LP levels in the population of students is required. At the most basic level, as in the example above, we expect the distribution of LP levels in the grade eight population to be skewed towards higher levels of the LP framework than the grade four population. Thus a clear "red flag" is raised if students in grade four respond at higher LP levels than students in grade eight. When a set of LP-based items is administered, the distribution of LP levels (for a given population) based upon a single item can be compared to the distribution across the set as a whole.

Inter-rater reliability (IRR). For items requiring scoring judgments (i.e., CR items), IRR is another important consideration in selecting items for an operational assessment. Large-scale assessments set IRR targets (minimum standards for scorer agreement) based on the number of score levels and scaling requirements. At the pilot/field test stage, items with IRR substantially below target require further revision to the items and/or scoring guides before inclusion in an operational assessment. During scaling, score levels that do not clearly discriminate may be collapsed into a single score level (e.g., partially correct and correct levels may be collapsed to convert a three-level item to a two-level item). While this situation is not ideal, it can be accommodated by the use of scaling models. However, collapsing of score levels is inappropriate for LP-based items since there is no defined LP level that corresponds to the collapsed score category. Therefore, considering how LP-based item scores are analyzed, interpreted and reported, the IRR requirements for LP-based items may be different than those for traditional CR items.

As noted above, LP-based items require a different type of scoring guide than traditional CR items. Traditional short CR items may have two score levels (correct and incorrect) or three score levels (correct, partially correct, and incorrect). Extended CR items may have four or five score levels. LP-based scoring guides, even for short CR items, necessarily have multiple score levels that correspond to those in the LP framework itself (typically four or five score levels). As described above, LP-based scoring guides define the "holistic" LP score levels, describe how these levels apply to the particular item, and illustrate how student responses map to each score level. For LP-based items, scorers do not evaluate the accuracy or completeness of particular components of a student's answer. Instead, scorers identify the LP level that most closely matches the nature of thinking revealed by a student's answer. This scoring procedure requires familiarity not only with the LP framework itself but also with how students are likely to express their thinking at each level.

Extensive scorer training will be required to achieve adequate IRR for LP-based items due to this different approach to scoring. In addition, as more assessment programs move towards computer-based testing, the use of automated scoring may be considered for CR items. However, automated scoring may not be possible for LP-based items. Rather than looking for specific components of a student's answer or the use of particular terminology, scorers of LP-based items need to evaluate the nature of students' thinking. Such evaluations seem beyond the capabilities of most automated scoring systems.

Differential item functioning (DIF). DIF analyses, including Mantel-Haenszel and IRT-based or logistic regression-based methods, are used to examine differential response patterns across groups of respondents. Large-scale assessments may include DIF analyses to evaluate any potential bias against key reporting subgroups (e.g., gender, race/ethnicity, sampling jurisdiction). Assessment development experts and subject matter specialists review all items identified as having DIF in order to decide if language or content unfairly makes the item more difficult for a particular subgroup.

Evaluating DIF for LP-based item sets may require additional considerations beyond those used for the standard subgroups. Since LP frameworks are often developed in the context of particular curricular environments—or in a narrow geographic region, in which curricula may exhibit little variation—it may be especially important to evaluate LP-based item sets for DIF related to students' exposure to different curricula. While particular curriculum materials may influence students' levels of understanding of a concept, students' performances should not be unduly influenced by particular experiences or by the way tasks are framed in particular curricula. An additional challenge is posed by the underlying theoretical assumption that a given LP framework is not the only pathway to full scientific understanding. Instead, multiple pathways are possible, and the pathway that a given student takes may be influenced by his/her prior knowledge and out-of-school experiences. As additional research is conducted to evaluate alternative pathways, DIF analyses must ensure that assessments accurately evaluate the progress that a range of students takes toward achieving scientific understandings.

Item evaluation summary. Thus, while typical item statistics can be calculated for LP-based items, they may not be the most appropriate and certainly are not the only method needed for evaluating these items. Most fundamentally, LP-based items cannot be evaluated using only individual item-level statistics; rather, they must be evaluated as a set. Then the question becomes not one of the performance of a particular item (according to criteria for its item statistics) but of the performance of the set. In particular, the evaluation must focus on the extent to which the set of items captures the thinking represented by the LP framework (and the extent to which the LP framework reflects actual student thinking). Ideally, a student's responses to the set of items would all be at roughly the same level of the LP framework. Deviations from this ideal should be examined to determine whether the LP framework misses or mischaracterizes some aspect of student thinking and/or whether particular items fail to accurately elicit student thinking with respect to the LP framework. Although traditional analyses and criteria may be inadequate or inappropriate for evaluating LP-based items, alternatives have not yet been fully developed. The chapters in the Modeling Learning Progressions section of this book, which focus on modeling students' responses to LP-based items, represent a first step towards identifying appropriate quantitative techniques and criteria for the evaluation of LP-based items.

In addition, the evaluation and modification of LP frameworks and associated items require—to a greater extent than more traditional items—the coordination of both quantitative and qualitative data about students' responses. In-depth analyses of students' written responses to CR and C-E items, as well as cognitive labs or other techniques for eliciting student thinking as they answer items, are needed. Student interviews may also be required to inform the language used to provide directions for LP-based item sets in test booklets. (This issue is discussed in more detail below.) Although cognitive labs and small-scale student try-outs are often conducted as part of large-scale assessment development, these studies are particularly important for the evaluation of LP frameworks and their associated assessment items. Such small-scale studies are essential for establishing the nature of students' thinking (for use in further developing and validating the LP framework) as well as for understanding how well the proposed items tap that thinking. The timeline for an operational assessment must consider that additional cognitive labs, interviews, and small-scale studies may increase the duration of the item development phase.

Design of operational assessments. The design of current large-scale assessments is an important consideration in using LP-based items as part of large-scale assessment systems. As noted previously, large-scale assessment frameworks reflect broad content coverage, and the item pool must adequately cover the framework to ensure content validity of the reporting scales. For example, NAEP science assessments typically include between 150 and 200 items at each grade that are assembled into 25-minute blocks containing items from each of the content subscales; these blocks are paired in multiple test booklets. Each student takes two blocks—about 25–30 items—that reflect only a fraction of the total assessment. NAEP uses IRT scaling and imputation methods to produce population and subgroup estimates based on the subset of item responses from each student in the sample. TIMSS uses a similar approach to assessment design to ensure broad content coverage while minimizing the test burden on individual students. In contrast, state assessment systems may use a single test at each grade level, with the use of anchor items to link the forms administered in different years.

In order to obtain a reliable measure of individual students' LP levels, they must each respond to a set of LP-based items focused on a particular science topic. While LP-based items may be distributed across booklets for an assessment (such as NAEP) designed to make claims about populations (rather than individuals), the number of items required to obtain a reliable measure of the LP level for those populations would certainly exceed the coverage that any particular objective normally receives on the assessment. Thus the decision to include LP-based items in large-scale assessments necessarily represents a commitment to over-representing particular areas of the content framework. Care is required to ensure that the over-representation of LP-specific content does not compromise the content validity of the resulting test scores, which should reflect the broad content coverage specified in the framework for a particular assessment. It is, therefore, important that any LP frameworks measured in the assessment focus on big ideas in science or on key concepts that have a high priority in the framework.

If LP-based items are included in test booklets with more typical items, an additional consideration is the way these items are presented to students. While OMC items may look identical to traditional MC items, they require a slightly different type of response from students. Rather than selecting the "correct" or "best" answer from fairly distinct options, students choose the option—from a set of options that may differ only subtly—that is most consistent with their thinking. Thus the two types of items may require different response processes. This situation raises concerns about presenting students with both traditional MC and OMC items in the same testing environment since the inclusion of OMC items may lead to confusion or unintended variability in how students interpret and respond to both types of MC items. The extent to which this concern is warranted is an empirical question that has yet to be answered. Keeping LP-based items in a different test section and providing a special set of general directions before the section may mitigate concerns about students' confusion when responding to OMC items.

Scoring and reporting. As discussed in the previous section, the design of an operational assessment requires decisions about the characteristics of the entire item pool and about the way the items are presented to students. These decisions have important implications for how results are reported. The inclusion of LP-based item sets introduces additional decisions and questions related to reporting. As mentioned in the previous section, some types of LP-based items may contribute to main reporting scales; however, in order to realize the vision reflected in the LP-based approach, additional LP-specific reports would also be required. This section addresses both typical large-scale reporting practices and modifications to these practices that LP-based items would require.

The types of large-scale assessments that are the focus of this chapter report scale scores that provide information about students' achievement relative to performance benchmarks for a particular grade level. Cut-scores are established to determine the points on the scale that correspond to each benchmark in order to enable reporting for individual students (or groups of students) by performance level. Descriptions and example items are used to illustrate what students know and can do at each level. For example, NAEP reports average scale scores and percentages of students for each of three achievement levels (basic, proficient, and advanced) at the national, state, or urban district level, overall and for key reporting subgroups (e.g., gender, race/ethnicity). However, NAEP does not report scores for individual students. TIMSS reports the percentage of students reaching four benchmarks (advanced, high, intermediate, and low) internationally and for each participating country (Martin, Mullis, & Foy, 2008). State assessments report group-level results by achievement level at the state and school levels and also produce student-level reports that may include percentile rankings and diagnostic scores that link performance levels on particular subscales to recommendations for future learning.

Thus large-scale assessment reports are based upon the categorization of students using achievement levels for their respective frameworks or assessment standards. Although, as described above, the levels in a LP framework have a different meaning than achievement levels in typical large-scale assessment frameworks, the reporting task is fundamentally similar: categorizing students into levels on the basis of their performance on a set of items. In this case, LP-specific scales would need to be created with cut-scores to place students at levels of the LP framework instead of at broader levels of science achievement.

While the overall approach to scaling and reporting is similar for traditional and LP-based scales, LP-specific reporting involves additional considerations. LP-based item sets may span multiple grades in order to report results both within and across grades. However, most, if not all, large-scale science assessments, including NAEP, do not report cross-grade scales. Scaling (and the establishment of cut-scores) for traditional large-scale assessments is typically performed separately at each grade level in order to make claims about how well students have mastered the content expectations for a given grade level. Since LP frameworks are designed to describe the progress that students make towards a particular concept across grade levels, scaling (and the establishment of cut-scores) for LP-based items should be conducted using data from students at different grade levels. This is because the performance indicative of a particular level of the LP framework should not depend on the grade level of the students. The range of grade levels that can reasonably be covered by a single LP framework must be compatible with the grade span of the assessment system (grades four, eight, and 12 for NAEP). Incorporating LP-based item sets may require some form of reporting across grades, although fully cross-grade scales may not be required.

Large-scale assessments such as NAEP typically use IRT models to produce scale scores. IRT methods are well-established for producing highly reliable individual student-level scores for state assessments or population estimates for NAEP. For the reporting of LP-specific scales, alternative measurement and reporting models are needed. The chapters in the Modeling Learning Progressions section of this book explore three possible measurement/reporting models: IRT (Wilson), Bayes Networks (West et al.), and the Attribute Hierarchy Method (Briggs & Alonzo). However, as these chapters explain, a number of challenges and questions remain regarding this work.

In addition, a critical issue—outside the scope of these chapters—is that the reliability of LP-specific scales depends upon the extent to which student thinking—particularly in the "messy middle" (Gotwals & Songer, 2010, p. 277) between naïve and scientific understandings—can be accurately described (by LP frameworks) and assessed (by LP-based items). Shavelson and Kurpius (chapter 2) raise some concerns about the extent to which LP frameworks can produce reliable scores, particularly in the middle of the scale.

If LP-based items can be scaled reliably, reports based upon LP levels may provide valuable information about the knowledge and practices of individual students and/or groups of students in key areas of an overall science assessment. This information has diagnostic capability and—particularly when interpreted in the context of the whole "story" of student learning represented by the LP framework—may be used to improve educational opportunities. However, the hypothetical nature of LP frameworks has important implications for another common use of large-scale state assessments: making high-stakes decisions about students, teachers, and/or schools. Recognizing that LP frameworks are hypotheses that do not necessarily include statements about the understanding that students should demonstrate at particular educational levels, it may inappropriate to use the LP-based scales in such high-stakes decisions.

The above discussion has focused on the use of responses to LP-based items for reporting students' achievement in the context of a LP framework. However, if these items are part of a larger assessment, designed to report overall achievement, one must consider whether and how the LP-based items contribute to the main scale. Two considerations are important. First, as discussed above, LP-based items are potentially scored differently from similar "regular" items. Second, LP-based items may contain content that is not included in the framework for a given grade level. (This is particularly true if items are administered to students at multiple grade levels.) However, scores on some LP-based items may be rescored or "mapped" onto a scoring scheme that is more consistent with the main scale and that necessarily includes variation across grade levels. For example, fourth grade students generally are not expected to provide a correct accounting of physical change at the atomic-molecular level. Thus the MTF item in Figure 2 could be rescored to focus only on the macroscopic properties (mass and volume). However, this rescoring could result in giving full or partial credit to an "incorrect" response if the content of the item exceeds the content expectations for the student's grade level. In addition, some LP-based items may be used for LP-specific scoring but may be unsuitable as contributions to the main reporting scale. For example, consider the OMC item in Figure 1. While fourth graders would not be expected to provide a correct accounting of physical change at the atomic-molecular level, there is no option that reflects the accounting expected of fourth graders.

CR or C-E items can be scored either to measure LP levels or to contribute to the main scale, depending on the nature of the scoring guide. Because LP-based scoring is designed to yield different information about a student's response, the direct mapping of LP-based scores to main scale scores may not be possible. In these cases, the LP-based item would be scored two ways—once with the LP-based scoring guide and once with a more traditional scoring scheme (e.g., as correct, partially correct or incorrect). For example, in the C-E item in Figure 4, a traditional scoring of this item might expect eighth or 12th graders to recognize that the density of ice is less than that of water; therefore, a correct answer differentiates between options B and E. In the LP-based scoring system, these responses are collapsed since scoring is based on the conservation of mass at a molecular-atomic level, without regard to volume. Thus traditional scoring focuses on different aspects of a student's response.

Assessments that measure and report trends in achievement (such as NAEP) pose particular challenges in the reporting of LP-based items. Since LP frameworks reflect data-based hypotheses, it is reasonable to expect that they will change over time; yet reporting of trends requires a stable framework. Therefore, trend reporting of LP levels may not be feasible. However, depending on the assessment design and item release plans, it still may be possible for some individual LP-based items to contribute to the main trend reporting.

Finally, as part of the reporting process, administrators of large-scale assessments typically release some proportion of items on a regular basis in order to show stakeholders what the assessment includes, inform interpretation of test scores, and identify potential areas of the curriculum that require more focus. LP-based items require special consideration in terms of release policies; these issues may be more important for assessments like NAEP and TIMSS than for state accountability systems. First, decisions must be made about whether to release the LP frameworks themselves. As products of both research and the large-scale assessment system, one might argue that the LP frameworks should be made public like standards and assessment frameworks. By releasing the LP frameworks, future research could refine these frameworks (outside of the large-scale assessment system). In addition, LP-based items are not very meaningful without the associated LP framework; therefore, release of LP-based items must be accompanied by release of the LP framework in order to fully inform interpretation of student performance on the released items. However, this procedure has implications if the same LP framework is used to guide future test administrations. If the assessment is designed to measure trends (as NAEP and TIMSS do), then multiple test administrations may need to include the same LP framework. Release of any information about the LP-based items may signal to stakeholders an area of the content framework that may be over-sampled in future assessments. While teaching to a particular LP framework may be desirable in terms of focusing attention on important content, this practice may occur at the expense of other content areas.

RECOMMENDATIONS

The incorporation of LP-based items into large-scale assessment systems requires close collaboration between researchers and assessment developers. In contrast with typical practice—in which assessment developers may draw upon research as one of many influences on the assessment framework and item development process—researchers and assessment developers must work together in an iterative fashion to refine the LP framework and its associated items. While large-scale assessment should incorporate only well-researched LP frameworks, as hypotheses, the LP frameworks may require revision in light of additional evidence. Researchers cannot "hand off" a validated (and static) LP framework to assessment developers, particularly since the administration of large-scale assessments (both pilot/field test and operational versions) provides opportunities to collect data that may be otherwise unavailable to researchers. Such collaboration may require reconsideration of policies and capacities related to the availability of information about large-scale assessment systems. Researchers need access to items and data, as well as the ability to share their findings about LP frameworks—and their instantiation in assessment items—with others in the research community. For assessments such as NAEP and TIMSS, test security is a particular concern. While the design of state assessments may permit greater release of information, the infrastructure required to provide data in forms useful to researchers may be more limited. State assessment systems are organized to report test scores for individual students and schools; personnel and data management systems may not be in place to provide the item-level data needed by researchers. In the following recommendations we assume that policies permit envisioned partnerships in which researchers and assessment developers play symbiotic although separate roles. In recognition of these different roles, we offer separate recommendations for researchers and assessment developers.

Researchers

In order for LP frameworks to influence large-scale assessments, much more research must be conducted to develop well-grounded LP frameworks. For instance, the examples of research-based LP frameworks described in the NAEP framework (NAGB, 2008) reflect only a small fraction of the total content covered in the assessment. However, as research moves forward, not all studies have equal potential to inform large-scale assessment development. Clearly, work on LP frameworks that address the big ideas in science are more likely to be consistent with large-scale assessment frameworks. The consensus in the science education community on the nature of LPs already reflects the idea that LP frameworks should focus on key areas of the science curriculum (e.g., Corcoran et al., 2009). However, agreement about those key areas may be lacking. This section outlines additional considerations for research that may inform large-scale assessment development. Other types of research have value as well, but large-scale assessment systems require specialized studies from the research community.

While much of the research on LP frameworks has a curriculum focus, it is important to recognize that large-scale assessments cannot reflect the development expected as a result of a particular curriculum. Thus LP frameworks that describe typical pathways students take as they progress from naïve to scientific thinking about a phenomenon are more useful to the large-scale assessment community. Over time, curriculum-based research may influence the way particular concepts are taught to students, and LP frameworks would need to be revised to reflect changes in typical learning patterns. However, for large-scale assessments, which represent a wide variety of students and educational experiences, the goal should always be to capture how typical students actually learn rather than how they might ideally learn from a particular curriculum. Research at grade levels assessed in large-scale assessments is particularly important.

Similarly, research is often conducted with convenience samples—groups of students whose teachers are willing to work with researchers. However, large-scale assessments must represent all students in a given geographical area. While researchers may not have access to the same population samples available to large-scale assessment developers, they should pay attention to the characteristics of students included in their research in order to develop LP frameworks that minimize the revisions necessary once data from a representative sample are available. Similarly, while cross-sectional studies are often more convenient, longitudinal explorations of the ways individual students' understandings develop over time are particularly relevant for the development and validation of LP frameworks across grades. Without longitudinal data, the hypotheses represented by LP frameworks—that students progress through the specified set of increasingly sophisticated understandings—cannot be fully tested.

In addition, research is needed that explores the technical aspects of the inclusion of LP frameworks in large-scale assessments. First, research that explores and evaluates alternative measurement models may identify LP-specific techniques for modeling students' responses to these sets of items. Measurement models that are currently used for scaling and imputing scores on large-scale assessments may be inappropriate for LP-based items. Alternative measurement models are currently being considered (see the chapters in the Modeling Learning Progressions section of this book for further discussion). Second, research to inform the design of LP-based assessments is needed. In particular, questions remain about the number of LP-based items that can be expected to produce a reliable LP-specific score and about the way these items should be distributed across students for different types of reporting. Research may also help answer questions about the effect of administering items that fall outside the content expectations for students' grade level or about the effect of including both OMC and traditional MC items in a single test administration. Finally, researchers may provide recommendations on the most effective and accurate reporting methods for providing meaningful interpretations of results from LP-based items for the broad range of large-scale assessment stakeholders.

Assessment Developers

Because the purpose of LP-based items differs from that of more traditional large-scale items, item-writers and scorers must understand the purpose of these items (to elicit responses that can be used to categorize students using the levels of the LP framework) and recognize the important features of these items (such as eliciting evidence of the nature of students' thinking at a range of different LP levels). Therefore, it is important that LP experts are involved in the item review and revision process. Scoring guides must be developed (and applied) that reflect the LP purpose. Rather than evaluating a student's response based upon the number of correct components or upon the extent to which it approximates the correct answer, LP-based items should be scored to reflect the LP level that most closely resembles the student's thinking. Since this procedure represents an important departure from typical scoring, scorers must be trained to avoid the assumption that responses naming a greater number of scientific concepts should be scored at a higher level. Additional time and training may be required for scorers to achieve acceptable levels of IRR.

Because LP frameworks represent hypotheses about the way student thinking develops over time, evaluation of LP-based items necessarily includes evaluation of the LP framework itself. Additional cognitive labs and student try-outs, in multiple cycles of revision and pilot/field testing, may be required for LP-based items prior to their inclusion in formal pilot/field testing. The typically short period of time allotted for this stage of the item development process may need to be lengthened in order to permit sufficient evaluation and testing of both the LP framework and its associated items. Once items are ready to be incorporated into pilot/field tests, careful consideration must be given to the distribution of items across test booklets. Since the pilot/field test stage may be the first opportunity to collect data from a representative sample, it is particularly important that individual students respond to a set of LP-based items; therefore these items cannot be distributed across multiple test booklets. The same is true if individual scores are to be reported in terms of the LP framework, since a set of items is needed to diagnose a student's LP level. The number of items that a student should be required to answer (for different reporting situations) is a question that can be explored in collaboration with researchers.

Traditional measures of item quality may not be appropriate for LP-based items. For the reasons described above, these items may not perform well when statistics such as frequency distribution and item discrimination are considered. In addition, statistics such as frequency distribution and item mean score must be evaluated in the context of a set of items (and the extent to which these items tell a consistent story about student performance) rather than in isolation. Therefore, these items must be evaluated according to rigorous standards in terms of their ability to capture the thinking described at each LP level. The LP framework itself must be validated in terms of its ability to describe levels of student thinking. Evidence of this thinking necessarily comes from both quantitative (i.e., item analysis) and qualitative (i.e., cognitive labs) data analyses.

Finally, assessment developers must consider the measurement model and analyses that will be used to report scores on LP-based items, particularly the way these items will interact with the main scale of a given assessment. Current IRT models may not be capable of fully capturing students' performance on LP-based items; therefore, alternatives should be explored in collaboration with researchers. In addition, careful consideration must be paid to how LP-based items are incorporated into any overall reporting.

CONCLUSIONS: LOOKING AHEAD

The NRC report *Systems for State Science Assessment* (2006) recommends explicit inclusion of cognitive models in the design of large-scale assessments and interpretation of their results. LP frameworks offer a promising new vision for achieving this goal. In this chapter, we have explored the challenges and questions raised by the modest goal of incorporating LP-based items into existing large-scale assessment systems. We view work towards this goal—which requires close collaboration among science education researchers, psychometricians, and assessment developers—as a critical foundation for a new vision of science assessment.

Looking forward, we envision significant advances in the use of LP frameworks to inform large-scale assessment systems. Researchers are working to develop LP frameworks across a wider range of the K-12 science curriculum. In addition, there are ongoing conversations about standards (e.g., Common Core Standards Initiative[9] and the NRC's effort to develop a conceptual framework for new science education standards;[10] see also Foster & Wiser, this volume). Both strands of work are likely to result in more targeted expectations for student learning, thus increasing the possibility that LPs cover the objectives of a particular science assessment. Work on psychometric techniques for modeling students' responses to LP-based items is currently underway (see the Modeling Learning Progressions section of this book) and should lead to more LP-specific criteria for evaluating items. Finally, the incorporation of computer adaptive testing (CAT) into the "machinery" of existing large-scale assessment systems holds particular promise for the administration of LP-based items. CAT allows students' responses to determine which items are presented later in the assessment. For example, linked sets of items could be developed in which a student's thinking about a specific phenomenon is investigated in an initial item and then explored in subsequent items (which are tailored to the response the student provided to the first item). Or a small set of items targeting the full range of LP levels could be followed by items that target a subset of LP levels indicated by a student's performance on the initial set of items. CAT has the potential to address several challenges related to LP-based assessments. First, as mentioned above, it can be difficult to write items that elicit responses at a range of LP levels. A CAT assessment requires fewer such items in order to provide an estimate of students' levels for determining the subsequent items to be administered. Second, this approach reduces the burden on a single item to clearly differentiate between levels. As mentioned above—and illustrated with the item in Figure 2—a particular item context may not allow for a clear distinction between responses at adjacent levels. However, such items may be used to provide an initial estimate of students' levels. Finally, CAT may reduce the number of LP-based items administered to students.

Eventually, we envision some assessment systems in which all items are linked to a small number of LP frameworks and results are reported in terms of students' knowledge/practices with respect to those frameworks. However, this is possible only when standards have been narrowed to focus on a smaller number of big ideas, when additional validated LP frameworks are available, when psychometric techniques have matured to the point that clear evaluation of LP-based items is possible, and when CAT systems are fully in place.

NOTES

1 These caveats are reflected in the discussion below and are described in the NAEP framework as: 1) "learning progressions are not developmentally inevitable"; 2) "there is no single 'correct order'; there may be multiple pathways by which certain understandings can be reached"; 3) "actual learning is more like ecological succession with changes taking place simultaneously in multiple interconnected ways"; and 4) "the learning progressions suggested in the framework and specifications are partly hypothetical or inferential because long-term accounts of learning by individual students do not exist"(NAGB, 2008, p. 90).

2 Trends in International Mathematics and Science Study (http://nces.ed.gov/timss/)

3 For example, the 2009 NAEP science assessment includes computer-based items (NAGB, 2007, 2008). Many US states have implemented or are exploring the use of computer delivery systems for administering accountability tests (Thurlow, Lazarus, Albus, & Hodgson, 2010). In addition, computer adaptive testing (CAT) capabilities are poised for inclusion in state accountability systems; the SMARTER Balanced Assessment Consortium (2010) will use CAT technology as part of a multistate effort to develop an assessment system based upon the new Common Core State Standards in mathematics and language arts. As discussed below, CAT represents a promising vehicle for administering LP-based items; however, the use of CAT is not currently a widespread practice in most large-scale assessments.

4 Toulmin (1972) used the term "intellectual ecology," drawing parallels with accounts of organic evolution. Influenced by Toulmin's work, Posner et al. (1982) coined the term "conceptual ecology," which is used here.

5 Information about NAEP practices is based upon *The Nation's Report Card: An Overview of Procedures for the NAEP Assessment* (2009) and the second author's work on the NAEP assessment system.

6 For example, in the 2009 NAEP science framework (NAGB, 2008), one can trace increasingly sophisticated expectations for students' understanding of matter from grade four to grade eight to grade 12. In grade 4, students should understand properties of the different states of matter and that

heating and cooling change matter from one state to another. In grade 8, students should have developed a particulate model of matter and should be able to use this model to explain both properties of different states of matter and changes between states of matter. In grade 12, students should understand that the arrangement of and forces of attraction between atoms, ions, and molecules explain properties of the different states of matter. They should also understand that changes of state require a transfer of energy.

7 Neither the LP framework nor the associated items should be used for research or other applications without significant revision as part of a comprehensive validation effort.

8 See Briggs et al. (2006) for further discussion of the relationship between OMC and CR items.

9 See www.corestandards.org; currently, there are Common Core State Standards in English language arts and mathematics.

10 See http://www7.nationalacademies.org/bose/Standards_Framework_Homepage.html

REFERENCES

Alonzo, A. C. (2009, April). Design criteria for learning progressions to support teachers‘ formative assessment practices. In A. C. Alonzo (Chair), *Learning progressions in science: Tools for teaching, learning, and assessment*. Symposium conducted at the annual meeting of the American Educational Research Association, San Diego, CA.

Anderson, C. W. (2008, February). *Conceptual and empirical validation of learning progressions Response to "Learning progressions: Supporting instruction and formative assessment."* Retrieved from Consortium for Policy Research in Education website: http://www.cpre.org/ccii/images/stories/ccii_pdfs/learning%20progressions%20anderson.pdf

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Grigg, W., Lauko, M., & Brockway, D. (2006). *The Nation's Report Card: Science 2005* (NCES 2006–466). Retrieved from National Center for Education Statistics website: http://nces.ed.gov/nationsreportcard/pdf/main2005/2006466.pdf

Martin, M. O., Mullis, I. V. S., & Foy, P. (with Olson, J. F., Erberber, E., Preuschoff, C., & Galia, J.) (2008). *TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hills, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from TIMSS website: http://timss.bc.edu/TIMSS2007/sciencereport.html

(2009). *The Nation's Report Card: An overview of procedures for the NAEP assessment* (NCES 2009–493). U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. Washington, DC: U.S. Government Printing Office. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/about/2009493.pdf

National Assessment Governing Board (2004, September). *Science framework for the 2005 National Assessment of Educational Progress.* Retrieved from http://www.nagb.org/publications/frameworks/s_framework_05/761907-ScienceFramework.pdf

National Assessment Governing Board.(2007). *Science assessment and item specifications for the 2009 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/publications/frameworks/naep2009-science-specs07.doc

National Assessment Governing Board.(2008, September). *Science framework for the 2009 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/publications/frameworks/science-09.pdf

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107–110.pdf

Posner, J., Strike, K., Hewson, P., & Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*, 211–227.

SMARTER Balanced Assessment Consortium. (2010). *A summary of core concepts*. Retrieved from http://www.k12.wa.us/smarter/pubdocs/SBACSummary2010.pdf

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations* (Synthesis Report 78). Retrieved from University of Minnesota, National Center on Educational Outcomes website: http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis78/Synthesis78.pdf

Toulmin, S. (1972). *Human understanding.* Princeton, NJ: Princeton University Press.

ALICIA C. ALONZO

ELICITING STUDENT RESPONSES RELATIVE TO A LEARNING PROGRESSION

Assessment Challenges

The assessing strand is critical to work on learning progressions. Obtaining evidence to support or revise a proposed learning progression requires assessments (methods to elicit student responses relative to the learning progression) in order to test hypotheses about student thinking and its evolution over time. In addition, many proposed applications of learning progressions involve assessments—either directly or indirectly. The most recent science framework for the National Assessment of Educational Progress (NAEP) calls for the inclusion of learning progressions in this influential national test (National Assessment Governing Board [NAGB], 2008). Learning progressions are promoted as offering support for teachers' formative assessment practices (e.g., Alonzo, 2009, 2011; Furtak, Thompson, Braaten, & Windschitl, this volume); links between formative and summative assessment at multiple levels of the educational system (Black, Wilson, & Yao, 2011); and coherence among curriculum, instruction, and assessment (National Research Council [NRC], 2006). The use of learning progressions to develop curricula requires assessments to evaluate their effectiveness (e.g., Gotwals, Songer, & Bullard, this volume; Schwarz, Reiser, Acher, Kenyon, & Fortus, this volume; Wiser, Smith, & Doubler, this volume). The use of learning progressions to develop standards (e.g., Foster & Wiser, this volume; NRC, 2011) will eventually lead to the design of assessments that evaluate student achievement relative to those standards.[1] Indeed, Corcoran, Mosher, and Rogat (2009) proposed that learning progressions are incomplete without associated assessment tools. While this chapter does not take Corcoran et al.'s position, it recognizes that their proposal points to the centrality of assessment to work on learning progressions.

Efforts to design learning progression assessments build on a rich foundation. As reflected in the chapters in this book, a 2001 NRC report made significant contributions to our understanding of the criteria for "knowing what students know." This report "considered implications of advances in the cognitive and measurement sciences for both classroom and large-scale assessment" (p. 18). First, the report stated that assessment design should be based on models of student cognition. The report's emphasis on the nature of expertise and its development in various domains is of particular relevance for work on learning progressions. Second, the report highlighted recent work on formal measurement models as "a particular form of reasoning from evidence" and called for "greater attention to the

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 241–254.

interplay between the statistical and cognitive aspects of assessment than has been customary" (p. 110). A key example identified in the report—the Berkeley Evaluation and Assessment Research (BEAR) assessment system—has greatly influenced the work on learning progressions. As described by Wilson (2005), the BEAR assessment system (BAS) provides a comprehensive means of designing, evaluating, and using assessments organized around a *construct map* that reflects an underlying continuum. Mislevy's work on evidence-centered-design (ECD; Mislevy, Steinberg, & Almond, 2003) provides another illustration of the NRC recommendations. ECD highlights the claims a test developer wishes to make about students and the evidence required to make those claims. ECD is also reflected in work on learning progressions and associated assessments (e.g., Gotwals & Songer, 2010; Stevens, Delgado, & Krajcik, 2010; West et al., this volume).

While we may draw important lessons from this foundational work, to date there has been little consideration of the assessment challenges specific to learning progressions. In its overview of these challenges, this chapter draws upon discussions at the Learning Progressions in Science (LeaPS) conference and upon other chapters in this book. The chapters in the assessing strand are most influential for this discussion. However, because the assessing strand is so tightly integrated with other strands of work on learning progressions, almost every chapter in this book mentions assessment. Therefore, this synthesis chapter also references chapters in other strands.

THEME 1: ASSESSMENT IS CRITICALLY DEPENDENT ON THE OTHER THREE STRANDS OF LEARNING PROGRESSION WORK

At the LeaPS conference, participants often qualified their responses to questions about assessment with the remark, "It depends." Such qualification was necessary because the assessing strand relies heavily on other aspects of learning progression work. One cannot consider learning progression assessments in isolation. The NRC (2001) assessment triangle illustrates this point by representing connections among cognition, observation, and interpretation. As several chapters in this book (Gotwals et al., this volume; Jin & Anderson, this volume; West et al., this volume) highlight, the assessment triangle provides a model for work on learning progressions and associated assessments. The iterative process of defining/refining a learning progression, designing associated assessment tasks, and interpreting student responses to assessment tasks in terms of the learning progression has connections to both the assessment triangle and to the four strands of work on learning progressions—defining, assessing, modeling, and using. The relationship between the assessment triangle and the four strands of work on learning progressions (depicted in Figure 1) is explored next.

The cognition vertex of the assessment triangle corresponds to the learning progression and, thus, to the defining strand. Decisions made in this strand are critical to the design of learning progression assessments. Learning progressions reflect assumptions about learning and choices about how to conceptualize progress. For example, decisions to highlight content (e.g., Plummer, this volume) or practices (e.g., Schwarz et al., this volume) or a combination of the two (e.g., Gotwals et al., this volume) have important implications for the types of assessments developed. Decisions about breath and grain size are also made in the defining strand. A narrow learning progression that focuses on specific conceptions (e.g., the sinking and floating learning progression described in the Shavelson and Kurpius chapter, this volume) requires assessment items that elicit student thinking about a particular phenomenon. In contrast, a learning progression that focuses on broad characteristics of student thinking (e.g., Gunckel, Mohan, Covitt, & Anderson, this volume; Jin & Anderson, this volume) requires assessment items about different phenomena in order to characterize student thinking across contexts. To assess accounts of carbon-transforming processes, Jin and Anderson present students with everyday phenomena representing three processes: organic carbon generation (i.e., photosynthesis), organic carbon transformation (i.e., digestion, biosynthesis), and organic carbon oxidation (i.e., combustion, cellular respiration). Accounts of only one process will not result in a full characterization of the range of thinking in the learning progression.

Because many learning progressions have approximately the same number of levels (4–6),[2] broader learning progressions usually have a larger grain size than narrower learning progressions. Thus, for broader learning progressions, greater distances between levels (in terms of increase in knowledge and/or practices) are more likely than for narrower learning progressions, and, overall, movement from the lowest to the highest level represents a greater accomplishment. This is true when comparing the two examples noted above. The floating and sinking learning

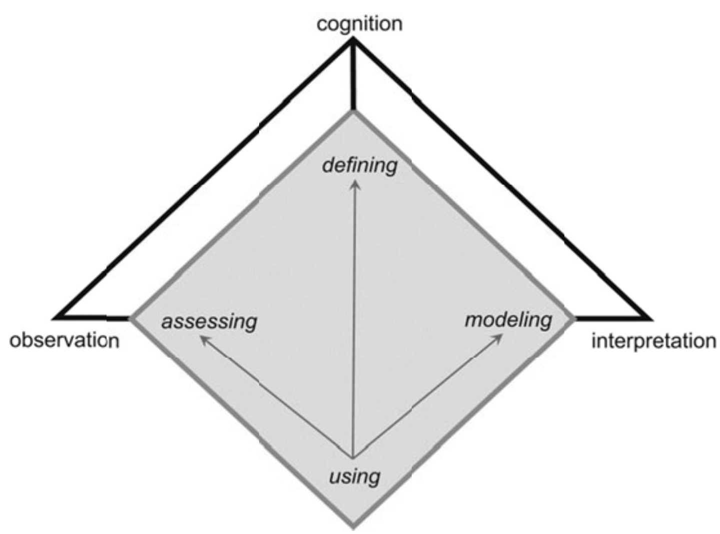


Figure 1. Relationship between the NRC (2001) assessment triangle (white triangle) and the four strands of learning progression work (gray square).

progression (Shavelson et al., 2008) spans a single instructional unit, in which student progress from one level to the next is expected after a few investigations. In contrast, the environmental literacy learning progressions (e.g., Gunckel et al., this volume; Jin & Anderson, this volume) span many years of instruction (upper elementary through high school); each level represents progress that may take years to achieve. As discussed in more detail below, broader learning progressions introduce additional challenges in terms of writing assessment items that are accessible to students at multiple levels.

The observation vertex, which corresponds to the assessing strand, lies between the cognition and interpretation vertices. The latter vertex corresponds to the modeling strand because interpreting student responses to assessment items—for example, to diagnose a student's learning progression level—often requires an appropriate measurement model. An important idea from both the LeaPS conference and the chapters in this book (e.g., Briggs, this volume; Gotwals et al., this volume; West et al., this volume) is that analysis of misfit is critical for informing revisions to learning progressions, assessment items, and measurement models. Misfit indicates a problem with the match between student thinking and the representation and interpretation of that thinking through the use of learning progressions and associated tools (assessment items and measurement models). Since an important goal of learning progression research is to capture student thinking and its development accurately, misfit analyses provide critical information for the iterative process of revising the learning progression, assessment items, and measurement model.

Misfit analyses require clear expectations for student performance and techniques to evaluate how well student performance matches expectations. Both the learning progression and the specific design of assessment tasks inform those expectations. (As discussed below, articulating expectations requires an understanding of what makes tasks more/less difficult for students.) While qualitative approaches are certainly essential for developing detailed understandings of student thinking, as learning progressions move beyond small-scale research studies, quantitative analyses become increasingly important. To ensure that learning progressions and associated tools are generalizable to contexts beyond those in which they were developed, large-scale studies are required. Quantitative analyses allow researchers to summarize large amounts of data for which qualitative analyses are impractical. In addition, quantitative analyses are critical as learning progressions begin to impact state and national assessments—through assessment frameworks (e.g., NAGB, 2008), new national standards (NRC, 2011), and state standards (Foster & Wiser, this volume). Quantitative analyses of misfit provide important evidence for evaluating the quality of items included in large-scale assessments. This evidence is particularly important for high-stakes assessments where the quality of items is of critical significance.

Finally, considerations of purpose (the using strand) underlie the assessment triangle. The use of assessment results is central to the concept of validity (e.g., Messick, 1989, 1995); thus it is impossible to evaluate the quality of a particular assessment without knowing how it will be used. It is clear that different uses may require different types of assessments. For some uses (e.g., large-scale assessments, summative evaluations of curricula) we may be interested solely in determining the learning progression level for a student (or a group of students). When we report assessment results by learning progression level, those results should be as reliable and valid as possible. For other uses (e.g., informing teachers' instructional decisions, evaluating a particular learning progression), we may need a much more detailed picture of student understanding. For these uses, assessment results should provide a more detailed accounting of student thinking about particular concepts and the patterns of their responses to particular assessment items.

The level of detail that can be reported depends on decisions about grain size made in the defining strand. These decisions are described differently in the defining synthesis (Mohan & Plummer, this volume) and in the modeling synthesis (Briggs, this volume). In the defining strand, grain size refers to the distance between the learning progression levels—the magnitude of the accomplishment required to progress from one level to the next. When assessments are used to diagnose student learning progression levels, we obtain much more detailed information for fine-grained learning progressions than for coarse-grained learning progressions. Diagnosing student levels using a fine-grained learning progression yields specific descriptions. For example, students at Level 2 of the sinking and floating learning progression "apply either mass or volume to [explain] sinking and floating" (Shavelson et al., 2008, p. 29). In contrast, diagnosing student levels using a coarse-grained learning progression yields much broader descriptions. For example, the NRC-commissioned learning progression on matter and atomic-molecular theory (Smith, Wiser, Anderson, & Krajcik, 2006) characterizes student reasoning within one of three grade bands (K–2, 3–5, or 6–8). A "big idea" for middle school students in the matter learning progression is that "[t]he properties of materials are determined by the nature, arrangement and motion of the molecules they are made of" (p. 14). Because this statement may apply to many properties of materials, the matter learning progression provides less specific information about student understanding of density than the sinking and floating learning progression does. However, as explored by Mohan and Plummer (chapter 7), if levels of fine-grained learning progressions cannot be reliably differentiated by assessments, a coarser-grained learning progression may be necessary.

As noted above, assessments are often designed to provide more information than simply student location on the learning progression. Whether more detailed information can be provided depends upon grain size as defined in the modeling synthesis (Briggs, this volume), where the focus is on the detail of the learning progression levels—how small the knowledge and practices are that comprise the levels. Here, too, decisions made in the defining strand are critical. For example, consider the Smith et al. (2006) learning progression. Although each level spans three years of schooling, these authors give a detailed accounting of how students at each level are expected to reason about different contexts (through *learning performances*). It is this detailed accounting that permits the authors to propose items that elicit student understanding of specific contexts at each level. Teachers and learning progression researchers can analyze responses to these items qualitatively in order to make instructional decisions and revise the learning progression. With an appropriate measurement model, such items can also provide more detailed information about student knowledge and practices in large-scale assessments. However, the assessment context may place constraints upon the type and number of items administered and, therefore, the level of detail that can be reliably obtained.

Thus it is not possible to define a learning progression (and perhaps associated curriculum) and only later think about assessment. Learning progression work, from its very beginning, requires substantive collaboration among those with varied expertise—cognitive scientists, scientists, science educators, and assessment and measurement experts.

THEME 2: LEARNING PROGRESSION ASSESSMENTS SHARE IMPORTANT FEATURES

Learning progression researchers work in many different contexts and have many different goals. Therefore, because of the integral relationships among a particular learning progression (as a model of student cognition), its proposed use, and its associated assessments, it is reasonable to expect variation in learning progression assessments. Yet, as revealed by comments at the LeaPS conference and by chapters in this book, there is some consensus about "ideal assessment systems" for learning progressions. Some features of these systems are not unique to the learning progression context, although they have particular relevance for learning progression assessments. Other features, however, are specific to the learning progression context.

General Assessment Features

Learning progressions have been promoted as a response to two calls for reform in science education. Learning progression assessments may play an important role in heeding these calls. The first call is for greater focus on a smaller number of core ideas. This call is a response to current curricula that have been described as "a mile wide and an inch deep" (Schmidt, McKnight, & Raizen, 1997, p. 122). The 2007 NRC report *Taking Science to School* recognized learning progressions as tools that can guide curriculum development and suggest "priorities in the curriculum" so that students learn "age-appropriate versions of core ideas with understanding" (p. 247). Assessments—particularly those with high-stakes—exert critical influence on the K-12 curriculum. Thus, as part of the large-scale assessment landscape, learning progression assessments that focus on core ideas may lead to more focused curricula. The second call is for greater coherence among parts of the educational system. This means that (a) curriculum, instruction, and assessment (e.g., NRC, 2001; Webb, 1997) and (b) classroom formative and summative assessments and large-scale accountability tests (e.g., NRC, 2001; Wilson, 2004) together support student learning. The 2006 NRC report *Systems for State Science Assessment* and Black et al's (2011) recent article advocate learning progressions as a way to achieve this coherence. Thus learning progression assessments may lead to a narrower focus (on core ideas) and increased coherence (across levels of the educational system).

General assessment guidelines recommend using different types of assessment items (e.g., Educational Testing Service, 2009; NRC, 2001). This recommendation may be particularly important for learning progression assessments, which (as discussed below) have a significantly more complex task (the characterization of student thinking) than do traditional assessments (the evaluation of the correctness of student responses). Although cognitive interviews are not practical in large-scale settings, they may be crucial for developing written assessments and for refining learning progressions that capture student thinking as accurately as possible. Several projects described in this book (Gotwals et al., this volume; Jin & Anderson, this volume; Schwarz et al., this volume; Wiser et al., this volume) use cognitive interviews for these purposes. In addition, different item types have been proposed for learning progression assessments. Scoring of these item types permits a fuller characterization of student thinking than a simple evaluation of correctness does. While constructed-response and multiple-choice items are commonly used in all types of assessments, more novel items—such as scaffolded items (e.g., Gotwals et al., this volume), ordered-multiple choice items (Briggs & Alonzo, this volume; Briggs, Alonzo, Schwab, & Wilson, 2006), multiple true-false items (e.g., Alonzo, Neidorf, & Anderson, this volume), and choose-explain items (e.g., Jin & Anderson, this volume)—may be particularly important in learning progression assessments. Examination of responses across item types may deepen our understanding of student thinking and of the affordances and constraints of non-traditional item types in eliciting and evaluating student thinking relative to learning progressions (e.g., Briggs et al., 2006; Alonzo & Steedle, 2009).

Learning-Progression-Specific Assessment Features

As Alonzo et al. (chapter 10) discuss, learning progression assessments have a fundamentally different purpose than more traditional assessments. Learning progression assessments are more diagnostically-oriented. Even in large-scale or summative contexts, in which the results may be reported as estimates of the learning progression level for a student (or for a group of students), the information is more diagnostic because the levels of the learning progression provide descriptions of student thinking. In this way, learning progression assessments reflect a *learning progression stance*, in which student learning is assessed more deeply than on the basis of whether they "get" a particular idea or can perform a particular practice. Rather, the learning progression stance assumes that student knowledge and practices in the "messy middle" (Gotwals & Songer, 2010, p. 277) and their "wrong" answers are essential for characterizing student thinking. The implication of this stance is that learning progression assessments must elicit knowledge and practices at various levels and must be sensitive to student progress towards a goal (rather than just assessing achievement of that goal).

As discussed by Jin and Anderson (chapter 8) and Gotwals et al. (chapter 9), the learning progression stance presents a significant challenge in the design of assessments. In particular, learning progression assessments must provide opportunities for students with different levels of knowledge and practice to demonstrate what they know and can do. Jin and Anderson explored a variety of assessment approaches and their ability to elicit accounts of phenomena from students at different levels of their learning progression. Gotwals et al. used carefully structured scaffolding to elicit students' scientific explanations with and without support. For both projects, the fundamental question is whether the assessments reveal student thinking accurately at different learning progression levels. More sophisticated students may not exhibit higher levels of understanding if the questions do not prompt a sophisticated response. For example, students at higher levels of Jin and Anderson's learning progression did not always recognize that questions about everyday events required accounts at the atomic-molecular scale. Students in the Gotwals et al. study may have thought it unnecessary to offer the details researchers require for evaluating their ability to construct scientific explanations. Scientific language may be required to cue older students to provide responses with the needed detail; however, this language may be confusing to younger students. In addition, the limited language skills of younger students may prevent them from expressing their understandings in writing.

As Alonzo et al. (chapter 10) briefly mentioned, and participants at the LeaPS conference discussed more extensively, computer-adaptive testing (CAT)[3] may be a promising approach to learning progression assessment. CAT could use students' responses to earlier items to estimate their learning progression levels and, thus, to present later items targeted at those levels. The later items could be used to obtain more refined estimates of student knowledge and practices. Because the later items mimic a teacher's ability to use follow-up questioning on the basis of initial responses by students, CAT may provide some approximation of the flexibility of classroom assessments in large-scale testing environments.

In addition to capturing student thinking at different levels, learning progression assessments should reflect the dynamic nature of student thinking. Although much learning progression work has been conducted using cross-sectional research, longitudinal studies are needed to determine if students progress through learning progressions in the ways hypothesized. Therefore, learning progression assessments should capture the change in individual students' thinking over time. This requirement relates to the grain size of the learning progression and the time period over which progress is monitored. For example, if a learning progression has broad levels, with significant shifts in understanding expected over years of instruction, assessments linked to the learning progression may not detect changes in student knowledge and practices during a single instructional unit. As discussed in the modeling strand (e.g., Briggs, this volume), the grain size issue also relates to the way learning progressions are conceptualized and modeled. Do we lose information about student progress by conceptualizing learning progressions in terms of discrete levels instead of as continuous pathways? If we focus only on big shifts in student knowledge and practices, do we miss the progress needed within a given level? Assessments that detect changes as student progress within a given level may provide the necessary information for making classroom-level decisions (for both curriculum development and teachers' formative assessment practices). For example, Schwarz et al. (chapter 6) discovered that scoring rubrics based upon the broadly-defined levels of their original learning progression did not capture important changes in student modeling practices. Therefore, they defined more detailed features of student modeling practices in order to capture these features in their assessments. Thus the requirement that learning progression assessments should capture change over time may lead to revisions in how learning progressions are defined and represented in measurement models.

THEME 3: THERE ARE A NUMBER OF IMPORTANT CHALLENGES IN LEARNING PROGRESSION ASSESSMENTS

Because learning progression assessments are integrally related to how the associated learning progressions are defined, the assessing strand is impacted by questions and issues from the defining strand, as discussed in this section.

Unresolved Questions about the Nature of Student Thinking

Much of the work on learning progressions in science makes a strong assumption linked to ongoing debates about the nature of student thinking. Diagnosing student learning progression levels using responses to assessment items requires that students reason consistently at a particular level.[4] This requirement is consistent with the claim that children's thinking is internally consistent and theory-like (e.g., Ionnides & Vosniadou, 2001; Vosniadou & Brewer, 1992). However, others claim that a knowledge-in-pieces model more accurately describes student thinking (e.g., diSessa, 1993; diSessa, Gillespie, & Esterly, 2004). If this second claim is true, we would not expect students to reason consistently across problem contexts. Indeed, student thinking in the messy middle may be rather fragmented and context-dependent, and therefore difficult to describe. Shavelson and Kurpius (chapter 2) caution that there is a danger in forcing students to fit a learning progression model. If student thinking is fragmented and context-dependent, attempts to characterize it using an ordered set of coherent models may result in mischaracterization. If we expect students to respond consistently to a set of learning progression assessment items, we may interpret inconsistencies in their responses as problems with the items, rather than as reflections of their thinking. It may be true that "misfit is our friend" since important lessons can be learned from student responses that do not make sense according to the learning progression model. However, eventually we may need to confront the question of whether it is possible to characterize student thinking using neatly ordered levels.

The nature of student thinking affects the number of assessment items needed to diagnose students (individually or in groups) with respect to learning progression levels. Researchers and assessment developers may seek rule-of-thumb guidelines for the number of items required for these diagnoses. However, guidelines that apply to traditional items may not work for learning progression items. It may be more difficult to obtain reliable measurements for learning progressions that apply to broad content areas because student thinking simply may be inconsistent across all phenomena included in the learning progression. This difficulty has important implications for broad learning progressions, such as those proposed by the new guidelines for national science education standards (NRC, 2011) and those that are expected to inform large-scale assessments such as NAEP. As discussed in Alonzo et al. (chapter 10)—at least so long as standards and standards-based assessments include many science topics—reliable learning progression assessments will likely require more items than are typically allotted to single content area.

Dependency of Learning Progressions on Students' Curricular Experiences and Related Knowledge/Practices

For traditional assessments, it may be possible to answer questions about item difficulty empirically. As long as an assessment covers the relevant framework and contains an adequate mix of easy and difficult items, it may not matter why one item is more or less difficult than another. While assessment developers may articulate item design models that include hypotheses about item difficulty (e.g., Embretson & Gorin, 2001), assessments may still include items that do not perform as expected (for example, so long as the items provide important content coverage). With learning progression assessments—in which the goal is to characterize student thinking—it may be essential to know what makes an item more or less difficult. Measurement models require a clear sense, as far as possible, of which items are more and less difficult for students—and why. Thus understanding item features and the way students' experiences, knowledge, and practices interact with item content is especially important.

One factor influencing item difficulty may be students' familiarity with particular item formats and contexts. This is largely a matter of transfer and, thus, relates to the issue discussed above—whether assumptions about the nature of student thinking lead to expectations that they can apply knowledge and practices to new contexts. (If student thinking is internally consistent, students should exhibit the same level of knowledge and practices in both familiar and unfamiliar contexts.) Students' curricular experiences may also influence their familiarity with item formats and contexts. While curriculum-dependence may be acceptable (and even desirable) for work within a particular curriculum (e.g., Gotwals et al., this volume), issues of fairness may arise when items are used in large-scale assessments or in comparisons of different curricular approaches.

A second factor influencing item difficulty may be the relationship between the knowledge and practices in a particular learning progression and other related knowledge and practices. To address this factor, Schwarz et al. (chapter 6) and Gotwals et al. (chapter 9) take different approaches to the assessment of scientific practices. Schwarz et al. assessed a practice (scientific modeling) that they hypothesize is transferrable to different science topics. Gotwals et al. intentionally assessed a cross between knowledge and practices. The authors of both chapters acknowledge the constraints that student knowledge of content may place on the assessment of their practices. They acknowledge that student scientific practices may be over- (or under-) estimated due to student familiarity (or unfamiliarity) with relevant science content. As Plummer (chapter 5) and Wilson (chapter 12) discuss, even learning progressions that seemingly have a straightforward focus on content rarely exist in isolation. Learning progressions often make assumptions about student mastery of related knowledge and practices, but these assumptions—if incorrect—can have important consequences for assessment (as well as for instruction) based upon learning progressions. If learning progression items require knowledge and/or practices not included in a certain learning progression, construct-irrelevant variance may be introduced, leading to uncertainty in the estimate of student learning progression levels.

Role of Language

As noted by Alonzo et al. (chapter 10), traditional assessments often reward students for their use of scientific language. In contrast, learning progression assessments look beyond student use of scientific language in order to uncover the underlying reasoning in their responses. To address student use of scientific language, Jin and Anderson (chapter 8) used results from an initial round of student assessments to redefine their learning progression in order to decouple student understanding from student use of scientific language. They also considered how scientific language in prompts impacts the elicitation of student understanding. When items do not signal to older students that details are expected in their responses, they tend to give lower level accounts for scientific phenomena. With younger students, scientific language may be intimidating or confusing. Jin and Anderson addressed these problems by presenting different items to older and younger students. However, in some cases, language may be so integral to the content of a particular learning progression that its use is unavoidable. In these cases, researchers and assessment developers must grapple with the reality that student understanding of language may develop in tandem with the content of a learning progression (e.g., Alonzo, 2010; Alonzo & Steedle, 2009).

CONCLUSIONS

Challenges in the assessing strand have important implications for all aspects of work on learning progressions, particularly the defining strand. The purpose of assessment is to elicit student thinking so that it can be examined and interpreted in terms of a cognitive model (in this case, a learning progression). This interpretation often involves use of a measurement model. Assessment results may provide evidence that is useful for addressing fundamental questions about learning progressions and, thus, about student thinking. The nature of student thinking and its relationship to familiarity with content, practices, and language are fundamental in the design of learning progressions. In addition, assessments may inform decisions about the grain size of learning progressions. Exploring and resolving these issues requires well-crafted assessments and a dynamic interplay between efforts to define and to assess learning progressions.

In order to provide the information needed to revise and validate learning progressions and to assess student knowledge and practices relative to these learning progressions, new assessment technologies may be required. In addition to meeting traditional quality criteria (validity, reliability, and fairness), learning progression assessments must meet additional criteria that may challenge assessment developers. Learning progression assessments seek to elicit performances from students at a range of levels and to characterize student knowledge and practices in the messy middle between naïve and sophisticated scientific performances. To meet these challenges, researchers and assessment developers should develop and evaluate new assessment items and new assessment platforms (such as CAT). This work should be undertaken in collaboration with psychometricians whose measurement models can be used to interpret student responses to assessments.

The NRC (2001) recommended using cognitive and measurement science to inform assessment. Learning progression assessments satisfy this recommendation. In addition, these assessments require new understandings of student thinking and methods for capturing that thinking using assessments and associated measurement models. As such, learning progression assessments may advance cognitive and measurement science beyond that envisioned by the NRC, leading to contributions that benefit the larger field of student assessment.

NOTES

1 Although not yet affecting science, development of Common Core State Standards (CCSS) "began with research-based learning progressions" (Common Core State Standards Initiative, 2010, p. 4). This connection to learning progressions has had a significant effect on the assessment designs considered by the two consortia that are creating assessments of the CCSS—Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium (SBAC)—and on work by the National Center and State Collaborative (NCSC) to develop assessments for students with significant cognitive disabilities (D. Briggs, personal communication, July 20, 2011).

2 Lehrer and Schauble (e.g., 2010), however, define learning progressions with many more levels.

3 For a brief description of CAT, see Wainer (2010). For a more detailed description, see Wainer (2000).

4 The work on learning trajectories in mathematics reflect different assumptions that may be important for learning progressions in science:

> For the most part they [levels of a learning trajectory] are thought to develop gradually out of the preceding level(s) rather than being sudden reconfigurations, and that means that students often can be considered to be partially at one level while showing some of the characteristics of the next… HI [Hierarchical Instructionalism] does not suggest that ways of thinking or operating characteristics [sic] of earlier levels are abandoned—rather students may revert to them if conditions are stressful or particularly complex, or perhaps as they "regroup" before they move to an even higher level. (Daro, Mosher, & Corcoran, 2011, p. 24)

REFERENCES

Alonzo, A. C. (2009, April). Design criteria for learning progressions to support teachers' formative assessment practices In A. C. Alonzo (Chair), *Learning progressions in science: Tools for teaching,* *learning, and assessment*. Symposium conducted at the annual meeting of the American Educational Research Association, San Diego, CA.

Alonzo, A. C. (2010). Discourse as a lens for reframing consideration of learning progressions. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010) - Volume 1, Full Papers* (pp. 588–595). Chicago, IL: International Society of the Learning Sciences.

Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research and Perspectives, 9*, 124–129.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Roadmaps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives, 9*, 71–123.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Common Core State Standards Initiative. (2010). *Common Core State Standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform*. (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Daro, P., Mosher, F. A., & Corcoran, T. (2011, January). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (CPRE Research Report #RR-68). Philadelphia, PA: Consortium for Policy Research in Education.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*, 105–225.

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*, 843–900.

Educational Testing Service. (2009, December). *Response to request for input on the Race to the Top assessment program*. Retrieved from http://www.ets.org/Media/Home/pdf/ETS_Response_RacetotheTopAssessment.pdf

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343–368.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Ioannides, C., & Vosniadou, S. (2001). The changing meanings of force: From coherence to fragmentation. *Cognitive Science Quarterly, 2*(1), 5–62. Retrieved from University of Athens website: http://www.cs.phs.uoa.gr/el/staff/vosniadou/force.pdf

Lehrer, R., & Schauble, S. (2010, March). *Seeding evolutionary thinking by engaging children in modeling its foundations*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Messick, S. (1989).Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003).On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

National Assessment Governing Board. (2008, September). *Science framework for the 2009 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/publications/frameworks/science-09.pdf

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Shavelson, R. J., Yin, Y., Furtak, E. M., Ruiz-Primo, M. A., Ayala, C. C., Young, D. B.., . . . Pottenger, F. (2008). On the role and impact of formative assessment on science inquiry teaching and learning. In J. E. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice*. (pp. 21-36). Arlington, VA: NSTA Press.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2010). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching, 47*, 687–715.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.

Wainer, H. (2000). *Computer-adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H. (2010). Computerized adaptive testing. In I. B. Weiner & W. E. Craighead (Eds.), *Corsini Encyclopedia of Psychology*. doi: 10.1002/9780470479216.corpsy0213

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Retrieved from University of Wisconsin, Wisconsin Center for Education Research website: http://facstaff.wcer.wisc.edu/normw/WEBBMonograph6criteria.pdf

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability: The 103rd yearbook of the National Society for the Study of Education, Part II*. Chicago, IL: The University of Chicago Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

MODELING LEARNING PROGRESSIONS

PATTI WEST, DAISY WISE RUTSTEIN, ROBERT J. MISLEVY, JUNHUI LIU, ROY LEVY, KRISTEN E. DICERBO, AARON CRAWFORD, YOUNYOUNG CHOI, KRISTINA CHAPPLE AND JOHN T. BEHRENS

A BAYESIAN NETWORK APPROACH TO MODELING LEARNING PROGRESSIONS

A central challenge in using learning progressions (LPs) in practice is modeling the relationships that link student performance on assessment tasks to students' levels on the LP. On the one hand, there is a progression of theoretically defined levels, each defined by a configuration of knowledge, skills, and/or abilities (KSAs). On the other hand, there are observed performances on assessment tasks, associated with levels but only imperfectly and subject to inconsistencies. What is needed is a methodology that can be used to map assessment performance onto the levels, to combine information across multiple tasks measuring similar and related KSAs, to support inferences about students, and to study how well actual data exhibit the relationships posited by the LP. In terms of the "assessment triangle" proposed by the National Research Council's Committee on the Foundations of Assessment (National Research Council [NRC], 2001), coherent theoretical and empirical connections are needed among the theory embodied in a progression (cognition), the tasks that provide observable evidence about a student's understanding relative to that progression (observation), and the analytic models that characterize the relationship between them (interpretation).

This chapter discusses the use of Bayesian inference networks, or Bayes nets for short, to model LPs. Bayes nets are a class of statistical models that have been adapted for use in educational measurement. At present, the use of Bayes nets in LP contexts is in its relative infancy. We describe the fundamentals of the approach and the challenges we faced applying it in an application involving a LP in beginning computer network engineering.

The first section of the chapter reviews our framework of model-based reasoning. Subsequent sections map the development of LPs and associated assessments onto this framework and show how Bayes nets are used to manage the problems of evidence and uncertainty in the relationship between LPs and assessment task performances. We then explain in more detail what Bayes nets are, how they can be used to model task performance in the context of LPs, and the challenges that we face in this work.

MODEL-BASED REASONING

The lens of model-based reasoning helps clarify the role Bayes nets can play in modeling LPs. A model is a simplified representation focused on certain aspects of

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 257–292.

a system (Ingham & Gilbert, 1991). The entities, relationships, and processes of a model provide a framework for reasoning about any number of real world situations, in each instance abstracting salient aspects of those situations and going beyond them in terms of mechanisms, causal relationships, and/or implications that are not apparent on the surface.

The lower left plane of Figure 1 shows phenomena in a particular real world situation. In the case of LP research, the situation is students' task performances. A mapping is established between this situation and, in the center of Figure 1, the semantic plane of the model; that is, structures expressed in terms of the entities, relationships, and properties of the model. The lines connecting the entities in the model represent causes, influences, mechanisms, and other relationships. The analyst reasons in these terms. In modeling LPs, this layer concerns progressions and their levels, relationships among different progressions, and expected performance on assessment tasks based on the features of tasks (what students are asked to do) and the features of their performances (what they actually do).

The real world situation is depicted in Figure 1 as fuzzy, whereas the model is well-defined. This suggests that the correspondence between real world entities and the idealizations in the model is never exact. The reconceived situation in the lower right plane of Figure 1 is a blend of selected aspects of the real world situation and elements of the model (shown in dotted form). The match between the real world and the data is not perfect, but a framework of meaning that the situation does not possess in and of itself can enhance our understanding of it (Suarez, 2004; Swoyer, 1991). It is here that descriptions, explanations, and implications for real world phenomena are formed. In the case of LPs, it is here that patterns of students' performance are interpreted in terms of their status or development with respect to the LP levels.

Symbol systems that are associated with some models further support reasoning, such as the algebraic and graphical representations of regression models shown above the semantic plane in Figure 1 as Representational Forms A and B. Similarly, Bayes nets provide mathematical and graphical representations to support reasoning about LPs, students' status on them, and evaluations of their performances across tasks.

DEVELOPMENT, ASSESSMENT, AND MODELING OF LEARNING PROGRESSIONS

When we speak of modeling a LP, we refer to a coherent set of elements: a progression defined in terms of the psychology and the substance of the domain under consideration, a specification of how real-world situations can be set up to evoke evidence about a student's status on the LP, and a measurement model (in our case, a Bayes net) that articulates the probabilistic relationship between student performances and status on the LP. These are the vertices of an "assessment triangle" (NRC, 2001): cognition, observation, and interpretation. Cognition refers to a theory about what students know and how they know it (the learning progression). Observation relates to the tasks we ask students to perform to gather evidence about what they know. Interpretation is the meaning we assign to these observations. Specifying and validating a probability model—Bayes nets in this case—helps analysts develop coherence among these elements in order to reason

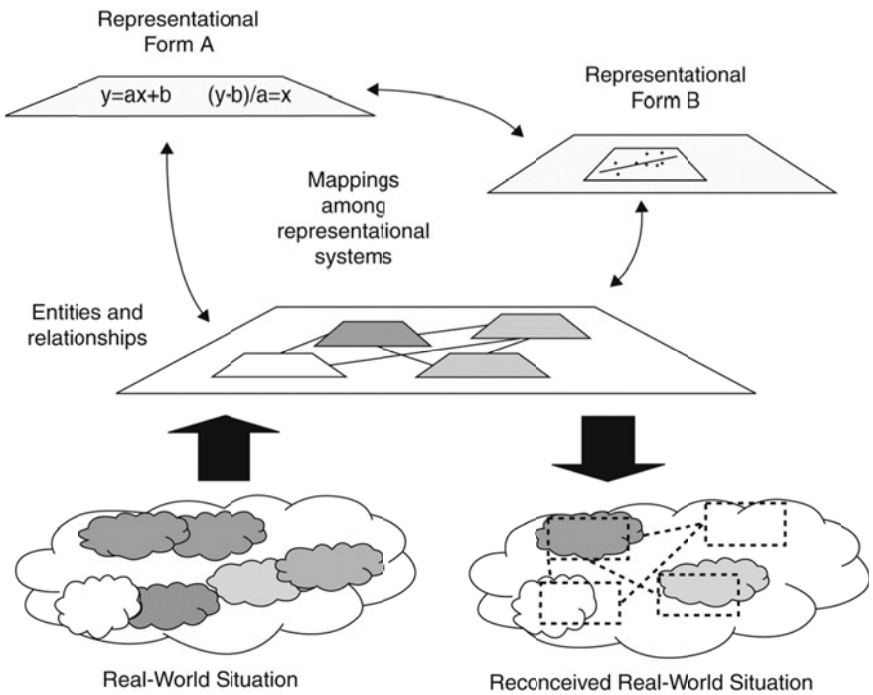


Figure 1. Reconceiving a real world situation through a model. From "Validity from the Perspective of Model-Based Reasoning," by R. J. Mislevy in The Concept of Validity: Revisions, New Directions, and Applications by R. L. Lissitz (Ed.) (p. 95), 2009, Charlotte, NC: Information Age Publishing. Copyright 2009 by Information Age Publishing. Reproduced with permission of Information Age Publishing.

about students' levels in the progression from their task performances. This section provides an overview of our work with LPs in the Cisco Networking Academy using this framework.

Learning Progression Development (Cognition)

LPs are empirically grounded and testable hypotheses about how a student's understanding and ability to use KSAs in a targeted area develop over time (Corcoran, Mosher, & Rogat, 2009). LPs are the pathways that bridge the gap between some starting point, such as the inability to connect two routers in computer networking, and a desired endpoint, such as the successful configuration of routers (National Research Council [NRC], 2007). There are five components of LPs (Corcoran et al., 2009; NRC, 2007): (1) learning targets or clear end points that are defined by social aspirations, (2) progress variables that identify the critical dimensions of KSAs that are developed over time, (3) levels of achievement or stages that define significant intermediate steps, (4) learning performances that are the operational definitions of what KSAs would look like at each stage, and (5) assessments that measure performances with respect to each key KSA over time. Although students may progress along different pathways, common paths can be tested and legitimated. It should also be noted that student learning and thinking progress in the context of instruction and experiences. This progression must be considered in creating, assessing, modeling, and interpreting LPs.

The following discussion describes the development of LPs in a specific content area: beginning computer networking. The context is the Cisco Networking Academy (CNA), a global program in which information technology is taught through a blended program of face-to-face classroom instruction, an online curriculum, and online assessments. These courses are offered in high schools, 2- and 3-year community college and technical schools, and 4-year colleges and universities. Since its inception in 1997, CNA has grown to reach a diverse population of approximately 900,000 students annually, in more than 160 countries (Levy & Murnane, 2004; Murnane, Sharkey, & Levy, 2004). Behrens, Collison, and DeMark (2005) discuss the framework that drives the ongoing assessment activity that, in turn, provides the data for this work.

In 2007, CNA updated and redesigned the curriculum for its primary network course offerings. A group of subject matter experts, working with psychometricians and educational psychologists, sketched out provisional LPs based upon several lines of work integral to the design of the curriculum (for details, see West et al., 2010). First, they conducted statistical analyses of student exams from the previous four-course curriculum. Classical and item response theory (IRT) analyses of end-of-chapter and final exam data revealed patterns in the difficulty of certain assessment tasks based upon their placement in the curriculum. For example, the same item used to assess IP addressing had different difficulty depending on whether the item was used before or after students learned basic routing concepts. Second, these patterns were considered in combination with the results of cognitive task analysis research into novice and expert performance in the domain (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2007; DeMark & Behrens, 2004). Finally, external research highlighting the real-world KSAs necessary for various job levels was used to validate the subject matter expert opinion and statistical analyses. Thus the initial LP framework was developed through the interaction of various experts using both theory and data.

To make this discussion more concrete, Table 1 presents an example of a LP in Internet Protocol (IP) Addressing, a key area taught in the four-semester Cisco Certified Network Associate (CCNA) course sequence. IP addressing is the mechanism by which all pieces of equipment in a network (PCs, routers, etc.) are given unique "addresses" so information sent to them knows where to go and information sent from them is properly labeled for return if necessary. An analogy is the street address of a house. A five-level progression is defined based on clusters of interrelated, assessable elements that describe a student's capabilities at each level. The levels reflect increasingly sophisticated understandings of IP Addressing. Table 1 presents an abridged version of the KSAs at each level.

Table 1. Sample of Knowledge, Skills, and Abilities in the IP_Addressing Progression.

Level 1 – Novice – Knows Pre-requisite Concepts: Can recall factual information and perform highly scripted activities • Student can navigate the operating system to get to the appropriate screen to configure the address. • Student can use a web browser to check whether or not a network is working.
Level 2 – Basic – Knows Fundamental Concepts: Able to understand reasoning behind actions, but can't apply in unknown situations • Student understands that an IP address corresponds to a source or destination host on the network. • Student understands that an IP address has two parts, one indicating the individual unique host and one indicating the network that the host resides on. • Student understands the default gateway is the address that data is sent to if the data is leaving the local network and why it must be specified. • Student understands how the subnet mask indicates the network and host portions of the address. • Student can create subnet masks based on octet boundaries.
Level 3 – Intermediate – Knows More Advanced Concepts: Able to apply concepts to actions • Student understands the difference between physical and logical connectivity. • Student can explain the process of encapsulation. • Student understands how Dynamic Host Control Protocol (DHCP) dynamically assigns IP addresses.
Level 4 –Advanced – Applies Knowledge and Skills: Able to apply concepts in context in an unscripted manner • Student can use the subnet mask to determine what other devices are on the same local network as the configured host. • Student can use a network diagram to find the local network where the configured host is located. • Student can recognize the symptoms that occur when the IP address or subnet mask is incorrect.
Level 5 – Expert – Applies Advanced Knowledge and Skills: Able to apply concepts in new contexts in an unscripted manner and predict consequences of actions • Student can recognize a non-functional configuration by just looking at the configuration information; no testing of functionality is required. • Student can interpret a network diagram to determine an appropriate IP address/subnet mask/default gateway for a host device. • Student can interpret a network diagram in order to determine the best router to use as a default gateway when more than one router is on the local network.

Task Design (Observation)

The CNA assessment development process follows an Evidence Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003) approach. ECD guides the assessment design process by addressing a series of questions: “What claims or inferences do we want to make about students?” “What evidence is necessary to support such inferences?” “What features of observable behavior facilitate the collection of that evidence?” At each level of the LP, a subject matter expert created multiple claims based on the set of related KSAs that define the level. In order to assess student performance with respect to these claims, the curriculum contains end-of-chapter tests and end-of-course final exams consisting of multiple-choice questions. Each chapter or course typically addresses multiple LPs. Our current focus is the case in which each item in an assessment is designed to measure one LP level. Work on modeling more complex assessment tasks that address multiple LPs is discussed later in the chapter.

In this example, IP_Addressing is called a student model variable (SMV) because it represents an aspect of a student’s proficiency. SMVs, like IP_Addressing, are latent variables, which means their values cannot be observed directly. However, students’ task performances provide evidence about them. Two items that provide evidence about a student’s level on IP_Addressing are shown in Figure 2. They both concern knowledge of the syntax of a router command. These two seemingly similar items provide evidence to distinguish between different levels of a LP due to a small but conceptually important difference in task features: Changing the stem from /24 to /28 requires students to have a more advanced IP Addressing skill, namely the skill to subdivide one of the octets. Item A distinguishes between Level 1 and Level 2 (students can create subnet masks based on octet boundaries), while Item B distinguishes between Level 3 and Level 4 (students can use the subnet mask to determine what other devices are on the same local network as the configured host).

Modeling Responses (Interpretation)

We can represent the different patterns of evidence provided by the sample items in Figure 2 with a Bayes net. First, a student’s level on the IP_Addressing LP can be represented with a variable called IP_Addressing. The variable has five possible values, one for each level of the LP. For each level, there is a probability that a student is at that level. Figure 3 represents ignorance about a student’s level, expressed as probabilities of .2 at each level. (The Netica program, Norsys Software Corp., 2007, shows these probabilities as percentages, hence 20 rather than .20.) This is called a *prior* probability distribution, reflecting the belief about a student before observing any of the student’s responses. We will see how observation of student responses allows us to update our beliefs and express them in a *posterior* probability distribution.

Item A	**Item B**
It is necessary to block all traffic from an entire subnet with a standard access control list. What IP address and wildcard mask should be used in the access control list to block only hosts from the subnet on which the host 192.168.16.43/**24** resides?	It is necessary to block all traffic from an entire subnet with a standard access control list. What IP address and wildcard mask should be used in the access control list to block only hosts from the subnet on which the host 192.168.16.43/**28** resides?
A.192.168.16.0 0.0.0.15	A.192.168.16.0 0.0.0.15
B.192.168.16.0 0.0.0.31	B.192.168.16.0 0.0.0.31
C.192.168.16.16 0.0.0.31	C.192.168.16.16 0.0.0.31
D.192.168.16.32 0.0.0.15	**D.192.168.16.32 0.0.0.15
E.192.168.16.32 0.0.0.16	E.192.168.16.32 0.0.0.16
**F.192.168.16.0 0.0.0.255	F.192.168.16.0 0.0.0.255

Figure 2. Sample items assessing levels of IP_Addressing. From A Bayesian Network Approach to Modeling Learning Progressions and Task Performances (CRESST Report 776) (p. 6), by P. West, D. W. Rutstein, R. J. Mislevy, J. Liu, Y. Choi, R. Levy, A. Crawford, K. E. DiCerbo, K. Chappel, and J. T. Behrens, 2010, Los Angeles, CA: University of California, CRESST. Copyright 2010 The Regents of the University of California. Reprinted with permission of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

To this end, our simple Bayes net also contains variables for responses to Item A and Item B. They are called observable variables (OVs) because we learn their values, usually with certainty, when we observe a student's responses. Their possible values are 1 and 0, scores for right or wrong responses, respectively. In our example, Item A distinguishes between Level 1 and Level 2, and Item B distinguishes between Level 3 and Level 4. The Bayes net indicates these relationships through conditional probability distributions for OVs, such as scores on Item A and Item B given a student's level on IP_Addressing, the latent SMV. The distributions indicate the conditional probability of getting Item A right or wrong at each level of the IP_Addressing LP.

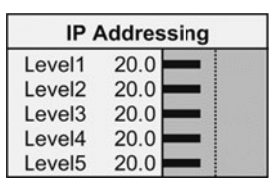


Figure 3. A depiction of a student model variable that represents a five-level learning progression, indicating equal probabilities that a student is at each level. (Note: This figure and the Bayes net graphics that follow were produced with the Netica computer program. Probabilities are displayed in the percent metric and thus sum to 100 rather than to 1.)

Table 2a specifies the relationship between IP_Addressing and Item A. Each row in the table is the conditional probability distribution for the values of the Item A OV, given the value of IP_Addressing. The row for Level 1, for example, says that a student at Level 1 has a probability of .8 of answering incorrectly and only .2 of answering correctly. (We discuss the source of these probabilities later in the chapter.) A student at Level 2 has a probability of .7 of answering correctly. Table 2 reflects the LP structure since students at Level 1 will probably get Item A wrong, but students at or above Level 2 will probably get Item A right. These expectations are probabilities rather than certainties because a student at Level 4 might miss Item A owing to carelessness, an arithmetic error, or a gap in knowledge. Although the capabilities at a given level are interrelated for both concepts and curriculum, students may be stronger on some elements at one level than on others. Table 2b, which gives conditional probabilities for Item B, shows a jump in conditional probabilities between Level 3 and Level 4.

Table 2. Conditional Probabilities for Item Responses Given the Level of IP_Addressing.

a) Item A

	Item A	
IP_Addressing	*Score 0*	*Score 1*
Level 1	80	20
Level 2	30	70
Level 3	20	80
Level 4	10	90
Level 5	10	90

b) Item B

	Item B	
IP_Addressing	*Score 0*	*Score 1*
Level 1	90	10
Level 2	80	20
Level 3	70	30
Level 4	20	80
Level 5	10	90

To summarize this section, theoretically defined levels of the learning progression provide information about what students know and how they know it. This is the cognition vertex in the assessment triangle. The theory and research underlying the LPs suggest how we might design tasks to elicit student performances that depend on their status in the progressions. This is the observation vertex. The interpretation vertex of the triangle addresses analytic models that connect assessment performances with the cognitive structure of the LP; these models are used to validate and improve the LP and task framework and to reason about individual students in that framework. The following sections explore this vertex using Bayes nets.

MODELING LPS USING BAYES NETS

Even the simplest LP structure poses issues of evidence and uncertainty since a student at a given level of a progression may provide responses that vary across levels from one task to the next. What degree of regularity should we expect in performance? How do we infer back from students' performances to the levels at which they likely work? How much evidence about a level does one task, or several tasks at different levels, provide us? Does a particular task operate differently than others when our theory says they should operate similarly? Additional complexities arise when we consider multifaceted clusters of concepts. Are there multiple typical paths students take that extend beyond the usual variability in performance? Are there identifiable strands of concepts that display their own regularities within a larger, less tightly structured progression? What hard prerequisites, or soft tendencies, seem to influence students' paths? How do we discover these patterns in noisy data?

Measurement models posit a relationship between (a) a student's status on inherently unobservable SMVs that characterize some aspects of their capabilities and (b) OVs that provide evidence about these SMVs. Specifying a measurement model becomes a matter of specifying these relationships and, in so doing, specifying how assessment data should be interpreted to yield inferences about students. In addition to Bayes nets, other modern measurement models with this same essential structure include latent class models (Dayton & Macready, 2007), cognitive diagnosis models (Rupp & Templin, 2008), and structured IRT models such as the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Adams, Wilson, & Wang, 1997). In the context of LPs, the SMVs correspond to LPs since students are presumed to be at a given but unobservable level on each LP. We obtain evidence about their level from performance on tasks built, based on the theory underlying the LP, for this purpose.

More specifically, modern measurement models facilitate inferences about students using two key features. The first is latent variables. These variables recognize that what we would ultimately like to know about students (i.e., their levels on the LP) is unobservable, and must be inferred from what we can observe, namely, performance on tasks. The relationship between a student's level on the LP—the latent SMV—and performance on the tasks—captured in OVs—is at the heart of the inference. By specifying which values of the OVs (i.e., the performances on tasks) are expected based on the value of the SMV (i.e., the level of the LP), the measurement model allows us to make inferences about the SMV from observed values of the OVs.

The second feature of modern measurement models is the use of probability models to express these relationships. Student performances (OVs) are modeled as probabilistically dependent on the student's level on the LP (SMV). A student may exhibit task performances that do not exactly agree with the expectations based on the model. For example, a student who has reached a given level of the LP might demonstrate a higher or lower level of performance on a particular task; task performance could be the result of chance, of inconsistency in applying concepts, or of the influence of factors not encoded in the model. This is why the conditional probabilities in the introductory example (Table 2) are not all ones and zeros.

Combining these two features produces a modern measurement model—performances on tasks (OVs) are modeled as probabilistically dependent on the unobservable level of the LP (SMV). Box 1 provides a formal definition of a measurement model formulation in the LP paradigm. Given such a model, we can characterize tasks' effectiveness at distinguishing between levels (through the patterns in the conditional probabilities as estimated from data), and we can draw inferences about the status of students on a LP (as we will see shortly, through posterior probability distributions once we observe students' performances). Further, probability theory helps a researcher explore the fit and misfit of a model to data and iteratively fine-tune both tasks and theories.

==

Box 1: Formal definition of a measurement model

To more formally define the measurement model structure used in the modeling of LPs, let θ denote an unobservable SMV. Further, let X_1, X_2,…, X_J represent some number J of OVs, the values of which summarize performance on tasks (e.g., scored item responses). A measurement model then specifies the *conditional probability* for each OV, denoted $P(X_j \mid \theta)$. The conditional probability expression yields different probabilities of values of the OV depending on the value of the SMV, capturing how a student's performance depends on his/her level of proficiency. Each OV is permitted to have its own conditional probability distribution given the SMV, as tasks may differentially measure the KSAs.

==

Bayesian Inference Networks

Bayes nets combine probability theory and graph theory to represent probabilistic relationships among variables. Bayes nets are so named because they support reasoning from any set of observations to any other variables (either latent or observable but not yet observed) in a network using algorithms that incorporate Bayes' theorem (Lauritzen & Spiegelhalter, 1988; Pearl, 1988). As a general modeling approach, Bayes nets focus on conditional probabilities in which the probability of one event is conditional on the probability of other events: in forecasting, for example, probabilities of tomorrow's weather given today's weather and climate patterns; in animal breeding, characteristics of offspring given characteristics of ancestors; in medical diagnosis, probabilities of syndromes given disease states and of test results given syndromes. In assessment, interest lies in item responses or features of performances given students' KSAs. Bayes nets can be used to structure relationships across large multivariate systems, allowing us, for example, to synthesize the results of many observed responses to support inferences about student thinking (Mislevy & Gitomer, 1996).

One way to represent these networks of variables and the resulting computations is with a graphical model (such as Figure 4) consisting of the following elements (Jensen, 1996):

- A set of variables, represented by ellipses or boxes and referred to as nodes. All the variables in the most widely used Bayes nets have a finite number of possible values, corresponding to a set of exhaustive and mutually exclusive states (e.g., the IP addressing example has five mutually exclusive LP levels that comprise all possible states of IP_Addressing as it is being modeled).
- A set of directed edges (represented by arrows) between nodes, indicating probabilistic dependence between variables. Nodes at the source of a directed edge are referred to as parents of nodes at the destination of the directed edge, their children. In our example, IP_Addressing is the parent of Item A and Item B. The direction of edges is often determined by theory, such as disease states as the parents of symptoms or the status of some indicator at Time *k* as a parent of status at Time *k*+1.
- For each variable without parents (such as IP_Addressing), there is an initial probability distribution for that variable. This could be uninformative, as in Figure 2, or based on background knowledge about a group or individual.
- For each variable with parents (such as Item A in Figure 4), there is a set of conditional probability distributions corresponding to all possible combinations of the values of the parent variables (as in the rows of Table 2a).

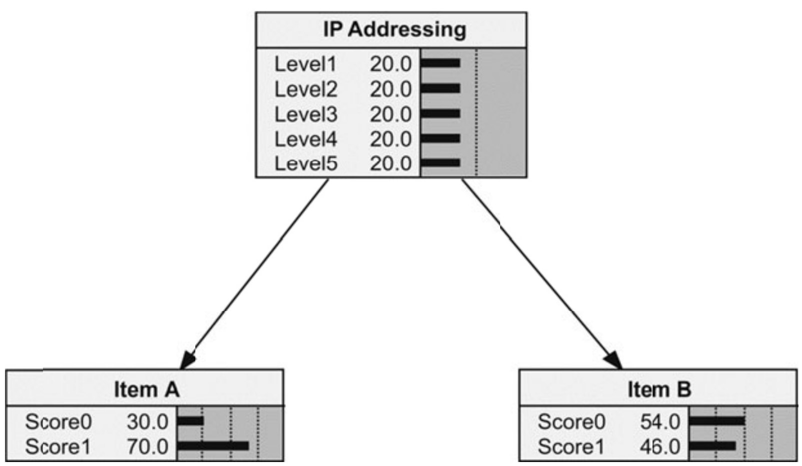


Figure 4. A Bayes net showing observable variables for two tasks. The performance is dependent on a student's level on the learning progression represented in the student model variable IP_Addressing.

Box 2 more formally describes Bayes nets as probability models. Because Bayes nets are framed in terms of the standard theory of probability and statistics, general approaches to model construction, parameter estimation, and model criticism are available to researchers seeking to model LPs or any other substantive situation.

==

Box 2: Bayes nets as probability models

In general, Bayes nets can be described as a probability model for the joint distribution of a set of finite-valued variables, say ($Y_1,\ldots Y_N$), represented recursively in terms of the product of conditional distributions:

$$P\left(Y_1,\ldots,Y_N\right)=\prod_j P\left(Y_j\,\middle|\,Pa\left(Y_j\right)\right), \qquad (1)$$

where $Pa(Y_j)$ refers to the subset of variables with indices lower than j upon which Y_j depends. These are the variables that have edges pointing from them to Y_j in the graphical representation of the network. Theory and experience suggest which variables should be considered parents of others. For example, in weather forecasting, variables for today's conditions are parents of variables for tomorrow's conditions. In genetics, variables representing genotypes of individuals are parents of variables representing the phenotypes of the same individuals, and variables for genotypes of literal parents are Bayes net parents of variables for the genotypes of their literal children. Theory and experience also provide information for determining if one set of variables should be modeled as independent from another set of variables, given the values of a third set of variables ("conditional independence"). For example, in Table 2, the probabilities of the Item A responses are independent of the probabilities of Item B responses given information about the levels of IP_Addressing; if we knew the value of a student's IP_Addressing variable, observing the value of the Item A response would not change our expectations for her response to Item B. When theory and experience suggest many conditional independence relationships, the variables in a Bayes net will have relatively few parents, and the diagram and the recursive expression simplify. The relationships among variables in the network can then be expressed in terms of interactions among relatively small clusters of variables.

Once such a representation has been built, one can update belief about any subset of the variables given information about any other subset using Bayes theorem. The rapid increase of the use of Bayes nets is due to efficient algorithms that allow these computations to take place in real time when the dependency structure is favorable.

In the context of Bayesian networks for LPs, θ is a latent discrete SMV with states that correspond to the levels of the progression. Formally, the values can be ordered, partially ordered, or unordered. A single LP would typically be represented by ordered levels. The OVs are discrete variables with states corresponding to the different possible scored performances on items or other tasks (e.g., a correct or incorrect response on an item, or levels or types of performance qualities in more complex tasks). The Bayes net specifies $P(X_j \mid \theta)$, a table of the conditional probabilities of observing different performances on tasks given the student's level on the LP. Multiple tasks yield OVs that may have different associated conditional probability tables. For example, one item may require a student to be at least at a low level of the progression in order to have a high probability of performing well, whereas another item requires the student to be at a higher level to have a high probability of performing well. In the case of the two items for IP_Addressing (Table 2), Item A requires a student to be at level 2 or above on IP_Addressing in order to have a high probability of getting a correct score, while Item B requires a student to be at level 4 or level 5 to have a high probability of answering correctly. The specification of the model is completed by defining an initial probability distribution for θ—i.e., a prior distribution—capturing how likely it is that a student is at each level of the progression. The prior may be uninformative, as in the introductory example (Figures 2 and 4), or based on other information such as student background data or instructors' expectations.

When a Bayes net is specified in this way to model assessment of a single discrete SMV, it can be viewed as a latent class model (Dayton & Macready, 2007; Lazarsfeld & Henry, 1968). A traditional formulation for a C-class latent class model (i.e., a model with C levels of a learning progression) specifies the probability that examinee i responds to item j yielding an OV value of r as

$$P(X_{ij} = r) = \sum_{c=1}^{C} P(\theta_i = c) P(X_{ij} = r \mid \theta_i = c), \tag{2}$$

where $P(\theta_i = c)$ is the prior probability that examinee i is in class c (i.e., level c of the progression) and $P(X_{ij} = r \mid \theta_i = c)$ is the conditional probability that an examinee in class c responds to item j in response category r. The usual restriction, $\sum_{c=1}^{C} P(\theta_i = c) = 1$, is imposed. Similarly, within latent classes the conditional probabilities over response categories are restricted such that $\sum_{r=1}^{R_j} P(X_{ij} = r \mid \theta_i = c) = 1$, where R_j is the number of distinct response categories for item j. The graphical representation contains edges from θ to each X (e.g., the edges from IP_Addressing to both items in Figure 4). The recursive representation is

$$P(X_1, \ldots, X_N, \theta) = \prod_j P\left(X_j \middle| Pa(X_j)\right) = \prod_j P\left(X_j \middle| \theta\right) P(\theta). \tag{3}$$

More complex cases can include multiple LPs as well as progressions that allow for different pathways so that the Bayes net must address a finer grain-size of KSAs to distinguish points along different pathways. In these cases, θ is vector-valued. Performance on a given observable X from a task can depend on more than one component of θ; that is, conditional probabilities for such an observable are estimated for possible combinations of its entire set of parent SMVs. In networking, for example, doing well on a certain troubleshooting task may require a student to be at Level 3 or higher in the IP_Addressing progression and at Level 2 or higher in the Connectivity (also called “Connect Networks”) progression.

==

To continue with our example, the graphical representation of a Bayes net depicts the structure of relationships among variables—in our case, how performance on tasks depends on LP levels—and probabilities that represent the analyst’s knowledge of a student at a given point in time. Probabilities that arise from (1) knowing nothing about a particular student’s level in IP_Addressing and (2) knowing the conditional probabilities of item responses from Table 2 are shown in Figure 4. The direction of the arrows reflects the direction of the conditional probabilities in the tables, namely, that item performance depends on the student’s level in IP_Addressing.

Once this probability structure has been built, we can reason in the other direction as well. We work back through an arrow to obtain a better estimate about an individual student’s level on the LP, given the response the student makes to a given item, and then revise what we would expect to see on other items as a result. Figure 5 shows that if a student answers Item A incorrectly, he or she is probably at Level 1. The probabilities for IP_Addressing are obtained by applying Bayes’ theorem as follows: Multiply the initial probabilities for each level of the LP (in this case, .2) by the corresponding conditional probabilities in the column for Score 0 of Table 2a; then normalize the result (i.e., rescale the results of the multiplications so they add to 100%). The result gives the posterior probabilities (IP_Addressing in Figure 5). These updated probabilities can then be used to obtain the probabilities for Item B. This updating is a simple example of Bayes’ theorem with just two variables. In more complicated networks, algorithms are used that build on Bayes’ theorem but take advantage of conditional independence structures to update many variables efficiently (Jensen, 1996).

Figure 6 shows that if the student answers item A correctly, s/he is probably at Level 2 or higher. Figure 7 shows that if the student who answers Item A correctly also answers Item B incorrectly, belief shifts to Levels 2 and Level 3. (The probabilities for IP_Addressing in Figure 4 have now been combined with the column for Score 0 in the Item B conditional probability table, Table 2b). If we wanted to sort out these possibilities, we would administer an item that focuses on capabilities that emerge in Level 3. Finally, if the student had answered B correctly, then our belief would shift to Level 4 and Level 5 (Figure 8).

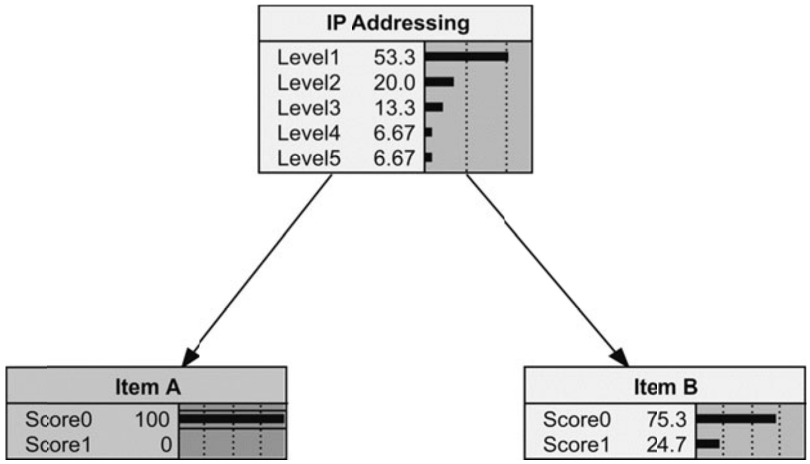


Figure 5. A Bayes net showing updated belief about IP_Addressing and expectations for Item B, after having observed an incorrect response to Item A.

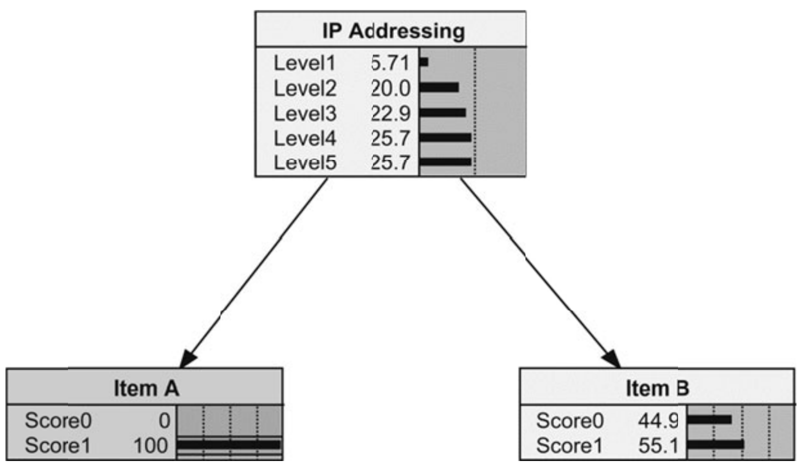


Figure 6. A Bayes net showing updated belief about IP_Addressing and expectations for Item B, after having observed a correct response to Item A.

We stated earlier that we would say more about the source of the numbers in the conditional probability matrices. At this point, we note the essential idea: We posit the basic structure of relationships (e.g., how LPs relate to each other and how LPs relate to tasks) and of the probabilities (e.g., where the jumps are) from our theory about learning and the way we construct tasks. We collect student responses to see how well theory fits data. If the model fits, we use the data to estimate the

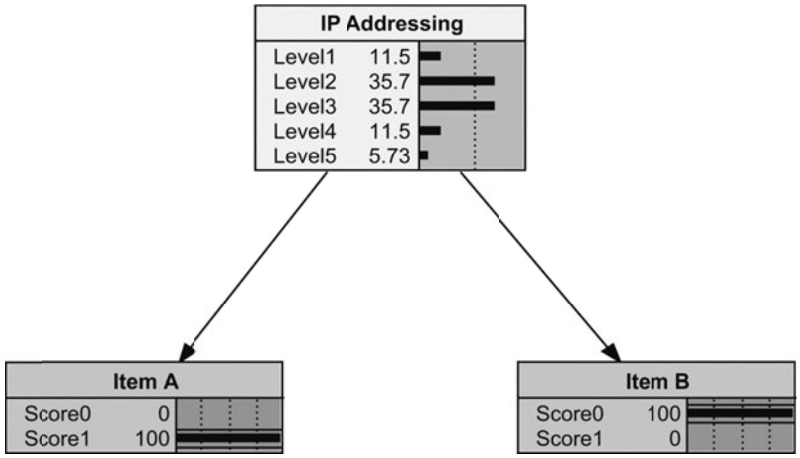


Figure 7. A Bayes net showing updated belief about IP_Addressing after having observed a correct response to Item A and an incorrect response to Item B.

probabilities. Bayesian model fitting and model-checking approaches such as described by Gelman, Carlin, Stern, and Rubin (1995) can be used. See Mislevy, Almond, Yan, and Steinberg (1999) and Levy and Mislevy (2004) for details on modeling and estimation approaches for fitting large conditional probability matrices for more complex Bayes nets. If the model doesn't fit, we use the data to revise the model, the theory, or the way we collect data (e.g., revision of tasks).

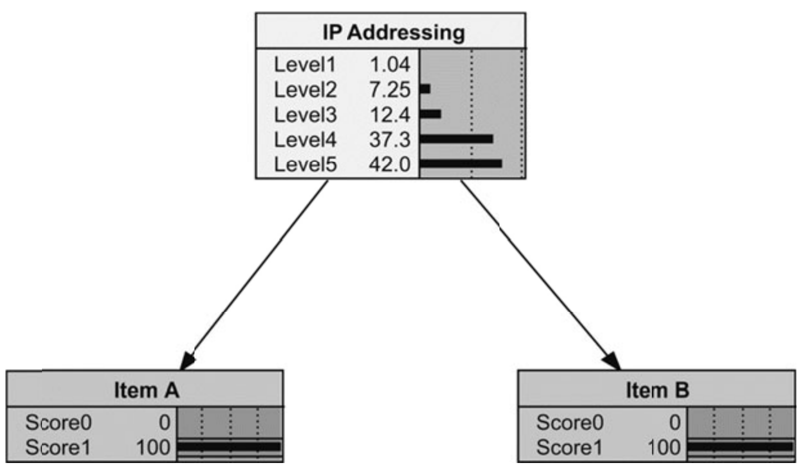


Figure 8. A Bayes net showing updated belief about IP_Addressing after having observed correct responses to both Item A and Item B.

CHALLENGES IN USING BAYES NETS TO MODEL LEARNING PROGRESSIONS

A number of challenges exist when using Bayes nets in modeling LPs and associated data from assessments targeting LPs. One challenge concerns the development of a toolkit of Bayes net techniques tuned for modeling assessment in the context of LPs. Bayes nets support probability-based inferences in complex networks of interdependent variables and are used in such diverse areas as forecasting, pedigree analysis, troubleshooting, expert systems, jurisprudence, intelligence analysis, and medical diagnosis (e.g., Pearl, 1988). When using Bayes nets in any new domain (like LPs), it is a challenge to develop an experiential base and modeling strategies using the general model-building, model-fitting, and model criticism tools of Bayes nets to address the relationships that are particular to that domain. For example, Bayes net fragments for properties of weather patterns and meteorological instruments can be assembled and tailored for weather forecasting applications (e.g., Edwards, 1998). Fragments concerning witness credibility and lines of argumentation also recur in Bayes nets in legal evidentiary arguments (e.g., Kadane & Schum, 1996). The unique features of LPs dictate that certain recurring structures will likely be present in applications of Bayes nets used to model LPs. We discuss three of these applications—(1) interrelationships among LPs; (2) KSA acquisition over time; and (3) evidence from complex tasks—which are likely to be part of sophisticated applications of modeling LPs.

A second, more local, challenge arises when one applies Bayes nets to model any specific substantive LP. In every application there are challenges in defining the LP, creating tasks, and iteratively fitting and improving the model and theory. In the NRC's (2001) terms, the cognition, observation, and interpretation components of an assessment must cohere. Bayes nets instantiate the last of these components. In and of themselves Bayes nets do not dictate the choices faced by researchers in any application, including the grain-size and number of levels in LPs. The definition and modeling of the middle levels of a LP present specific challenges in connecting cognition, observation, and interpretation.

There is a continual interplay between these two kinds of challenges. A Bayes net toolkit for LP research, at any stage of development, aids the analyst in all projects. Every project has its unique wrinkles, offers the possibility of insights about model structures or modeling strategies that may be more broadly useful for successive projects, and, as such, motivates expressing these new understandings in resources for the toolkit. Since Bayes net analysis of LPs is relatively new, we note in the following discussion the local challenges we faced. These challenges highlight recurring patterns that the field may expect to encounter more broadly in modeling LPs. For the interested reader, the Appendix gives details of an application of Bayes nets to modeling a learning progression in the CNA context.

Interrelationships Among LPs

The IP_Addressing example we discuss in this chapter concerns a single LP. In any complex domain, however, multiple KSAs must be developed not only with respect to sophistication in and of themselves but also in terms of their connections to other KSAs, and jointly as the basis for more integrated understandings. The knowledge maps in the *Atlas of Science Literacy* (American Association for the Advancement of Science [AAAS], 2001, 2007) suggest that such relationships are common and may become the object of study in LP research. This phenomenon occurs in the CNA curriculum. Therefore, we can use our experience to illustrate the broader challenge of modeling the interrelationships among LPs. The relationships we build in Bayes nets can, when schematized, be starting points for future researchers who tackle LP modeling challenges that resemble ours.

In learning computer network skills, the student goes beyond understanding isolated concepts to a synthesis of related KSAs. Figure 9 is a graphical representation of the KSAs required for computer networking in the CNA curriculum. It was created from discussions with subject matter experts and instructors in the curriculum. This is not a Bayes net. Rather, it is a kind of concept map that is similar to the maps in the *Atlas of Science Literacy* (AAAS, 2001, 2007). The map is one source of information we use in building LPs in the CNA domain and in building the Bayes nets for modeling them. The map suggests that a student's capability is directly related to some KSAs that are specific to particular networking devices and to other KSAs that are more general. For example, in Figure 9, IP_Addressing depends on Basic Math Skills and Understanding Local & Remote Network Concepts.

In a model of a domain consisting of multiple LPs, the structure and strength of the relationships among different LPs can be incorporated in a Bayes network. These relationships can be straightforward, such as when two LPs are correlated or when mastery of one LP is a prerequisite for mastery of a more advanced LP. Other relationships can be more complicated. For example, exploratory analysis in the CNA curriculum suggests that to master certain levels of the IP_Protocol_Rules progression, learners must be at a certain level of understanding in the IP_Addressing progression. It can be challenging to determine how to model the relationships between the LPs. While there are methods to learn the structure of a Bayesian network just from data, it is often useful to hypothesize the structure first and then use data to verify or to revise this model.

Using Bayes nets to model the hypothesized structure of multiple LPs, we structure the joint distribution among a set of LPs by constructing relationships among latent variables in a multivariate system. As previously discussed, under a Bayes net approach, each LP is represented as a discrete latent variable (node) with categories corresponding to different levels of KSAs in the LP. In the graphical representation, directed edges connect latent variables according to a model structure suggested by subject matter experts or exploratory analyses; for example, Figure 10 indicates that there is a dependence between IP_Addressing and IP_Protocol_Rules as discussed above. That is, the arrow from IP_Addressing to IP_Protocol_Rules indicates that the probabilities of the levels of the IP_Protocol_Rules SMV are different, depending on the level of IP_Addressing.

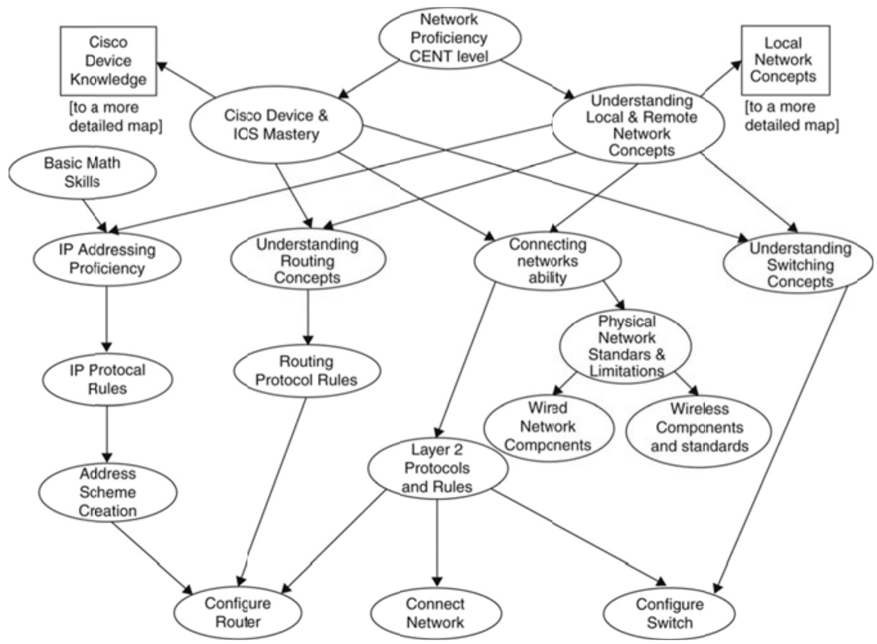


Figure 9. Concept map of the CNA curriculum. This model displays the relationship between different networking KSAs of which IP Addressing Proficiency is a part. A Bayes Net could be constructed with a student model variable corresponding to each node. From A Bayesian Network Approach to Modeling Learning Progressions and Task Performances (CRESST Report 776) (p. 23), by P. West, D. W. Rutstein, R. J. Mislevy, J. Liu, Y. Choi, R. Levy, A. Crawford, K. E. DiCerbo, K. Chappel, and J. T. Behrens, 2010, Los Angeles, CA: University of California, CRESST. Copyright 2010 The Regents of the University of California. Reprinted with permission of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

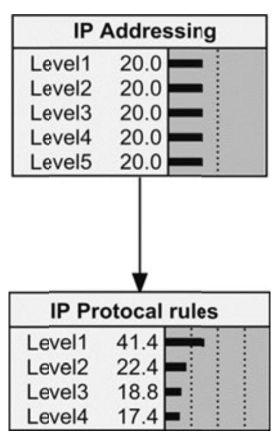


Figure 10. Graphical representation indicating that there is a relationship between two LPs. The nature of the relationship is specified by the conditional probability matrix in Table 3.

The edge in this graph only indicates that there is a relationship, not its nature or strength. This information is contained in the conditional probabilities. Table 3 shows one possible relationship. Reading conditional probability distributions across rows, we see that if a student is at Level 1 or Level 2 of IP_Addressing, there is a high probability the student will be at Level 1 of IP_Protocol_Rules. However, a student at Level 3, Level 4, or Level 5 of IP_Addressing has very similar probabilities of being at any level of IP_Protocol_Rules. The interpretation of this structure is that students at Level 1 or Level 2 of IP_Addressing are usually at Level 1 of IP_Protocol_Rules. For students at or above Level 3 in IP_Addressing, there is only a mild positive association between the two variables.

Table 3. A Conditional Probability Table with a Dependency Estimated from Data.

	IP_Protocol_Rules			
IP_Addressing	*Level 1*	*Level 2*	*Level 3*	*Level 4*
Level 1	0.72	0.14	0.08	0.06
Level 2	0.59	0.21	0.11	0.09
Level 3	0.32	0.28	0.22	0.18
Level 4	0.23	0.25	0.26	0.26
Level 5	0.21	0.24	0.27	0.28

Conditional probabilities can be determined based on data alone or in conjunction with constraints suggested by subject matter experts. Table 4 is very similar to Table 3, but Table 4 has theory-based constraints. The zeros in Table 4 imply that a person who is at Level 1 or Level 2 on IP_Addressing cannot be at a high level of IP_Protocol_Rules. That is, Level 3 of the IP_Addressing LP is a prerequisite for being at Level 3 of the IP_Protocol_Rules LP. Such a structure could be suggested by the substantive relationship between the KSAs at the levels of the two LPs. With data, a statistical test could be applied to test whether constraining the probabilities at the upper right of Table 4 to zero provides acceptable fit.

Table 4. A Conditional Probability Table with Constraints on Conditional Probabilities that Affect a Prerequisite Relationship.

	IP Protocol Rules			
IP_Addressing	*Level 1*	*Level 2*	*Level 3*	*Level 4*
Level 1	0.8	0.2	0*	0*
Level 2	0.7	0.3	0*	0*
Level 3	0.25	0.25	0.25	0.25
Level 4	0.25	0.25	0.25	0.25
Level 5	0.25	0.25	0.25	0.25

* Constrained value.

Multiple Time Points

Another challenge related to the Bayes net implementation of LPs is modeling change over time. As mentioned above, LPs can be characterized as measurable pathways that students may follow in building their knowledge and in gaining expertise over time. The Bayes nets we have discussed in this chapter have only addressed student status at a single point in time. Dynamic Bayes nets can be used to model student LPs over multiple time points where a student's level may change from one set of observations to another.[1] At each time point there are (a) one or more SMVs representing the LP(s) and (b) OVs with probabilistic dependence upon the SMVs. In addition, there is a copy of the SMVs for each time point. We model the relationship between unobservable LPs over time with conditional probability distributions that reflect transition probabilities. Transition probabilities indicate the probability of moving from a particular level at one measurement occasion to the other levels at the next measurement occasion.

Figure 11 shows an example of modeling LPs with a dynamic Bayes net. The Bayes net contains two parts: (1) four SMVs, which are actually the same LP but assessed at four successive time points where each measurement occasion modeled is dependent on the previous one, and (2) four OVs at each time point that are dependent on the SMV for that time point. Different patterns of transition matrices can be considered that depend on the developmental theory that grounds the LPs and on the students' experiences between measurement occasions. For example, the effectiveness of instructional treatments can be compared in terms of the transition probabilities they produce. Figure 11 depicts a situation in which observations have been made at all four time points. At each occasion, the results of four tasks were observed. This student was most likely at Level 1 of the SMV on the first occasion, at Level 2 on the second occasion, at Level 3 on the third occasion, and at Level 4 on the fourth occasion.

Complex Tasks

In the examples discussed thus far, each observable variable depends on only one SMV (i.e., LP). More complex tasks, however, may require jointly employing the KSAs that are modeled to reflect levels in more than one LP. Conducting an investigation in Mendelian inheritance, for example, may require KSAs from both a LP for the concepts in Mendelian genetics and the skills in a LP for proficiency in scientific inquiry.

In computer networking, students solving real-world network design and troubleshooting problems often encounter tasks that require them to draw upon multiple KSAs. Figure 9 suggests that assessing a student's capabilities in configuring a router involves the student's understanding of IP addressing and router concepts plus the student's ability to connect networks. While tasks can be defined to measure just one skill (and most of the multiple choice questions in end-of-chapter tests are so designed), in order to determine whether students can solve problems in real-world environments we must design tasks that require KSAs from multiple LPs.

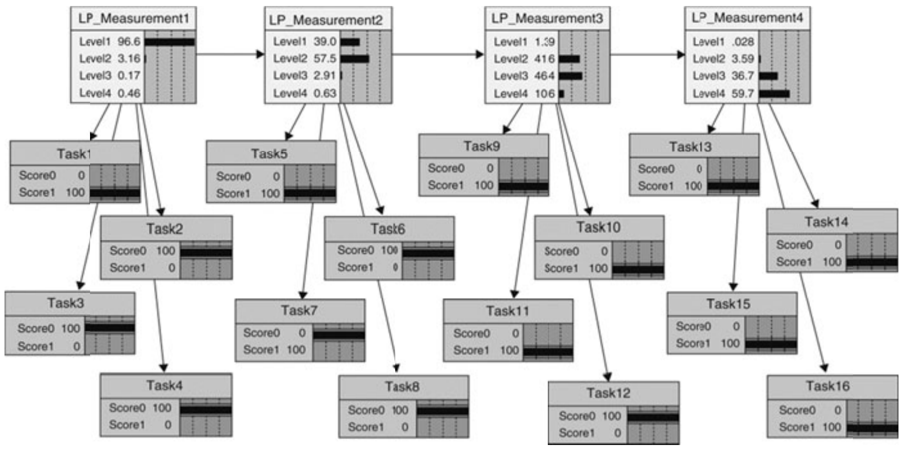


Figure 11. A dynamic Bayes net.

Cisco has developed a simulation environment for creating complex assessment tasks using a tool called Packet Tracer (Frezzo, Behrens, & Mislevy, 2009). Students interact with these tasks as they would with a real network environment. As in a real troubleshooting or network design problem, students must draw on KSAs from multiple LPs to complete these tasks. These relationships are incorporated in a Bayes net by having multiple SMV parents of an observable variable. Figure 12 shows a hypothetical example. Some tasks depend only on capabilities in the Connectivity LP, some depend only on capabilities in the IP_Addressing LP, and others depend on capabilities in both. Table 5 shows the conditional probability table for an observable variable for such a task, identified as ConAddTask1. Each row gives the probabilities of a right and wrong response (Score1 or Score0) assuming that a student is at a given combination of values for the SMV parents, Connectivity and IP_Addressing. The probabilities represent students' status on two LPs, namely, Connectivity with three levels and IP_Addressing with five. By choosing task features and performance requirements in accordance with the definitions of the LPs, the author of ConAddTask1 designed the task so that, to be likely to succeed, students need to be at Level 2 or higher on Connectivity and at Level 4 or higher on IP_Addressing. The pattern of conditional probabilities reflects this structure.

The challenges in modeling complex performance tasks involve determining not only how the relationships from the model are to be structured but also what type of observations are most appropriate for incorporation as evidence in the model. In Packet Tracer, the students' log files include a record of all commands that they have performed as well as a final network configuration that would resolve the problem in a troubleshooting task or meet the client's requirements in a configuration task. Currently the final configuration is evaluated by comparing it to

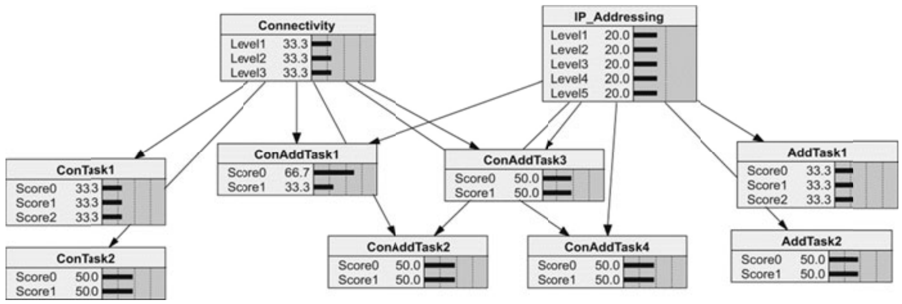


Figure 12. A graphical representation of a task dependent on multiple learning progressions. Note that some tasks depend strictly on Connectivity (ConTasks), some on IP_Addressing (AddTasks), and some on both (ConAddTasks). Adapted from A Bayesian Network Approach to Modeling Learning Progressions and Task Performances (CRESST Report 776) (p. 13), by P. West, D. W. Rutstein, R. J. Mislevy, J. Liu, Y. Choi, R. Levy, A. Crawford, K. E. DiCerbo, K. Chappel, and J. T. Behrens, 2010, Los Angeles, CA: University of California, CRESST. Copyright 2010 The Regents of the University of California. Reprinted with permission of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

a target configuration on the quality and features of a number of network aspects. Additional information is being collected, however, that can be used for evaluation. Earlier research with the NetPASS prototype (Williamson et al., 2004) used sequences and timing of actions recorded in the log files as evidence of student capabilities. We are currently studying whether other types of evidence can be used operationally.

Grain-size of a LP

The grain-size of a LP refers to how broadly or finely focused the progression of the described KSAs is. There are tradeoffs between modeling LPs at coarser or finer levels of detail; the choice depends on the purpose of the research or the assessment project at hand. Fine-grained LPs may be appropriate for studying changes in a narrowly defined concept, but a more broadly cast LP might be preferable for summarizing students' general levels of reasoning in a large-scale educational survey. The LPs we have worked with in CNA are built to help instructional designers, task developers, and teachers coordinate their work to define the curriculum and guide instruction at the level of lessons and exercises. The IP_Addressing LP shown in Table 1, for example, is at a medium grain-size. Each level is associated with a set of KSAs that are related to IP_Addressing. Students operating at a given level in this LP are likely to display most of the KSAs at lower levels and few of the KSAs at higher levels. However, more finely-grained LPs that focus more closely on one KSA in the set are possible—that is, IP_Addressing may be broken into a number of finer-grained LPs. On the other hand, a coarser LP may be defined by combining IP_Addressing with, say, Connectivity, Routing, and Switching for a broader description of students' KSAs. In this example, the grain-size was chosen for reasons related to subject matter and data.

Table 5. Conditional Probability Table for the Observable Variable ConAddTask1, Which Has Two SMV Parents (IP_Addressing and Connectivity).

Connectivity	*IP_Addressing*	*ConAddTask1*	
		Score 0	*Score 1*
Level 1	Level 1	90	10
Level 1	Level 2	90	10
Level 1	Level 3	90	10
Level 1	Level 4	90	10
Level 1	Level 5	90	10
Level 2	Level 1	90	10
Level 2	Level 2	90	10
Level 2	Level 3	90	10
Level 2	Level 4	20	80
Level 2	Level 5	20	80
Level 3	Level 1	90	10
Level 3	Level 2	90	10
Level 3	Level 3	90	10
Level 3	Level 4	20	80
Level 3	Level 5	20	80

In terms of subject matter, the groupings reflected in the IP_Addressing LP are based on clusters of related concepts that are taught and practiced together as variations on a "key idea" that is addressed in instruction and built on in subsequent levels. Two data-driven lines of reasoning influenced our choice of grain-size: (1) analyses of existing test data and subsequent identification of patterns of stability in that data and (2) variation in performance across two different organizations of the CNA curriculum.

Analysis of end-of-chapter test data revealed items with similar difficulties in terms of statistics and clustering of students in accordance with latent classes that represented those who "got the idea" and those who did not—usually one central concept, sometimes two, in a chapter. We conducted exploratory analyses using unconstrained latent class models (see Haertel, 1989) to identify structures that may suggest portions of LPs. These exploratory analyses and additional latent class analyses revealed dependencies across chapters that reflect curriculum developers' beliefs that certain concepts build on others. Tracking these dependencies revealed linear progressions of concepts across chapters that formed a LP, such as IP_Addressing. There was instructional value in defining a LP at this grain-size because the central theme in a given LP level (as discussed in connection with the IP_Addressing example) could account empirically for a cluster of related KSAs addressed in the chapter and the associated learning exercises. We also found cases in which knowledge at a given level of one LP was necessary for advancing to a given level of a different LP.

We gained further insights by comparing results across different presentations of material. Different classes present information in different sequences. Our analyses are still underway, but it appears that the patterns of performance in different courses can be understood in terms of the different orders in which the LP levels are addressed. In other words, modeling at a coarser grain-size would produce a very "messy middle" because after, say, two courses, students would have very different performance profiles. Modeling at the medium grain-size allows us to understand the middle in terms of different profiles across the same set of LPs.

The "Messy Middle"

Another challenge related to the definition of LPs may arise when modeling middle levels of proficiency. It is typically easiest to define the endpoints of a LP, where the lowest level refers to a novice state and the highest level refers to an expert state. In the simple LP described in the beginning of this chapter, learning generally proceeds as the successive attainment of KSAs in a single order, as shown in Table 1. In such cases, it is possible to define a LP in terms of ordered levels of a single SMV (as shown in Figure 13a).

It is more challenging to model the intermediate levels in the LP, however, when there are multiple pathways a student may follow in acquiring the KSAs associated with the various levels. To illustrate some of these possibilities, Figure 13b depicts alternative structures for the sequencing of KSA acquisition. In each sequence, the nodes represent different KSAs associated with the LP. Note that KSA 1 and KSA 5, at the beginning and the end of the sequences, represent the lowest and highest endpoints of the LP, respectively. The sequence on the left represents the acquisition of KSAs 1–5 in a particular order: students acquire KSA 1, followed by KSA 2, followed by KSA 3, followed by KSA 4, and finally KSA 5. The sequence in the middle offers a similar structure in which KSAs are acquired in an ordered fashion although the order differs. The sequence on the right depicts a different structure in which students acquire KSA 1 and then KSA 2. They then may acquire either KSA 3 or KSA 4, both of which must be acquired before KSA 5.

As Figure 13 illustrates, numerous patterns of KSA acquisition are possible. It is often unclear which sequence or pattern holds, or, as may be possible, if students experience different sequences of KSA acquisition. The difficulty in defining a single sequence that applies to all students, or of enumerating all the sequences that students experience—to say nothing of identifying which sequence students progress along—is what is referred to as the "messy middle" (Gotwals & Songer, 2010, p. 277). Approaches for modeling multiple sequences in the "messy middle" can be found in the psychometric literature on diagnostic and classification models (e.g., Haertel & Wiley, 1993; Leighton & Gierl, 2007; Rupp & Templin, 2008; Tatsuoka, 2002). These approaches can be expressed in Bayes net structures by extending the ideas discussed in the previous section.

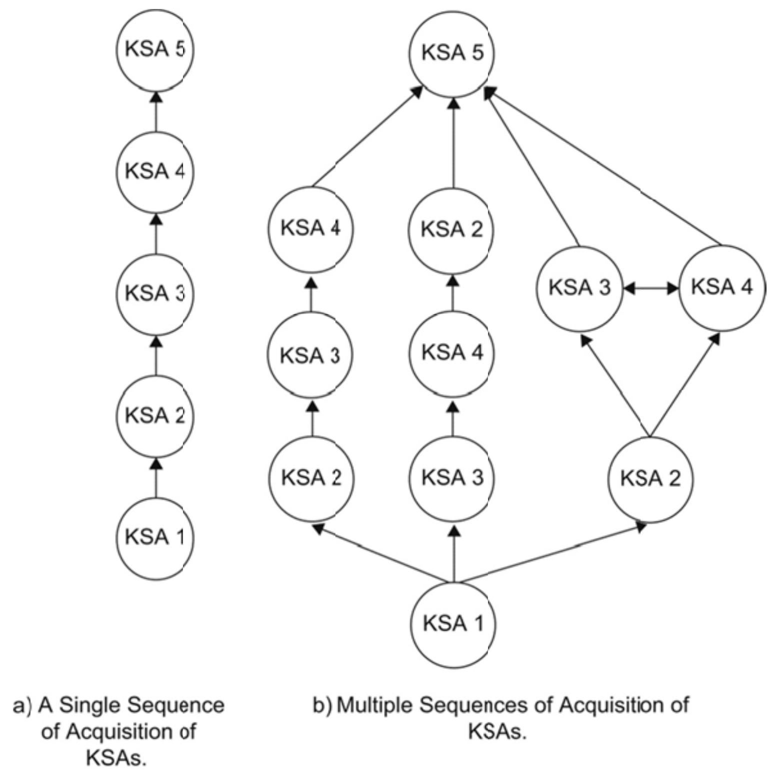


Figure 13. Single and multiple sequences of acquisition of KSAs.

In particular, a Bayes net for the multiple path scenario can be written in terms of multiple nodes. Figure 14 shows a Bayes net structure that accommodates all the paths in Figure 13b, using SMVs with two values for each KSA. This structure posits that, in students' development, the acquisition of KSA 1 tends to occur first and the acquisition of KSA 5 last. However, different orderings of KSAs 2, 3, and 4 may occur along the way. The conditional probability matrices corresponding to the arrows indicate if paths are common. If a single ordering is required, say, for reporting purposes, it is possible that the state with KSA 1 only is at a low level of a LP; any state with KSA 1 and KSA 2, 3, or 4 is at a second level; any state with KSA 1 and two of KSA 2, 3, and 4 is at a third level; a state with KSAs 1–4 is at a fourth level; and a state with five KSAs is at the highest level. According to this definition, students at the same broad level of the LP could have different profiles of probabilities for tasks, depending on their states within levels.

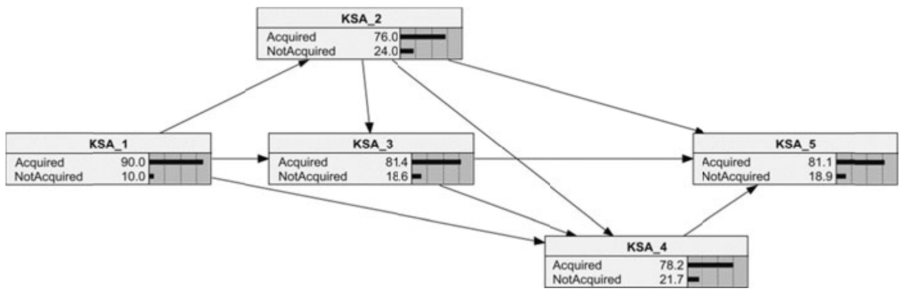


Figure 14. Bayes net for LP with multiple sequences of acquisition of KSAs. A student could progress from KSA_1 to KSA_2, 3, or 4 and then through any of the various paths indicated by the arrows.

DISCUSSION

Benefits of Using Bayes Nets to Model LPs

Why use Bayes nets to model LPs? One strength of the Bayes net approach is that it builds on a structure that can be based on theory and expert opinion; initially, conditional probabilities can be based on expert opinion and fragmentary data. This means a network can be used immediately for low-stakes inferences. As data accumulate, however, the probability-based reasoning foundations allow for coherent improvement of the estimates and more powerful investigations of model fit and subsequent model improvement. A particular advantage of Bayes nets is the flexibility in modeling complex and often subtle evidentiary relationships (Schum, 1994) among LP levels and between LPs and task performances. This is especially important if multiple interrelated progressions are considered and if KSAs from multiple progressions are involved in complex tasks or multiple aspects of a performance are evaluated (Mislevy, 1994; Mislevy & Gitomer, 1996). Estimates for students' LP levels can be updated in real time. Fragments of Bayes nets can be pre-built and assembled on the fly in response to an evolving situation, which is important in adaptive testing and in games- and simulation-based assessment (Almond & Mislevy, 1999; Conati, Gertner, & VanLehn, 2002; Martin & VanLehn, 1993; Shute, Ventura, Bauer, & Zapata-Rivera, 2009).

Although the work of Bayes nets is focused on modeling student performance in light of LPs, Bayes nets also make LPs more useful in some practical and important ways. Using Bayes nets to model LPs can lead to efficient and valid task design. The process of identifying initial LPs focuses test developers on the theory of cognition in the domain and defines the characteristics of individuals at various levels of the LP. Bayes nets confirm these levels and progressions, allowing designers to specify the levels of KSAs at which they aim assessment items. For example, it is possible to state that a task is designed to target Level 3 of one progression and requires knowledge at Level 2 of another progression. This helps make task design more principled, more planful, and ultimately more valid. Bayes nets also help connect curriculum to assessment. For example, curriculum designers can use information from a Bayes net structure to make decisions about which content areas to emphasize so that students have a greater probability of mastering future KSAs (DiCerbo & Behrens, 2008).

An area we continue to explore is how Bayes nets can provide feedback to students and instructors (DiCerbo, 2009). Such feedback could be achieved in at least two ways. First, students could receive reports that update their estimated levels on various KSAs, given their assessment performance. Based on these reports, students could be directed to other activities. This is the idea behind intelligent tutors or, when wrapped in a "fun" scenario, behind games (Shute et al., 2009). Teachers can use the structure of Bayes nets the same way that the curriculum designers (mentioned above) do when making decisions about content emphasis. In addition, teachers can diagnose student problems. For example, if a student is struggling in one area, teachers can look backwards to a network of variables to see what prerequisite KSAs the student probably lacks.

Challenges for the Community

Model-building is, by nature, iterative. Progressions are hypothesized, and models are built, based on understandings of the substantive area and data at a given point in time. As discussed above, building a Bayes net for a particular application involves encoding the relationships and making hypotheses of interest from the domain in the Bayes net. The Bayes net is fit to the data, and data-model fit and related model-checking tools are used to identify strengths and weaknesses of the Bayes net in terms of overall features, subsets of variables, or subsets of examinees (Levy, 2006; Sinharay, 2006; Sinharay & Almond, 2007; Williamson, Mislevy, & Almond, 2001). The results of these analyses have several interpretations. In a statistical sense, adequate data-model fit indicates that the probabilistic relationships in the Bayes net account for what actually takes place in terms of the data at hand, whereas data-model misfit indicates the relationships in the Bayes net do not reflect what actually takes place. More substantively, because the Bayes net is explicitly built to reflect domain-specific hypotheses, adequate data-model fit constitutes support for those hypotheses, whereas data-model misfit constitutes evidence against the hypotheses. Data-model misfit might indicate that some approximations or choices made in the model are not precise enough, or that certain relationships are poorly understood, or that the hypotheses and relationships hold for certain students but not for others.

As noted above, Bayes nets are flexible statistical models applicable to a wide variety of problems. Assessment and assessment in the context of LPs constitute just a few of these problems. The unique features of assessment of LPs, however, dictate that certain recurring features of the model are likely present in applications of Bayes nets to LP assessments. At present, the development of Bayes nets to accommodate these aspects is in its relative infancy. Similarly, related aspects of modeling need to be tuned to the particular features of assessment in the context of

LPs. For example, efficient data-model fit procedures to evaluate hypotheses about sequences of KSA acquisition through the "messy middle" need to be developed.

As discussed above, new challenges arise in every application in which a researcher models a LP in a particular substantive area. A comprehensive approach to assessment for LPs develops Bayes nets in concert with the specification of the desired inferences and tasks. This process, which is often iterative, is always localized to the specific situation as defined by the purpose of the assessment. Any serious application of Bayes nets involves the interplay between the methodological tools of Bayes nets and the substantive expertise required to build appropriate model approximations for the domain.

NOTE

[1] Because the LP variables are unobservable, the resulting Bayes net is formally a hidden Markov model (Cappé, Moulines, & Rydén, 2005; Langeheine &Van de Pol, 2002).

REFERENCES

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223–237.

American Association for the Advancement of Science. (2001, 2007). *Atlas of science literacy* (Vols. 1–2). Washington, DC: Author.

Behrens, J. T., Collison, T. A., & DeMark, S. F. (2005) The seven Cs of comprehensive assessment: Lessons learned from 40 million classroom exams in the Cisco Networking Academy program. In S. Howell & M. Hricko (Eds.), *Online assessment and measurement: Case studies in higher education, K-12 and corporate* (pp. 229–245). Hershey, PA: Information Science Publishing.

Behrens, J. T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2007). Structural, functional and semiotic symmetries in simulation-based games and assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59–80). New York, NY: Routledge.

Cappé, O., Moulines, E., & Rydén, T. (2005). *Inference in hidden Markov models*. New York, NY: Springer.

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User modeling and user-adapted interaction, 12*, 371–417.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Dayton, C. M., & Macready, G. B. (2007). Latent class analysis in psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 421–446). Cambridge, MA: Elsevier.

DeMark, S. F., & Behrens, J. T. (2004). Using statistical natural language processing for understanding complex responses to free-response tasks. *International Journal of Testing, 4*, 371–390.

DiCerbo, K. E. (2009, April). *Communicating with instructors about complex data analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

DiCerbo, K. E., & Behrens, J. T. (2008, March). *Identifying keystone concepts: Big ideas in computer networking*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist, 53,* 416–428.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2009). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *The Journal of Science Education and Technology, 19*, 105–114.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London, UK: Chapman & Hall.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement test items. *Journal of Educational Measurement, 26,* 301–321.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale, NJ: Erlbaum.

Ingham, A. M., & Gilbert, J. K. (1991). The use of analogue models by students of chemistry at higher education level. *International Journal of Science Education, 13*, 193–202.

Jensen, F. V. (1996). *Introduction to Bayesian networks*. New York, NY: Springer.

Kadane, J. B., & Schum, D. A. (1996). *A probabilistic analysis of the Sacco and Vanzetti evidence.* New York, NY: Wiley.

Langeheine, R., & Van de Pol, F. (2002). Latent Markov chains. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 304–341). Cambridge, UK: Cambridge University Press.

Lauritzen, S., & Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, 50*, 157–224.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks.* (Doctoral dissertation, University of Maryland). Retrieved from http://drum.lib.umd.edu/bitstream/1903/3483/1/umi-umd-3311.pdf

Levy, R., & Mislevy, R. J. (2004). *Specifying and refining a measurement model for a simulation-based assessment* (CSE Technical Report 619). Retrieved from University of California Los Angeles, National Center for Research on Evaluation, Standards, & Student Testing website: http://www.cse.ucla.edu/products/reports/r619.pdf

Levy, F., & Murnane, R. J. (2004). *The new division of labor: How computers are creating the next job market*. Princeton, NJ: Princeton University Press.

Linzer, D. A., & Lewis, J. (2007). *poLCA: Polytomous variable latent class analysis*. *R package version 1.1*. Retrieved from http://userwww.service.emory.edu/~dlinzer/poLCA

Martin, J. D., & VanLehn, K. (1993). OLEA: Progress toward a multi-activity, Bayesian student modeler. In S. P. Brna, S. Ohlsson, & H. Pain (Eds.), *Artificial intelligence in education: Proceedings of AI-ED 93* (pp. 410–417). Charlottesville, VA: Association for the Advancement of Computing in Education.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence* (pp. 437–446). San Francisco, CA: Morgan Kaufmann.

Mislevy, R. J., & Gitomer, D. (1996).The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253–282.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Murnane, R. J., Sharkey, N. S., & Levy, F. (2004). A role for the internet in American education? Lessons from the Cisco Networking Academies. In P. A. Graham & N. G. Stacey (Eds.), *The* *knowledge economy and postsecondary education: Report of a workshop*. (pp. 127–158). Washington, D.C.: National Academy Press.

Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus User's Guide* (4th ed.). Los Angeles, CA: Author.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.

Norsys Software Corp. (2007). *Netica manual*. Retrieved from http://www.norsys.com

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*: *A Multidisciplinary Journal, 14,* 535–569.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Kaufmann.

R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*, 219–262.

Schum, D.A. (1994). *The evidential foundations of probabilistic reasoning.* New York, NY: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Philadelphia, PA: Routledge.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*, 1–34.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement, 67,* 239–257.

Suarez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71*, S767–779.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*, 449–508.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Applied Statistics, 51,* 337–350.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., . . . Behrens, J. T. (2010, August). *A Bayesian network approach to modeling learning progressions.* (CRESST Research Report 776). Retrieved from University of California Los Angeles, National Center for Research on Evaluation, Standards, & Student Testing website: http://www.cse.ucla.edu/products/reports/R776.pdf

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. (2004). Design rationale for a complex performance assessment. *International Journal of Measurement, 4,* 303–332.

Williamson, D. M., Mislevy, R. J., & Almond, R. A. (2001, April). *Model criticism of Bayesian networks with latent variables*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

APPENDIX

An Example of Building and Fitting a Bayesian Network for a Learning Progression

This example demonstrates how assessment data can be used to help validate a learning progression (LP) using statistical modeling in the form of a Bayes net. The data are from scored responses to 35 items written to target specified levels of the IP_Addressing progression. The hypothesized structure is a Bayes net with these 35 items as conditionally independent observable variables, dependent on a single discrete latent variable with values that indicate LP levels.

Owing to the connection between a Bayes net of this description and latent class analysis (Box 2), a series of latent class analyses were conducted using the poLCA package (Linzer & Lewis, 2007) in R (R Development Core Team, 2008). These were exploratory analyses that did not constrain the solution to finding the theoretical levels that motivated the item writers or to yielding conditional probabilities that reflected jumps that accorded to levels. Rather, they were unconstrained latent class analyses for 2-Class, 3-Class, 4-Class, 5-Class, and 6-Class solutions. Furthermore, it was not required that classes be ordered or that the conditional probability matrices for items would show jumps at levels targeted by the item writers. The structure that emerged would be driven by patterns in the data. As the ensuing discussion shows, the latent class structure that emerged empirically closely reflected the theoretical structure of LP levels and conditional probabilities for items with jumps at intended levels.

The 4-Class model demonstrated the best fit to the data, based on statistical fit in terms of the BIC (Schwarz, 1978) and the bootstrapped likelihood ratio test (McLachlan & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007) conducted in M*plus* (Muthén & Muthén, 1998–2006). In addition, this model offered the best interpretability of the classes in terms of class membership proportions and consistently ordered patterns of class performance across items. The four classes identified in the analysis corresponded to increasing levels of performance on the items and were interpretable as increasing levels of KSAs. A hypothesized further distinction at the high end of the LP was not realized due to the small number of items targeted at this level. In other words, there was insufficient information in the data set to differentiate students at the two highest theorized levels. A Bayes net representation of a model with a single SMV containing four levels (classes) was then constructed in Netica (Norsys Software Corp., 2007), represented in Figure A1.

INFERENCES REGARDING ASSESSMENT ITEMS

An item was classified as "at the level" of a certain class if it supported an interpretation that students at that level would be able to solve or complete the task,

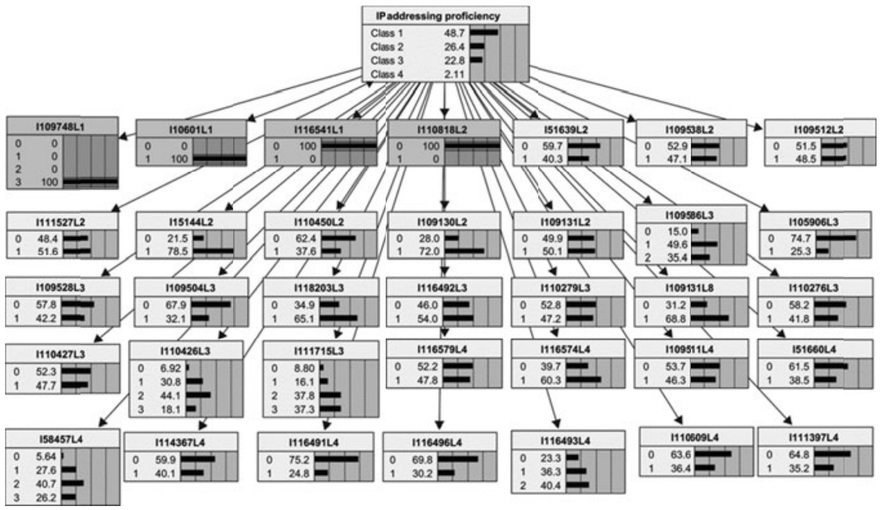


Figure A1. Graphical representation of the Bayes net. Adapted from A Bayesian Network Approach to Modeling Learning Progressions and Task Performances (CRESST Report 776) (p. 17), by P. West, D. W. Rutstein, R. J. Mislevy, J. Liu, Y. Choi, R. Levy, A. Crawford, K. E. DiCerbo, K. Chappel, and J. T. Behrens, 2010, Los Angeles, CA: University of California, CRESST. Copyright 2010 The Regents of the University of California. Reprinted with permission of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

whereas students at lower levels would likely be unsuccessful. To classify items, the conditional probability tables were examined. For each item, the odds of answering the item completely correctly were calculated in each class, and odds ratios were calculated to compare adjacent classes. These odds ratios capture the power of the items to discriminate between classes. To construct an odds ratio for the first class, the probability that a complete novice (our interpretation of the lowest class emerging from the analysis—a class with low conditional probabilities of correct responses to all items) would get the item right was defined as the probability of getting the item right by guessing. Each item was assigned to a level based on considerations of (a) the size of these odds ratio between successive classes, (b) the criterion that the probability of responding correctly at the assigned level exceeded .50 for dichotomously scored items, and (c) the distribution of probability across the response categories for polytomously scored items.

The results indicated that many items discriminated strongly between classes. For example, Figure A2 contains the conditional probability table for an item that only students in the fourth (highest) class are likely to solve successfully. Statistically, this item aids in distinguishing students in the fourth latent class from those in the remaining classes. Substantively, the item exemplifies one aspect of what it means to be at the fourth level of the LP. Students at the fourth level have acquired the KSAs necessary to have a strong chance of correctly answering this item; students at the lower levels have not.

IP_Addressing_Proficiency	Score0	Score1
Class1	84.200	15.800
Class2	74.420	25.580
Class3	62.440	37.560
Class4	12.480	87.520

Figure A2. Conditional probability table for an item that discriminates well between the fourth and the other classes. From A Bayesian Network Approach to Modeling Learning Progressions and Task Performances (CRESST Report 776) (p. 19), by P. West, D. W. Rutstein, R. J. Mislevy, J. Liu, Y. Choi, R. Levy, A. Crawford, K. E. DiCerbo, K. Chappel, and J. T. Behrens, 2010, Los Angeles, CA: University of California, CRESST. Copyright 2010 The Regents of the University of California. Reprinted with permission of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Other items were more ambiguous in terms of their levels. For example, Figure A3 shows the conditional probability table for an item on which students in the second class have a .88 probability of earning partial or full credit but only a .52 probability of earning full credit, whereas students in the third class have a .86 probability of earning full credit. A simple classification of this item in terms of one level is insufficient to fully capture its connection to the classes. A richer characterization of the item, recognizing that it discriminates well between multiple adjacent classes, states that once a student reaches class two, she is very likely to earn at least partial credit but needs to reach class three (or four) in order to be as likely to earn full credit.

Across all items, the results were largely consistent with the experts' expectations. Ten items exhibited clear and distinct patterns that distinguished between classes exactly as predicted by experts. That is, these items were "located" at the expected level. Figure A2 shows an example of one such item; the expert prediction of this item as a level 4 item is strongly supported by the empirical results.

IP_Addressing_Proficiency	Score0	Score1	Score2
Class1	40.300	49.000	10.700
Class2	11.870	36.280	51.850
Class3	2.350	12.050	85.600
Class4	0.000	5.110	94.890

Figure A3. Conditional probability table for an item that discriminates differentially at multiple points. From A Bayesian Network Approach to Modeling Learning Progressions and Task Performances (CRESST Report 776) (p. 19), by P. West, D. W. Rutstein, R. J. Mislevy, J. Liu, Y. Choi, R. Levy, A. Crawford, K. E. DiCerbo, K. Chappel, and J. T. Behrens, 2010, Los Angeles, CA: University of California, CRESST. Copyright 2010 The Regents of the University of California. Reprinted with permission of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Five items were scored polytomously; these distinguished roughly well at the level predicted by experts. This is seen in terms of differential probabilities between the targeted LP levels at one score level and the two other LP levels at another score level. This phenomenon is illustrated for the item whose conditional probabilities appear in Figure A3. This item was expected to be a level 4 item. This item is located at class 2 with respect to being able to obtain a score of 1 as opposed to 0; it is also is located at class 3 in terms of being able to obtain a score of 2 as opposed to 1.

Overall, eighteen items were located at a level adjacent to the predicted level (e.g., an item expected at level 4 was located at class 3). One item was located adjacent to the predicted class and was also located at another class not adjacent. Only one item was clearly located at a class that was not equal to or adjacent to the predicted level. Initial reviews of these results indicated revisions that would help the items more sharply target the concepts at their intended levels.

INFERENCES REGARDING STUDENTS

The conditional probability tables also reveal how inferences regarding students are conducted in the Bayes net. For example, observing a correct response for the item in Figure A2 is strong evidence that the student is in class 4; observing an incorrect response for the item in Figure A2 is relatively strong evidence that the student is not in class 4. The use of a Bayes net approach supports inferences regarding students by collecting and synthesizing the evidence in the form of observed values of variables. That information is then propagated through the network via algorithms based on Bayes' theorem to yield posterior distributions for the remaining unknown variables (Pearl, 1988), including the SMV corresponding to the LP. For example, Figure A1 contains the Bayes net for a student who has completed four items. The student correctly answered the first two items and incorrectly answered the next two items. On the basis of this evidence, the posterior distribution for his/her latent skill variable indicates that this student has a probability of being in classes 1–4 of .487, .264, .228, and .021, respectively. On this basis we may infer that the student is almost certainly in one of the first three classes (i.e., is at one of the first three levels of the progression) and is more likely in the first class than either the second or third. Yet there still remains considerable uncertainty. The collection and inclusion of more data would lead to a more refined inference.

COMMENT ON THE EXAMPLE

The results of the modeling offer a data-based interpretation of the development of KSAs that constitute the LP. In some cases, the results for items confirm the experts' expectations. For other items, the results are more ambiguous or offer an alternative to the experts' expectations. To take a more comprehensive perspective on assessment of LPs, the results of the statistical analyses will be submitted to the subject matter experts for consultation and possible refinements in terms of the definition of the LP, the items that assess the aspects of the LP, and the utility of additional items for modeling students' progression.

DEREK C. BRIGGS AND ALICIA C. ALONZO

THE PSYCHOMETRIC MODELING OF ORDERED MULTIPLE-CHOICE ITEM RESPONSES FOR DIAGNOSTIC ASSESSMENT WITH A LEARNING PROGRESSION

An appealing feature of learning progressions is their potential to facilitate diagnostic assessment of student understanding. In this context, diagnostic assessment hinges upon the development of items (i.e., tasks, problems) that efficiently elicit student conceptions that can be related to a hypothesized learning progression. Briggs, Alonzo, Schwab, and Wilson (2006) introduced Ordered Multiple-Choice (OMC) items for this purpose. OMC items combine the efficiency of traditional multiple-choice items with the qualitative richness of responses to open-ended questions. The potential efficiency results because OMC items offer a constrained set of response options that can be scored objectively; the potential qualitative richness results because OMC response options are designed to correspond to students' answers to open-ended questions and are explicitly linked to a discrete level of an underlying learning progression. The OMC format belongs to a broader class of item formats in which the interest is the diagnosis of students' reasons for choosing incorrect answers (cf. Minstrell, n.d., 1992, 2000). An attractive aspect of OMC items is that they are consistent with the spirit of learning progressions, which, at root, muddy the notion that students either "get something" or they don't.

This chapter describes some challenges inherent in the psychometric modeling of learning progressions, using the context of a specific learning progression and an associated set of OMC items. Formal psychometric modeling is important for learning progression work for two reasons. First, a psychometric model may be used to draw probabilistic inferences about unobserved (i.e., latent) states of student understanding. With such inferences, it is possible to quantify how well a student has mastered the content of a learning progression. Second, the process of specifying a model and evaluating its fit may offer a systematic way to validate and refine a hypothesized learning progression. However, a psychometric model may fail to make diagnostic inferences because it does not support reliable diagnoses or because diagnoses do not correspond to other evidence of student understanding. Then questions arise that learning progression development teams must address. Is there a problem with the assessment instrument? Or should the levels of the hypothesized learning progression be revised? Or has an inappropriate psychometric model been specified?

The rest of the chapter is divided into five sections. First, we briefly describe the development of a relatively simple learning progression and an associated set of

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 293–316.

OMC items. Then we present descriptive statistics from a recent administration of these OMC items to a convenience sample of high school students. In this section we also review some limitations of using classical item statistics to make inferences about student understanding. In the third section we discuss the inherent challenges associated with choosing an approach for modeling student responses when the goal is to make diagnostic inferences about student understanding. We distinguish between approaches based on item response theory models and on diagnostic classification models. In the fourth section we introduce the Attribute Hierarchy Method (AHM; Leighton, Gierl & Hunka, 2004) as a relatively novel approach for modeling OMC items. The AHM is a diagnostic classification model that builds on the seminal work of Tatsuoka (1983, 2009) who developed the Rule Space Method for cognitive assessment. To our knowledge, the AHM has not been applied in the context of a learning progression in science. We illustrate the steps that are necessary to apply the AHM to OMC item responses in order to produce diagnostic student classifications. In the last section we speculate about strengths and weaknesses of the AHM.

BACKGROUND

In previous studies we developed learning progressions in earth science, life science, and physical science (Alonzo & Steedle, 2009; Briggs et al., 2006). In this chapter we use a learning progression that focuses on conceptual understanding of the Earth in the Solar System (ESS) as the context for the modeling discussion that follows. The ESS learning progression describes students' developing understanding of target ideas in earth science. According to national science education standards, students should understand these ideas by the end of eighth grade (American Association for the Advancement of Science [AAAS], 1993; National Research Council [NRC], 1996). However, there is substantial evidence that typical instruction has not succeeded in achieving this goal. In fact, even many college students have misconceptions about these target ideas (e.g., Schneps & Sadler, 1987).

Our initial development of the ESS learning progression followed the same process we used for other learning progressions (for more detail, see Briggs et al., 2006). We began by defining the top level of our learning progression, relying upon national science education standards (AAAS, 1993; NRC, 1996). By the end of eighth grade, students are expected to use accurate models of the relative motion of the Earth and other objects in the Solar System to explain phenomena such as the day/night cycle, the phases of the Moon, and the seasons. We defined lower levels of the learning progression (i.e., novice understanding, intermediate understanding, etc.) using the research literature on student understanding of the targeted ideas (e.g., Atwood & Atwood, 1996; Baxter, 1995; Bisard, Aron, Francek, & Nelson, 1994; Dickinson, Flick, & Lederman, 2000; Furuness & Cohen, 1989; Jones, Lynch, & Reesink, 1987; Kikas, 1998; Klein, 1982; Newman, Morrison, & Torzs, 1993; Roald & Mikalsen, 2001; Sadler, 1987, 1998; Samarapungavan, Vosniadou, & Brewer, 1996; Stahly, Krockover, & Shepardson, 1999; Summers & Mant, 1995;

Targan, 1987; Trumper, 2001; Vosniadou, 1991; Vosniadou & Brewer, 1992; Zeilik, Schau, & Mattern, 1998). In defining these lower levels, we used information about both misconceptions and productive—although naïve—ideas that could provide a basis for further learning. While the target level of understanding at the top of this learning progression is linked to the AAAS and NRC expectations for eighth grade students, the lower levels represent understandings that students are expected to develop in kindergarten through the middle grades.

At this point it is important to note two key limitations of the research available for the construction of this (and most other) learning progressions. Although the aim of learning progressions is to describe how student understanding develops, the research evidence is primarily cross-sectional. While we have important information about the prevalence of particular ideas at different ages, there is little documentation of the progress individual students make through these ideas over time as a result of instruction. In addition, much of the ESS research concerns students' misconceptions, generally focusing on isolated (incorrect) ideas rather than on the relationship between students' ideas (both correct and incorrect). Since the ESS learning progression encompasses multiple phenomena—the Earth orbiting the Sun, the Earth rotating on its axis, and the Moon orbiting the Earth—the definition of levels requires grouping ideas about these phenomena, using both experience and experts' logical reasoning. Thus the ESS learning progression represents a hypothesis both about how students progress towards targeted levels of understanding and about how ideas about different phenomena "hang together." Testing this hypothesis requires further evidence from the iterative process of developing a learning progression and its associated assessment items. The learning progression informs the development of assessment items; these items are used to collect data about student thinking; the data are linked to the initial progression using a psychometric model; and revisions are made to the items and to the learning progression.

The current version of the ESS learning progression is depicted in Figure 1. The science education community is very interested in learning progressions that specify different levels of student knowledge and also describe how students may demonstrate that knowledge. Smith, Wiser, Anderson, and Krajcik (2006) have called for learning progressions that specify "learning performances" (p. 9). In the ESS learning progression, such learning performances are implied: students are expected to use the targeted knowledge to explain or predict phenomena such as the day/night cycle, the phases of the Moon, and the seasons.

Figure 2 shows two examples of OMC items developed to assess students' location on the ESS learning progression. At first glance, the OMC items resemble the typical multiple-choice items found on most traditional exams. There is a difference, however. For OMC items, each response option is intended to represent a qualitatively distinct level of understanding of the ESS learning progression. Although each OMC item has a response option considered the "most" correct, partial credit is given for responses that reflect developing understanding of the phenomenon in the item stem. (For more detail on how these items were developed, see Briggs et al., 2006.) Here we note three features of OMC items that impede diagnostic inference:

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: • the day/night cycle • the phases of the Moon (including the illumination of the Moon by the Sun) • the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that • the Earth is both orbiting the Sun and rotating on its axis • the Earth orbits the Sun once per year • the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky • the Moon orbits the Earth once every 28 days, producing the phases of the Moon COMMON ERROR: Seasons are caused by the changing distance between the Earth and Sun. COMMON ERROR: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the Moon.
3	Student knows that: • the Earth orbits the Sun • the Moon orbits the Earth • the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. COMMON ERROR: It gets dark at night because the Earth goes around the Sun once a day.
2	Student recognizes that: • the Sun appears to move across the sky every day • the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis. COMMON ERROR: The Sun travels around the Earth. COMMON ERROR: It gets dark at night because the Sun goes around the Earth once a day. COMMON ERROR: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, "darkness") covers the Sun. COMMON ERROR: The phases of the Moon are caused by clouds covering the Moon. COMMON ERROR: The Sun goes below the Earth at night.
0	No evidence or off-track

Figure 1. A learning progression for student understanding of Earth in the Solar System. From "Diagnostic Assessment with Ordered Multiple-Choice Items," by D. C. Briggs, A. C. Alonzo, C. Schwab, and M. Wilson, 2006, Educational Assessment, 11, p. 42. Copyright 2006. Reprinted by permission of the publisher (Taylor & Francis Ltd), granted by Copyright Clearance Center Inc. on behalf of Taylor and Francis.

2) Which is the best explanation for why we experience different seasons (winter, summer, etc.) on Earth?

	Level
A. The Earth's orbit around the Sun makes us closer to the Sun in the summer and farther away in the winter.	4
B. The Earth's orbit around the Sun makes us face the Sun in the summer and away from the Sun in the winter.	3
C. The Earth's rotation on its axis makes us face the Sun in the summer and away from the Sun in the winter.	3
D. The Earth's tilt causes the Sun to shine more directly in the summer than in the winter.	5
E. The Earth's tilt makes us closer to the Sun in the summer than in the winter.	4

3) Which best describes the movement of the Earth, Sun, and Moon?

	Level
A. The Sun and Moon both orbit the Earth; the Earth rotates on its axis.	2
B. The Moon orbits the Earth; the Earth orbits the Sun; the Earth rotates on its axis.	4
C. The Moon orbits the Earth; the Earth orbits the Sun.	3
D. The Earth, Sun, and Moon do not move, but other objects in the sky orbit around them.	1
E. The Earth rotates on its axis.	3

Figure 2. OMC items associated with ESS learning progression. Item 2 has been revised from "Diagnostic Assessment with Ordered Multiple-Choice Items," by D. C. Briggs, A. C. Alonzo, C. Schwab, and M. Wilson, 2006, Educational Assessment, 11, p. 43. Copyright 2006. Reprinted by permission of the publisher (Taylor & Francis Ltd), granted by Copyright Clearance Center Inc. on behalf of Taylor and Francis.

1. For some items, it is impossible to write (and consequently, for students to select) a response option at the highest levels of the ESS learning progression. This impossibility creates a ceiling effect at the item level. For example, for item 3 in Figure 2, the highest possible response option is at Level 4.
2. For some items, it is impossible to write (and consequently, for students to select) a response option at the lowest levels of the ESS learning progression. This impossibility creates a floor effect at the item level. For example, for item 2 in Figure 2, the lowest possible response option is at Level 3.
3. Many items feature more than one response option linked to the same level. This duplication increases the likelihood that students who guess might select an option indicating a particular level of understanding by chance. For example, for both of the items in Figure 2, two of the five possible response options are at Level 3.

These features may appear in OMC items owing to practical considerations or to design preference. From a practical standpoint, it may be impossible to find response options that span the full range of a learning progression. The highest level of the learning progression may not be required to explain a particular context fully, or more complex contexts may be inaccessible to students at the lower levels of the learning progression. In addition, the highest-level response option often includes scientific words that too easily suggest the most correct response compared to lower-level response options that do not include these words. In terms of design preference, a ceiling effect may be built into the design of a subset of OMC items if the assumption is that students have not yet been taught the skills and concepts needed to respond at the highest level. Or a floor effect may be built into a subset of OMC items if the assumption is that students have already learned the skills and concepts at lower levels. Or an OMC item may include multiple response options at the same level because the options are qualitatively different but cannot be ordered. Although these options do not add additional information about a student's level on the learning progression, they may provide qualitative information about nuances in students' thinking. Therefore, such options may be included in the design of the OMC items if there are multiple ways of thinking that are consistent with a particular level of the learning progression.

DATA AND CLASSICAL ITEM STATISTICS

In the 2008–2009 school year we administered a science test to 1,088 high school students (grades 9–12) at six high schools in rural and suburban Iowa. Any student enrolled in a science course at these schools was eligible to participate in the study. However, not all science teachers granted permission for data collection in their classes. Participating students were enrolled in 68 science classes at a range of levels—from freshmen-level to upper-level courses. The test consisted of 28 OMC items: 12 items associated with a hypothesized learning progression for ESS and 16 items associated with a hypothesized learning progression on the topic of force and motion. We focus here on the results from the 12 ESS OMC items.

Students answered the OMC items in their science classes. The average participation rate across all classes was 83%. The sample was fairly evenly divided between male and female students (52% male; 48% female). High school students were chosen for the study to minimize guessing. Most high school students should have been exposed to the ideas in the two learning progressions and therefore would not need to guess at answers. However, a drawback of the sample was that we were less likely to find students choosing options consistent with the lower ends of the learning progressions. After completing the ESS OMC items, students answered the following question: "Was the content of [these] questions covered in a science class you've taken?" While 46% of the students answered "yes," 25% answered "no," 28% answered "I am not sure," and 2% did not respond. A partial explanation of these results may be that ESS is not always taught in high school science curricula.

Table 1 shows the distribution of student OMC item responses mapped to the levels of the ESS learning progression. The items are arranged from easiest to hardest, where "easiness" is defined as the proportion of students selecting a response option at the highest possible level. For example, 74% of students selected the highest possible response option for item 11 ("Which picture best represents the motion of the Earth (E) and Sun (S)?"). Thus item 11 is the easiest item. Only 20% of students selected the highest possible response option for item 8 ("Which is the best explanation for why we see a full moon sometimes and a crescent moon other times?"). Thus item 8 is the hardest item. The shaded cells in Table 1 represent levels where there is no corresponding response option for the item. Only three items (items 9, 2, and 8) permitted a response at the highest level of the ESS learning progression. For five items (items 11, 6, 4, 7, and 2), response options were linked to only three of the five possible levels. Cells with two numbers expressed as a sum represent items with two response options associated with the same score level. For example, for item 11, 7% of students chose response option A and 14% chose response option B, but both response options are linked to Level 2 of the ESS learning progression. Finally, the bottom row of Table 1 gives the item to total score correlations (i.e., point-biserials) associated with the highest possible response option for each item.

Table 1. Observed Distribution of ESS OMC Item Responses.

	Easiest		←			*OMC Items*				→	*Hardest*	
Level	*11*	*6*	*3*	*1*	*4*	*12*	*10*	*5*	*9*	*7*	*2*	*8*
5									43		28	20
4	74	72	66	64	63	62	61	59	41	29	34+15	35
3	5	16+4	15+5	14+6	21	12	18	14	7	47+14	10+14	18
2	7+14	7+2	12	12	11+6	14	7+5	7	9	11		15+11
1			3	5		12	9	14+5				
pt-bis	0.59	0.69	0.60	0.64	0.54	0.69	0.68	0.68	0.33	0.12	0.36	0.21

Note. Values in cells are the percentage of students choosing a response option linked to a level of the underlying learning progression. Some columns may not sum to 100 due to rounding error.

Table 1 presents information about classical item statistics that some would use to evaluate item quality. For example, from item to item, a majority of students chose response options linked to the highest two available levels of the ESS progression (i.e., Levels 3 and 4 or Levels 4 and 5). The three items that allow for a Level 5 response had point-biserials less than 0.4, a value that is typically considered a cutoff for a "good" item in traditional testing contexts. In other words, the students who chose the Level 5 response option were not necessarily students who performed the best on the other items. Item 7 seems to stand out as a problematic item because there is a very low correlation between a choice of the Level 4 response option and the total score on the other items (point-biserial = .12).[1] Finally, one can use true score theory to estimate the reliability of the total scores derived from these items. An estimate based on Cronbach's alpha coefficient suggests a reliability of 0.67. This implies that about one-third of the variance in OMC scores across students may be attributable to measurement error. If the results from these items were used to support high-stakes inferences about individual students, this would be a cause for concern. On the other hand, if the scores were used for formative purposes or to compare group means, a reliability of 0.67 might be less worrisome.

There is certainly nothing "wrong" with the analysis above, which seems to suggest that some items from the test should be rewritten or that the links between the response options and ESS levels should be reconsidered. However, these interpretations are somewhat arbitrary because, as is well-known, they are highly dependent on the particular sample of students taking the exam. If, for example, we found that the highest proportion of students chose response options associated with Levels 2 and 3, would this finding indicate a problem with the ordering of the item options? Or would it reflect the fact that the students have not been exposed to this content in their curriculum? Furthermore, the analysis above might provide equivocal diagnostic information when it is disaggregated at the student level. Consider the following two randomly selected student response vectors for two students— "Liz" and "Andrew"—shown in Figure 3.

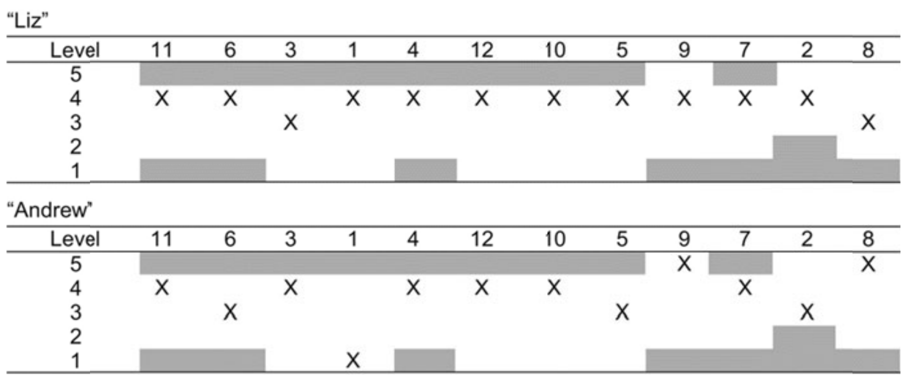


Figure 3. Observed ESS level classifications for two students' item-level responses.

If each student's set of item responses were summarized by a report of central tendency, the summary would show that the median score level for both students is 4, with Liz's arithmetic average (3.83) just slightly higher than Andrew's (3.67). Yet clearly, the greater variability in Andrew's item responses implies greater uncertainty about the diagnostic utility of the information. While teachers can draw informal, qualitative inferences on a student-by-student basis, this may be a very subjective and time-consuming process. Thus there are some advantages to engineering the diagnostic process using a psychometric model. Such a model can test the validity of the hypothesized learning progression and provide formal, probabilistic inferences about student understanding. In the next section we discuss some challenges inherent in this endeavor before presenting one possible solution—the AHM—in detail.

CHALLENGES TO MODELING OMC ITEM RESPONSES TO SUPPORT DIAGNOSTIC INFERENCES

We now focus on two important and related challenges that arise after making the decision to model OMC item responses probabilistically in order to make diagnostic inferences. The first challenge is to choose the functional form of the model; the second is to make plausible assumptions about whether the latent variable (or variables) “measured” is (are) discrete or continuous.

“Modeling” item responses refers to the activity of making a formal statement about the factors involved when a student interacts with an assessment item. In item response theory (IRT; De Boeck & Wilson, 2004; van der Linden & Hambleton, 1996), this formal statement is made in terms of an *item response function* (IRF). Let the variable X_{pi} represent possible responses to assessment item *i* that could be given by student *p*. An IRF provides a mathematical expression for the probability of observing an item response in score category k as a function of one or more parameters (i.e., dimensions) specific to respondents ($\boldsymbol{\theta}_p$), and one or more parameters specific to items ($\boldsymbol{\xi}_i$):

$$P(X_{pi}=k)=f(\boldsymbol{\theta}_p,\boldsymbol{\xi}_i). \tag{1}$$

The very general expression in Equation 1 accommodates IRFs that range from the very simple (a single parameter for each student and a single parameter for each item) to the very complex (multiple parameters per individual student and item). One well-known example of an IRF, which is often applied when modeling student responses to the traditional multiple-choice items found on most large-scale assessments, is the three-parameter logistic model (3PL; Birnbaum, 1968):

$$P\left(X_{pi}=1 \mid \theta_p\right)=c_i+\left(1-c_i\right)\frac{e^{a_i\left(\theta_p-b_i\right)}}{1+e^{a_i\left(\theta_p-b_i\right)}}. \tag{2}$$

The 3PL model is so-named because three distinct parameters are specified for every item that a student answers (a_i, b_i, c_i). Values of the parameter a_i affect the *slope* of the IRF. The larger the value of a_i, the steeper the curve. This means that on items with relatively large values of a_i, a small change in (unidimensional) $_p$ will produce a large change in the probability of a correct response. Because such items appear better at discriminating between respondents with different underlying values of θ, the parameter a_i is sometimes referred to as the item discrimination parameter. In contrast, values of the parameter b_i affect the *location* of the IRF. The larger the value of b_i for an item, the larger the value of θ needed for a respondent to have a high probability of answering the item correctly. The parameter b_i is often referred to as the item difficulty parameter. Finally, values of c_i, which in theory may range between 0 and 1, establish a lower asymptote for the IRF. The larger the value of c_i, the higher the “floor” on the probability that a respondent will answer the item correctly. Because it is intended to capture the possibility that respondents have answered an item correctly by guessing, the parameter c_i is often referred to as the guessing parameter. The inclusion of a single student-specific variable θ in Equation 2 brings the total number of parameters to four.

An example of an IRF with a simpler functional form is the Rasch Model (Rasch, 1960):

$$P\left(X_{pi}=1\,|\,\theta_p\right)=\frac{e^{\left(\theta_p-b_i\right)}}{1+e^{\left(\theta_p-b_i\right)}}. \tag{3}$$

Upon inspection, the mathematical difference between the IRFs in Equations 2 and 3 is that in Equation 3, the item parameters a_i and c_i have been constrained to be equal to 1 and 0, respectively. Yet there are also important differences between the two IRFs that are more philosophical than mathematical. In the Rasch tradition, items are developed to fit the model because when the data fit the Rasch model, it is possible to make invariant comparisons between respondents. That is, comparisons of students do not vary as a function of the specific items chosen for a test instrument, just as comparisons of items do not depend upon the specific sample of students who respond to them. The alternative tradition is to view the data as fixed and to choose an IRF that bests fits the data—whether this leads to the Rasch Model or something much more complex. (An extended discussion of these two positions is outside the scope of this chapter; for additional details, see Andrich, 2004; Bock, 1997; Thissen and Wainer, 2001; Wilson, 2005; Wright, 1997). We raise this issue in order to make the broader point that even for assessment items with traditional score formats, the selection of an IRF using an IRT-based approach is not straightforward. The problem is that there are two criteria for optimality. First, there is the technical need to model observed item responses as faithfully as possible. Second, there is the practical need to choose models that are parsimonious and readily interpretable. The Berkeley Evaluation and Assessment Research (BEAR) Assessment System, which has been previously applied to model learning progressions in science, is an example of an IRT-based approach that prioritizes the latter criterion (Wilson, 2009).

Because learning progressions attempt to distinguish between multiple levels of understanding, their associated items often (if not usually) must be scored in more than two ordinal categories (i.e., polytomously). The OMC format described in this chapter is an example. At minimum, the specification of an IRF for OMC items should take this added complexity in scoring into account. In addition, a parameterization must be selected to address the obstacles to score interpretations noted previously: floor and ceiling effects, as well as multiple response options linked to the same score level. De Boeck and Wilson (2004) offer one paradigm for such decision-making in their book *Explanatory Item Response Models.* This edited volume gives many examples of conventional IRFs of IRT models that are expanded by the addition of new variables and parameters that explain variability in observed item responses. Another example is the "mixture model" that proposes a solution for the problem that students may guess the correct OMC response option (Mislevy & Verhelst, 1990; Wilson, 1989). This model posits two populations of students: those who guess when they don't know the most sophisticated item response and those who do not guess in the same circumstance. Using this model, it is possible to specify two distinct IRFs, one for each hypothetical population of students. Yet another example deals with the fact that certain items have multiple response options at the same level while others do not. In the "logistic latent trait model" (Fischer, 1977, 1983) item difficulty is modeled as a function of both the levels of possible response options and the number of these options.

In recent years several authors have argued that even these elaborated IRFs may not be ideal if the purpose of the model is to make diagnostic classifications of students (Junker & Sijtsma, 2001; Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010). IRT models posit that the latent variable (or variables) underlying a student's item responses is (are) continuous. As a result, the estimation of student-specific values for these variables does not lead to direct classification of students into discrete categories. Rather, a second step is needed in which "cut-points" are established along the student-specific latent variable θ. This step typically requires some degree of subjective judgment as is the case when criterion-referenced standards are established for student performance on large-scale assessments.

This argument emphasizes the murkiness in defining θ as a "latent variable" or student "ability." In the context of learning progressions, one may say that θ represents at least one attribute that becomes more sophisticated as students receive instruction. In this sense, θ is some unknown (i.e., latent) variable that assumes values that span multiple levels of a hypothesized learning progression. But what is the mapping between the values of the latent variable and the levels of the learning progression? If θ is assumed to be continuous while the levels of the learning progression are discrete, then to some extent there is a mismatch between the granularity of the hypothesis that underlies the design of assessment items and the granularity of the latent variable that underlies the design of the psychometric model. Such a mismatch seems inherent when θ is defined as a continuous latent variable in an IRT-based approach.

An alternative is the specification of what Rupp et al. (2010) describe as *diagnostic classification models* (DCMs). DCMs can be distinguished as models in which the latent variables of interest are discrete rather than continuous, and the objective of the models is to provide a profile of knowledge and skills based on statistically derived classifications (Rupp & Templin, 2008). A taxonomy of models that fit this definition is outside the scope of this chapter (see Rupp et al., 2010 for these details). However, in the next section we illustrate the basic principles of a DCM-based approach by showing how the AHM (a specific DCM-based approach) can model OMC item responses according to the hierarchy implied by the ESS learning progression.

Before proceeding, we emphasize that we do not argue that IRT-based approaches are invalid for making diagnostic inferences. Even when the assumptions of an IRT model are wrong, and those of a DCM are right (and unfortunately, the truth is never known a priori), the former may provide a first-order approximation of the latter, and vice versa when the conditions are reversed. This is an empirical question that we do not address in this chapter. A latent variable is, after all, by definition unobservable, so assumptions are unavoidable. Nevertheless, in our view, the assumption that a latent variable has a continuous structure (implicit in IRT) is much stronger and less plausible than the assumption that the variable has an ordinal structure (implicit in a DCM).[2] This view provides some motivation for the approach described in the next section. We consider other pros and cons of taking a DCM-based approach rather than an IRT-based approach in the final section of the chapter.

APPLYING THE ATTRIBUTE HIERARCHY METHOD TO OMC ITEMS

Background on the AHM

There has been an explosion in the development of DCMs for cognitive diagnostic assessment in the past decade.[3] The pioneering work by Kikumi Tatsuoka, originating in the early 1980s, has inspired much of the interest in such models. Tatsuoka's premise is fairly simple: the score derived from a set of items (i.e., θ in IRT) often obscures important diagnostic information about more fine-grained "attributes" that students use to solve problems within a given domain. To address this problem, Tatsuoka developed the idea of a Q matrix that allows for the formal specification of a hypothesized linking between attributes and items. Specification of a Q matrix makes it possible to generate expected item response patterns associated with specific knowledge states where these states are defined by the attributes that a student does or does not have. Given these expected response patterns and students' actual response patterns, Tatsuoka developed the Rule Space Method as a pattern-matching technique for probabilistic diagnostic classification.

More recently, Leighton et al. (2004) introduced an extension of Tatsuoka's Rule Space Method called the Attribute Hierarchy Method (AHM). The AHM takes as its starting point the assumption that the construct of measurement is comprised of finer-grained "attributes" that have an ordered, hierarchical relationship. The specification of this relationship precedes and guides the specification of a "reduced form" Q_r matrix. While many DCM applications assume that all attributes are independent and/or non-hierarchical, in the AHM, a hierarchical dependence among attributes is central to the theory. In our view, this feature makes the AHM an appealing candidate for modeling learning progressions since learning progressions specify the hierarchical distinctions in student understanding as it becomes more sophisticated. Applications of the AHM to date have involved traditional multiple-choice items that are scored dichotomously (Gierl, Wang, & Zhou, 2008; Leighton et al., 2004). The application of the AHM to polytomously scored OMC items is a novel extension of this modeling approach.

There are two stages to the AHM. In the first stage, an attribute hierarchy is specified based upon the construct of measurement. The attribute hierarchy is then used to characterize the cognitive features of items through a Q_r matrix. This stage allows the generation of distinct expected item response patterns that characterize the pre-specified attribute combinations comprising the hierarchy. In the second stage, expected response patterns are compared to observed response patterns using either a parametric or a nonparametric statistical classification approach. The result is a set of probabilities that characterize the likelihood that a student with a given item response pattern has a level of understanding consistent with a hypothesized level in the attribute hierarchy. In addition to the calculation of these probabilities, one can also generate hierarchy fit indices and estimates of reliability at the attribute level. Next we illustrate the first stage of the AHM as it could map to the ESS learning progression and associated OMC items. We then illustrate the essence of the second stage and provide a hypothetical illustration of how the results from this stage could be used diagnostically.[4]

Stage 1: Specifying a Learning Progression as an Attribute Hierarchy

We begin by translating the qualitative descriptions that distinguish the levels of our existing ESS learning progression (Figure 1) into attributes that can be coded dichotomously as either present or absent in a student.

A1: Student recognizes that objects in the sky move systematically.

A2: Student knows that the Earth orbits the Sun, the Moon orbits the Earth, and the Earth rotates on its axis.

A3: Student can coordinate apparent and actual motions of objects in sky.

A4: Student can incorporate the motions of the Earth and Moon into a complete description of motion in the Solar System that explains the day/night cycle, phases of the Moon, and the seasons.

The proper grain size of these attributes will always be a matter for debate. For example, the attribute A2 could easily be split into three smaller attributes. The more finely specified the attributes, the easier it is to code them as present or absent. However, the larger the number of attributes, the harder it is to distinguish them with a finite number of test items, and the more difficult it is to summarize them as a diagnostic assessment of student understanding. We return to this issue in the concluding section of the chapter.

Next we specify a hierarchy among these attributes. In this example, the hierarchy is fairly straightforward and mirrors the hierarchy implicit in the original ESS learning progression: A1 → A2 → A3 → A4. These attributes are conjunctive—a student must possess an attribute lower in the hierarchy (e.g., A1) in order to possess a higher attribute (e.g., A4). The combinations of these four attributes may be used to define the levels of the ESS learning progression.

Level 1 = No attributes

Level 2 = A1

Level 3 = A1 & A2

Level 4 = A1 & A2 & A3

Level 5 = A1 & A2 & A3 & A4

The simple attribute hierarchy above leads to the specification of two matrices. An "adjacency" matrix

$$A = \begin{vmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

and a "reachability" matrix

$$R = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

The *A* matrix represents all the direct dependencies between attributes, where each row and column combination above the main diagonal represents a unique attribute combination. In this example, the first row of the matrix has the following interpretation: The knowledge that the Earth orbits the Sun, the Moon orbits the Earth, and the Earth rotates on its axis (A2, second column) depends directly on recognizing that objects in the sky move systematically (A1, first row). There is a similar interpretation for the other nonzero cells in the *A* matrix. The *R* matrix represents both direct and indirect dependencies. Hence row 1 of the *R* matrix indicates that attributes A2, A3, and A4 depend on attribute A1. For A2, the dependency is direct (as indicated in the *A* matrix); for attributes A3 and A4, the dependency is indirect. The *A* and *R* matrices can be manipulated using Boolean algebra to specify the "reduced" Q matrix Q_r. In applications of the AHM with traditional multiple-choice items, a Q_r matrix has dimensions *a* by *I*, where *a* represents the number of attributes, and *I* represents the number of items. Because OMC items are scored polytomously, the associated Q_r matrix is considerably more complicated. For our set of items, there is a 4 attribute by 55 item option matrix instead of a 4 attribute by 12 item matrix since each item-specific option has a separate column (and items in this set contain four or five options). For ease of presentation, Table 2 shows an excerpt of the Q_r matrix for the ESS OMC items using only items 2 and 3 from Figure 2.

Table 2. Excerpt of the Q_r matrix Associated with ESS Attribute Hierarchy.

	2A	2B	2C	2D	2E	3A	3B	3C	3D	3E
A1	1	1	1	1	1	1	1	1	0	1
A2	1	1	1	1	1	0	1	1	0	1
A3	1	0	0	1	1	0	1	0	0	0
A4	0	0	0	1	0	0	0	0	0	0
Level	4	3	3	5	4	2	4	3	1	3

Note. Columns represent OMC item options; rows represent hypothesized attributes that must be present for students to select each item option. The row and column labels, as well as the indication of the learning progression level corresponding to each item option, are included to make the matrix easier to interpret.

The interpretation of the Q_r matrix for OMC item 2 ("Which is the best explanation for why we experience different seasons?") option A ("The Earth's orbit around the Sun makes us closer to the Sun in the summer and farther away in the winter.") is as follows: In order to select this response option, a student should possess attributes A1, A2, and A3. However, attribute A4 is not a prerequisite for selecting option A. The columns for the other possible response options have similar interpretations. The presence of a "1" in a row indicates that the associated attribute is a prerequisite for the response; the presence of a "0" indicates an attribute that is not a prerequisite. The Q_r matrix leads naturally to the specification of an expected response matrix for OMC items, where each row of the matrix represents the expected response to each OMC option for students with each conceivable attribute combination. Note that the expectation for the level associated with each item option depends upon the accuracy of the central hypothesis regarding the attribute structure and its relationship to items. Table 3 shows the excerpt from an expected item response matrix that corresponds to the Q_r matrix in Table 2.

Table 3. Excerpt from an Expected Response Matrix for ESS OMC Items.

Hypothetical Student	*Expected Responses by Item [2] [3]*	*Attributes [A1 A2 A3 A4]*	*ESS Level*
1	$\tfrac{1}{5}\tfrac{1}{5}\tfrac{1}{5}\tfrac{1}{5}\tfrac{1}{5}\ \ [00010]$	0000	1
2	$\tfrac{1}{5}\tfrac{1}{5}\tfrac{1}{5}\tfrac{1}{5}\tfrac{1}{5}\ \ [10000]$	1000	2
3	$0\tfrac{1}{2}\tfrac{1}{2}00\ \ \ 00\tfrac{1}{2}0\tfrac{1}{2}$	1100	3
4	$\tfrac{1}{2}000\tfrac{1}{2}\ \ [01000]$	1110	4
5	$[00010]\ [01000]$	1111	5

Note. The row and column labels and the last column ("ESS Level") are included to make the matrix easier to interpret.

The expected responses for OMC items 2 and 3 are given in brackets in the second column of Table 3. For example, consider a hypothetical student with a Level 1 understanding of ESS according to our learning progression. This student does not yet have attributes A1 through A4. Yet, for item 2, all possible response options require at least one of these attributes. Therefore, we may reasonably assume that the student will guess among the available response options; hence we insert a 1/5 for each expected response. (An alternative procedure is to assign the item options associated with fewer attributes—or lower levels of the learning progression—higher probabilities than those with more attributes.) In contrast, item 3 includes a response option (D) that requires no attributes. Hence the expected response string for this hypothetical student is [00010].

Note that it is the combination of the attribute hierarchy (the *A* matrix) and the Q_r matrix (Table 2) that are used to generate conditional expected item responses—the student by item response combinations we expect to observe if the hypotheses underlying both the *A* and Q_r matrices are true.

In this example, a strategy for modeling OMC items with floor effects, ceiling effects, and multiple options comes into clearer focus.

- When the ability of a student is below that of the lowest available OMC option, assume that the student is guessing (e.g., expected response patterns of hypothetical students 1 and 2 for item 2).
- When the ability of a student is above that of the highest available OMC option, assume that the student will choose the highest available option (e.g., expected response pattern of hypothetical student 5 for item 3).
- When there are multiple options at a student's level, assume the student has an equal chance of selecting either option (e.g., expected response patterns of hypothetical students 3 and 4 for item 2).

Stage 2: Classifying Students Probabilistically into Attribute Profiles

Establishing the expected item response matrix marks the culmination of the first stage of the AHM. In the second stage one must establish the criteria used to classify students into learning progression levels on the basis of their observed item response patterns. The purpose of this stage is to facilitate the probabilistic mapping of observed responses to the expected responses for students at each level of the ESS learning progression. A starting point is to simulate item responses for hypothetical students at each level of the learning progression (i.e., with each possible combination of attributes). We simulate data under the constraint that the learning progression is true in order to compare simulated responses with item responses from real students who are, of course, unlikely to give responses that perfectly match our initial hypothesis.

To illustrate the process of simulating such a dataset, suppose we wished to simulate item responses for N students, uniformly distributed across the five levels of the ESS progression. (Note that no assumption is made that students in actual school settings are uniformly distributed across all five levels—our aim is to characterize all possible item response patterns that could, in theory, be observed.) The item responses that are expected for students at each level of the learning progression appear in the excerpt for an expected response matrix associated with items 2 and 3 in Table 3. We return to this example in the context of simulating item response vectors. By a "vector" we mean a sequence of item responses. For the test as a whole, each vector consists of a sequence of option choices for the 12 items; in our example the sequence consists of only two items.

For item 2 ("Which is the best explanation for why we experience different seasons on Earth?"), we expect students who are at Level 5 of the learning progression to choose answer D, which is a Level 5 response. For item 3 ("Which best describes the movement of the Earth, Sun, and Moon?"), we expect the Level 5 students to select the highest-level option (B). For these two items, the response vector we expect for all students at Level 5 is DB.[5] We simulate this item response pattern for N/5 students in our dataset. For students at Level 4, there is a complication. For item 3, there is only one response option at Level 4 (B), but for item 2 there are two possible response options at Level 4 (A and E). It follows that there are two equally plausible response vectors: AB and EB. Each vector must be simulated for half of the N/5 students generated at Level 4 in the dataset. Now consider students at Level 3. On both items 2 and 3 there are two possible response options at Level 3. This means that four item response vectors are equally plausible: BC, BE, CC, CE. Each vector is simulated for one-fourth of the N/5 students generated at Level 3 in the dataset. Finally, for students at Levels 1 and 2, item 2 has no response options available at their levels. However, for item 3, there is one associated response option per learning progression level (option A is Level 2; option D is Level 1). In simulating item responses for these students, we assume that when their level of understanding is below the available response options, they will guess. Hence, for students at both Levels 1 and 2 there are five plausible item response vectors: AD, BD, CD, DD, ED for Level 1; AA, BA, CA, DA, EA for Level 2. Each vector is simulated for one-fifth of the N/5 students generated at Levels 1 and 2 in the dataset. Table 4 summarizes the simulated data set that results from this process.

This example illustrates that the simulation of distinct item response vectors corresponding to each hypothesized learning progression level in the OMC context is more and more complicated with increases in (a) the number of items, (b) the complexity of the attribute structure, (c) the number of item floor effects, and (d) the number of items with multiple options linked to the same attributes/levels. The last column of Table 4 shows the total score when the scored item responses (learning progression levels) for each expected response vector are added. One can see from this that the total score could be a potentially misleading statistic if it were to be used for diagnostic classification, as it does not necessarily provide an accurate ranking of these simulated students in terms of the learning progression levels used to generate the simulated data.

Table 4. Simulated Dataset Based on Idealized Item Responses to Items 2 and 3.

Distinct Item Response Vector	*Learning Progression Level (Attributes)*	*Simulated Sample Size*	*Plausible Item Response Vector*	*Total Score (Item 2 Level + Item 3 Level)*
1	5 (A1 & A2 & A3 & A4)	N/5	DB	9
2	4 (A1 & A2 & A3)	N/10	AB	8
3	4 (A1 & A2 & A3)	N/10	EB	8
4	3 (A1 & A2)	N/20	BC	6
5	3 (A1 & A2)	N/20	BE	6
6	3 (A1 & A2)	N/20	CC	6
7	3 (A1 & A2)	N/20	CE	6
8	2 (A1)	N/25	DA	7
9	2 (A1)	N/25	AA	6
10	2 (A1)	N/25	EA	6
11	2 (A1)	N/25	BA	5
12	2 (A1)	N/25	CA	5
13	1 (None)	N/25	DD	6
14	1 (None)	N/25	AD	5
15	1 (None)	N/25	ED	5
16	1 (None)	N/25	BD	4
17	1 (None)	N/25	CD	4

The step from simulating a dataset with deterministic item responses to using the information in this dataset as a basis for classifying the likelihood of attribute patterns associated with observed item response vectors can be rather complicated. Multiple approaches have been suggested (e.g., Gierl et al., 2008; Leighton et al., 2004). While the details are outside the scope of this chapter, the basic idea can be explained by returning to the example of the two students we met previously, Liz and Andrew. If we consider only items 2 and 3, the observed response vector for Liz is AC (which is scored as a Level 4 and a Level 3 response), and the observed response vector for Andrew is BB (which is scored as a Level 3 and a Level 4 response). Neither of these item response vectors is among those expected if the attribute hierarchy is true. If both students were actually at Level 3 of the learning progression (i.e., they have mastered attributes A1 and A2 but not A3 and A4), then (according to the model) they each chose one answer that constitutes an "error" in a positive direction. If both students were actually at Level 4 (i.e., they have mastered attributes A1, A2 and A3 but not A4), then (according to the model) they each chose one answer that constitutes error in a negative direction. To determine which scenario is more plausible, more information is needed about the overall probabilities that students will "slip" (give a response that is lower than expected) or "guess" (give a response that is higher than expected); analysis of the item response patterns for the complete sample of students would provide this information.

If, after comparing expected and observed response vectors, element by element, across all items and students, we find few matches between expected and observed responses, this finding would provide evidence against the hypothesized attribute hierarchy. That evidence would then raise questions about the validity of the learning progression. To evaluate this possibility, one can compute a Hierarchical Classification Index (HCI). The HCI takes on values between -1 and +1. According to simulation work by Cui and Leighton (2009), values above 0.7 are interpreted as an indication of acceptable fit. When misfit is found, then one must revise the attribute hierarchy, revise the hypothesized relationship between items and this hierarchy, revise the items themselves, or revise all of the above.

Imagine now that we have computed the HCI for these OMC items and have convinced ourselves that, on the basis of the gathered data, the hypothesized learning progression is at least tenable. How could we use the results from applying the AHM to facilitate diagnostic inferences about Andrew's understanding of ESS? Figure 4 shows an example of a diagnostic profile display we could give Andrew's teacher. This display indicates the probability that Andrew has each of the attributes that comprise the ESS learning progression. In examining the display, the teacher may conclude that Andrew is likely to have attributes A1 and A2 and, therefore, is generally thinking about ESS with a Level 3 understanding on the learning progression. There is some evidence that Andrew is beginning to coordinate the apparent and actual motion of objects in the sky (attribute A3), but he is not yet doing so consistently. It is not surprising that this conclusion is unlikely to differ substantially from the one that would be reached through a more subjective visual inspection of Andrew's item response pattern (Figure 3). The purpose of the model in this context is to complement rather than replace teacher judgment. If the teacher's judgment, based on a careful inspection of the item observed responses, disagreed substantially with the probability profile, there would be reason for concern.

Probability	A1	A2	A3	A4
1.00				
.90	***			
.80				
.70		***		
.60				
.50				
.40			***	
.30				
.20				
.10				
.00				***

A1: Student recognizes that objects in the sky move systematically.
A2: Student knows that the Earth orbits the Sun, the Moon orbits the Earth, and the Earth rotates on its axis.
A3: Student can coordinate apparent and actual motions of objects in sky.
A4: Student can incorporate the motions of the Earth and Moon into a complete description of motion in the Solar System that explains the day/night cycle, phases of the Moon, and the seasons.

Figure 4. Andrew's attribute profile for the ESS learning progression.

DISCUSSION

In this chapter we illustrated a novel method for the psychometric modeling of OMC items. At heart, building any psychometric model is about comparing observed and expected student item responses. The process of delineating what is expected forces the developer of a learning progression to make some formal commitments about the appearance of more or less sophisticated expressions of conceptual understanding. In this chapter we described how this process might unfold when applying a specific DCM, the AHM.

One strength of the AHM is that it requires the developer of a learning progression to be very explicit about the specific elements of student understanding—the "attributes"—that change as a student progresses from naïve to sophisticated levels of understanding. This specification essentially involves breaking down level descriptors into a sequence of binary codes. Combinations of the binary codes define movement from one level to the next. This process generates a Q_r matrix that formally maps assessment items to the specific attributes students are expected to have in order to answer each item correctly. Use of the AHM focuses attention on the link between hypothesized levels of a learning progression and the corresponding expectations for item response patterns.

In our application involving the OMC format, we noted the challenges presented by floor and ceiling effects and by multiple response options linked to the same learning progression level. Briggs et al. (2006) suggest two IRT-based approaches for the psychometric modeling of OMC items: the Ordered Partition Model (Wilson, 1992) and the Multiple-Choice Model (Thissen & Steinberg, 1997). It is still possible to take one of these approaches in stage 2 of the AHM after simulating a sample of expected item response vectors under the preliminary assumption that the specified attribute hierarchy is correct. In such a scenario, the AHM may be viewed as a complement to an IRT-based approach. However, this view negates one motivation (as noted earlier) for applying a DCM: There is no assumption of continuity for the construct of measurement. A different tactic for stage 2 of the AHM modeling approach is to view the activity as "pattern matching" and then to invoke a neural network approach or Tatsuoka's Rule Space Method to classify students into learning progression levels. In this case, the AHM constitutes a genuine alternative to an IRT-based approach.

Because all DCMs (of which the AHM is a specific example) take a confirmatory modeling approach, the ability to evaluate model fit is critical. Although considerable progress has been made recently, indices of model fit (e.g., the HCI) and their interpretation are not yet well-established for DCMs. When a DCM produces output that suggests a low probability of classifying a student at any level of a learning progression, an important question arises concerning the fit of the student to the model, and vice versa. Qualitative investigation of these discrepancies is needed to advance our understanding of how students learn about scientific phenomena. In addition to internal evaluations of model fit, an alternative approach is to compare student classifications that result from more exploratory model specification. For example, Steedle and Shavelson (2009) used an exploratory modeling approach that did not begin with an a priori learning progression hypothesis (i.e., an exploratory latent class model). In their case, the results showed diagnostic classifications with substantively different interpretations of what students appeared to know and could do compared to the results from a more confirmatory diagnostic model similar to the AHM.

A potential weakness of taking a DCM-based approach is that this class of models is intended for applications in which there is a desire for very fine-grained diagnoses and for which the attributes can be very precisely specified as "present" or "absent." It is unclear whether such fine-grained specification is possible (or even desirable) for some learning progressions under development in science education. In general, the more qualitative and holistic the learning progression, the less amenable it is to a DCM-based approach. For example, we found that the force and motion learning progression (Alonzo & Steedle, 2009) is much harder to map using the AHM than is ESS learning progression described in this chapter. Taking an IRT-based approach is sometimes viewed as a solution to this problem because IRT is thought to provide for inferences at a larger grain size (since constructs are typically specified in terms of multiple attributes). However, it may be harder to defend the diagnoses that result from an IRT-based approach after the continuum has been segmented through a process that may or may not follow from substantive theory (i.e., standard-setting panels).

Regardless of the approach chosen for the psychometric modeling of responses to an assessment item format such as OMC, the approach has to satisfy at least two criteria. First, the approach must facilitate diagnostic classifications according to an underlying learning progression. The classification should have formative utility for classroom instruction. Second, the approach must enable the developer of a learning progression to evaluate whether the initial hypothesis of the learning progression, and its instantiation using assessment items, can be supported empirically. Hence a program of study on the use of a DCM to model OMC items requires at least two distinct strands: one that depends upon the technical quality of the model specified (in part, through simulation work) and another that depends upon an examination of the extent to which stakeholders (e.g., teachers) use the diagnostic information the model provides. The evidence from these two strands of research will move the learning progression concept from a merely interesting idea to a validated idea.

ACKNOWLEDGEMENTS

Initial development of the Earth in the Solar System learning progression and associated items was conducted at the University of California, Berkeley, in collaboration with Cheryl Schwab and Mark Wilson. This work was supported with funding from the National Science Foundation (#REC-0087848, Research in Standards-based Science Assessment) as part of a collaboration between the BEAR Center and WestEd. Additional data collection was funded by the University of Iowa Department of Teaching & Learning and the Iowa Mathematics & Science Education Partnership (Grant #IMSEP1-10-08). Any opinions, findings, conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily represent the official views, opinions, or policy of the National Science Foundation or other funding agencies.

NOTES

[1] The stem for item 7 was "A solar eclipse is possible because," and the Level 4 response option was "The Sun is much bigger than the Moon and much further away from the Earth." The Level 3 response chosen more frequently was "The Moon is always closer to the Earth than the Sun is."

[2] For detailed arguments in support of this perspective, see Michell (1990, 2008).

[3] For books, see Leighton and Gierl (2007); Tatsuoka (2009); Rupp et al. (2010). For an example of journal articles, see the special issue of the *Journal of Educational Measurement* co-edited by DiBello and Stout. (2007). For conference symposia, see the programs of the annual meeting of the National Council for Measurement in Education between the years 2007 and 2010.

[4] The actual implementation of the AHM with these OMC items is beyond the scope of this chapter. A forthcoming manuscript will address this topic.

[5] To make this presentation easier to follow, we have simplified matters by expressing the response to each OMC item in terms of the response choices A to E. For the underlying mathematical specification of the model, the actual response vector for "DB" is written in binary code as <[00010][01000]> as indicated in Table 3.

REFERENCES

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith, Jr. & R. S. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143–166). Maple Grove, MN: JAM Press.

Atwood, R. K., & Atwood, V. A. (1996). Preservice elementary teachers' conceptions of the causes of seasons. *Journal of Research in Science Teaching, 33,* 553–563.

Baxter, J. (1995). Children's understanding of astronomy and the earth sciences. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 155–177). Mahwah, NJ: Lawrence Erlbaum Associates.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novicks (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bisard, W. J., Aron, R. H., Francek, M. A., & Nelson, B. D. (1994). Assessing selected physical science and earth science misconceptions of middle school through university preservice teachers: Breaking the science "misconception cycle." *Journal of College Science Teaching, 24,* 38–42.

Bock, R. D. (1997) A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*(4), 21–32.

Briggs, D. C., Alonzo, A. C., Schwab, S., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429-449.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

DiBello, L. V., & Stout, W. (Eds.). (2007). IRT-based cognitive diagnostic models and related methods [Special issue]. *Journal of Educational Measurement, 44*(4).

Dickinson, V. L., Flick, N. B., & Lederman, N. G. (2000). *Student and teacher conceptions about astronomy: Influences on changes in their ideas*. (ERIC Document Reproduction Service No. 442652)

Fischer, G. H. (1977). Linear logistic trait models: Theory and application. In H. Spada & W. F. Kampf (Eds.), *Structural models of thinking and learning* (pp. 203–225). Bern, Germany: Huber.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *46*, 59–77.

Furuness, L. B., & Cohen, M. R. (1989, April). *Children's conceptions of the seasons: A comparison of three interview techniques*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, CA. (ERIC Document Reproduction Service No. ED306103)

Gierl, M., Wang, C., & Zhou, J. (2008) Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *The Journal of Technology, Learning and Assessment, 6*(6). Retrieved from http://www.jtla.org

Jones, B. L., Lynch, P. P., & Reesink, C. (1987). Children's conceptions of the Earth, Sun, and Moon. *International Journal of Science Education, 9,* 43–53.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Kikas, E. (1998). The impact of teaching on students' definitions and explanations of astronomical phenomena. *Learning and Instruction, 8,* 439–454.

Klein, C. A. (1982). Children's concepts of the Earth and the Sun: A cross cultural study. *Science Education, 65,* 95–107.

Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, *6*, 7–24.

Minstrell, J. (n.d.). *Facets of students' thinking*. Retrieved from University of Washington website: http://depts.washington.edu/huntlab/diagnoser/facetcode.html

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110–128). Kiel, Germany: IPN/Institute for Science Education.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44–73). Washington, DC: National Academy Press.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different persons employ different solution strategies. *Psychometrika, 55*, 195–215.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Newman, D., Morrison, D., & Torzs, F. (1993). The conflict between teaching and scientific sense-making: The case of a curriculum on seasonal change. *Interactive Learning Environments, 3,* 1–16.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Roald, I., & Mikalsen, O. (2001). Configuration and dynamics of the Earth-Sun-Moon system: An investigation into conceptions of deaf and hearing pupils. *International Journal of Science Education, 23,* 423–440.

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 219–262.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Sadler, P.M. (1987). Misconceptions in astronomy. In J. Novak (Ed.), *Proceedings of the second international seminar on misconceptions and educational strategies in science and mathematics* (pp. 422–425). Ithaca, NY: Cornell University.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35,* 265–296.

Samarapungavan, A., Vosniadou, S., & Brewer, W. F. (1996). Mental models of the earth, sun, and moon: Indian children's cosmologies. *Cognitive Development, 11,* 491–521.

Schneps, M. H., & Sadler, P. M. (1987). *A private universe* [DVD]. Available from www.learner.org

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

Stahly, L. L., Krockover, G. H., & Shepardson, D. P. (1999). Third grade students' ideas about the lunar phases. *Journal of Research in Science Teaching, 36,* 159–177.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*, 699–715.

Summers, M., & Mant, J. (1995). A survey of British primary school teachers' understanding of the Earth's place in the universe. *Educational Research, 37*, 3–19.

Targan, D. S. (1987). A study of conceptual change in the content domain of the lunary phase. In J. Novak (Ed.), *Proceedings of the second international seminar on misconceptions and educational strategies in science and mathematics* (pp. 499–511). Ithaca, NY: Cornell University.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the Rule Space Method*. New York, NY: Routledge.

Thissen, D. & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–65). New York, NY: Springer-Verlag.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Trumper, R. (2001). A cross-age study of science and nonscience students' conceptions of basic astronomy concepts in preservice training for high school teachers. *Journal of Science Education and Technology, 10,* 189–195.

van der Linden, W. J., & Hambleton, R. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

Vosniadou, S. (1991). Conceptual development in astronomy. In S. M. Glynn, R. H. Yeany, & B. K. Britton (Eds.), *The psychology of learning science* (pp. 149–177). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the Earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*, 535–585.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16*, 309–325.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*, 716–730.

Wright, B. D. (1997) A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.

Zeilik, M., Schau, C., & Mattern, N. (1998). Misconceptions and their change in university-level astronomy courses. *The Physics Teacher*, 36, 104–107.

MARK WILSON

RESPONDING TO A CHALLENGE THAT LEARNING PROGRESSIONS POSE TO MEASUREMENT PRACTICE

Hypothesized Links between Dimensions of the Outcome Progression

This chapter discusses an important challenge that is posed to measurement practice by the concept of a learning progression. In typical measurement practice, one concentrates on uni-dimensional constructs. If there are multiple dimensions, then multidimensional approaches are taken, such as factor analysis or multidimensional item response modeling. However, learning progressions often include not only multidimensional hypotheses but also the presence of "links" from one dimension to another. Responding to this challenge, in this chapter I (a) offer a description of how the BEAR Assessment System (BAS) can be seen as providing a sensible modeling approach for the uni-dimensional case (and, by straightforward extension, for multidimensional cases) and (b) discuss how this approach can then be expanded to respond to the challenge of hypothesized links between dimensions.

Thus the chapter first summarizes the elements of the BAS, emphasizing the central concept of a construct map, and describes how the idea of a construct map can be helpful in the context of a uni-dimensional learning progression. The chapter then focuses on some of the more complex ways to see the relationship between a set of construct maps and a learning progression (see Wilson, 2009, for a more complete set). Here the chapter uses an example based on the Molecular Theory of Matter in middle school science. This provides the context for a discussion of how a structural equation model (SEM) can be useful for modeling a learning progression and also introduces the structured constructs model (SCM) that goes further than a SEM in modeling aspects of a learning progression. The chapter then discusses some strengths and limitations of this conceptualization and suggests further elaborations. In this chapter, the manner in which the measurement approach supports the learning progression is referred to as the *outcome progression* for the learning progression.

LEARNING PROGRESSIONS: CHALLENGES TO ASSESSMENT

At a recent meeting of researchers working on the topic of learning progressions, the following broad description was suggested by a group consensus:

> Learning progressions are descriptions of the successively more sophisticated ways of thinking about an important domain of knowledge and practice that

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 317–343.

> can follow one another as children learn about and investigate a topic over a broad span of time. They are crucially dependent on instructional practices if they are to occur. (Corcoran, Mosher, & Rogat, 2009, p. 37)

The description is deliberately broad, allowing a wide possibility of usage, but, at the same time, it is intended to reserve the term to mean something more than just an ordered set of ideas, curriculum pieces, or instructional events. As well, the group saw it as a requirement that the learning progression should indeed describe the "progress" through a series of levels of sophistication in the student's thinking. At the same time, the learning progression should still be broad enough to allow for complications such as non-linearity in the ordering and the possibility that the order of levels might differ for different subgroups of students.

Although the idea of a learning progression has links to many older and venerable ideas in education, the history of the specific term "learning progression" in the context of science education is a relatively brief one (Corcoran et al., 2009), starting with the publication of *Systems for State Science Assessment* (National Research Council [NRC], 2006). That report was focused on assessment in K-12 education, and hence the connections between learning progressions and assessment have been there right from the start. A second NRC (2007) report, *Taking Science to School*, also featured the concept of learning progressions, highlighting classroom applications. Several assessment initiatives and perspectives are discussed in these reports, including references to the seminal NRC (2001) report *Knowing What Students Know*. Among the assessment programs highlighted there, probably the most prominent are the work on *progress variables* by the Australian researcher Geoff Masters and his colleagues (e.g., Masters, Adams, & Wilson, 1990; Masters & Forster, 1996) and the closely-related work on the somewhat more elaborated BAS (Wilson, 2005; Wilson & Sloane, 2000). In this chapter I draw on the latter as the core set of assessment perspectives and practices to relate to learning progressions.

It has been pointed out (R. Lehrer, personal communication, October 2010) that there are at least three aspects of the idea of a learning progression that one ought to keep in mind: (a) the developmental aspects, which have to do with the changes in students' understandings as they pass through the learning progression; (b) the instructional aspects, which have to do with the pedagogic activities and environment that are designed to further that progression; and (c) the outcome aspects, which have to do with how one knows where a student is located along the progression at any particular moment in time. In keeping with this set of distinctions, I will refer to the object of scrutiny in this chapter as the "outcome progression," whereby I focus on these outcomes without forgetting that the outcomes only make sense when one takes into account an underlying idea of student development and an accompanying pattern of instruction that might bring about student progress along the learning progression.

From the definition of an outcome progression above, one can see the challenge to measurement methods and practices that come from the idea of a set of "successively more sophisticated ways of thinking." First, for simplicity, assume that there is just one such fully-ordered set of these ways of thinking. In this case, the challenge is to relate the underlying construct to the description of these ways of thinking. What is needed is a way to "bootstrap" the resulting measures back to the qualitative features of the cognitive structure through the observed characteristics of the students' responses to the tasks or other data-generating methods. The BAS (described in some detail below) provides one approach to this simple uni-dimensional version of an outcome progression. However, it may be too simplistic to assume a complete ordering of the ways of thinking. The topic of the outcome progression may involve sub-concepts, or sub-dimensions, each of which has its own ordering. In addition, the sub-dimensions may have complex relationships with one another, or links from one level of one dimension to a level of another dimension. Thus, in the second half of the chapter, the ideas embodied in the BAS are extended to respond to this challenge.

THE BEAR ASSESSMENT SYSTEM (BAS)

The BAS is based on the idea that good assessment addresses the need for sound measurement through four principles: (1) a developmental perspective; (2) a match between instruction and assessment; (3) the generation of quality evidence; and (4) management by instructors to allow appropriate feedback, feed forward, and follow-up. These four principles, plus four building blocks that embody them, are shown in Figure 1. Below I take up each of these principles and building blocks in turn. See Wilson (2005) for a detailed account of an instrument development process that works through these steps. Note that the place of the curriculum is quite prominent in these four building blocks: (1) the construct map will necessarily be shared by the curriculum and the assessment; (2) the contexts and styles of the assessments should relate to the contexts and styles of the curriculum; (3) the evidence for validity and reliability of the assessments will be gathered in the context of one or more specific curricula; and (4) the professional development to support teachers' use of the curriculum and the assessments would necessarily have common elements.

While discussing all four building blocks, this section highlights the first building block—the construct map—and one potential relationship with the idea of an outcome progression. I have labeled this as the assessment structure. It might seem a waste of time to describe all four building blocks when only the first is used in the rest of the chapter; the concern, however, is that unless the place of the construct map in the entire BAS approach is understood, its relevance and importance in the following discussion would be misunderstood. At relevant points in the discussion, issues concerning the items, the outcome space, and the measurement model are also mentioned. But the main focus of this chapter is on the conceptual relationship between the construct map and an outcome progression; hence, these other matters, although they are of great importance for any actual realization of a construct map, are not fully explored.

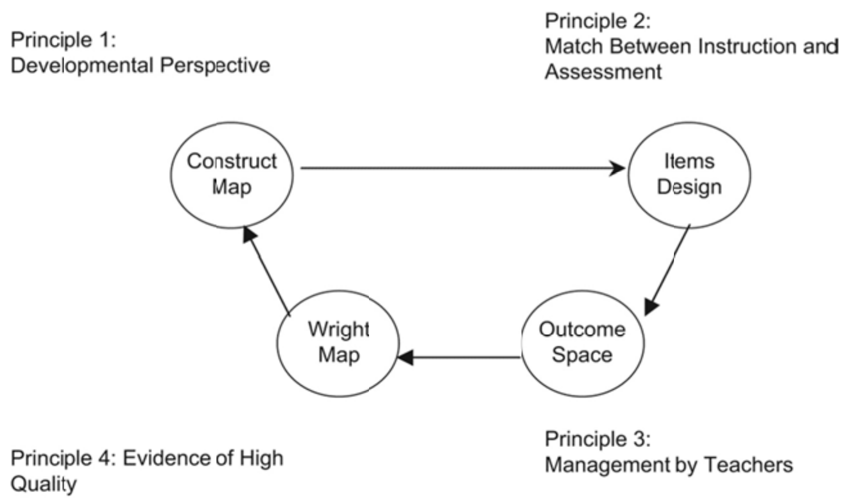


Figure 1. The principles and building blocks of the BEAR Assessment System. From "Measuring Progressions: Assessment Structures Underlying a Learning Progression," by M. Wilson, 2009, Journal for Research in Science Teaching, 46, p. 718. Copyright 2009 by Wiley Periodicals, Inc. Reproduced with permission of Wiley Periodicals, Inc. via Copyright Clearance Center.

Principle 1: A Developmental Perspective

A "developmental perspective" regarding student learning means assessing student understanding of particular concepts and skills longitudinally, as opposed to, for instance, making a single measurement at some final or supposedly significant time point. Decisions about what to assess and how to assess it, whether to focus on generalized learning goals or domain-specific knowledge, and the implications of a variety of teaching and learning theories all impact which approaches might best inform developmental assessment. Taxonomies such as Bloom's Taxonomy of Educational Objectives (Bloom, 1956), Haladyna's (1994) Cognitive Operations Dimensions, and the Structure of the Observed Learning Outcome (SOLO) Taxonomy (Biggs & Collis, 1982) are among many attempts to concretely identify generalizable frameworks. One issue is that as learning situations vary, and their goals and philosophical underpinnings take different forms, a "one-size-fits-all" developmental assessment approach rarely satisfies educational needs. Much of the strength of the BAS is in providing tools to model many different kinds of learning theories and learning domains. What is to be measured and how it is to be valued in each BAS application is drawn from the expertise and learning theories of the teachers, curriculum developers, and assessment developers involved in the process of creating the assessments.

Building block 1: Construct maps. Construct maps (Wilson, 2005) embody this first of the four principles: that of a developmental perspective on the assessment of student achievement and growth. A construct map is a well thought-out and researched ordering of qualitatively different levels of performance, focusing on one characteristic, such as, say, conceptual understanding of floating and sinking (Kennedy & Wilson, 2007). Thus a construct map defines what is to be measured or assessed in terms general enough to be interpretable within a curriculum, and potentially across curricula, but specific enough to guide the development of the other components of the BAS (which are described below). When instructional practices are linked to the construct map, the construct map also indicates the aims of teaching. Construct maps are one model of how assessments can be integrated with instruction and accountability systems. They provide a way for large-scale assessments to be linked in a principled way to what students are learning in classrooms, while at least having the potential to remain independent of the content of a specific curriculum. In the simplest case, a uni-dimensional outcome progression can be thought of as a construct map.

This approach assumes that student performance can be traced over the course of a given curriculum, facilitating a developmental perspective on student learning. Assessing the growth of students' understanding of particular concepts and skills requires a model of how student learning develops over a certain period of (instructional) time. A growth perspective helps one to move away from "one shot" testing situations and cross-sectional approaches to defining student performance toward an approach that focuses on the process of learning and on an individual's progress through that process. Clear definitions of what students are expected to learn and a theoretical framework for how that learning is expected to unfold as the student progresses through the instructional material (i.e., in terms of learning performances) are necessary to establish the construct validity of an assessment system.

The idea of using construct maps as the basis for assessments offers the possibility of gaining significant efficiency in assessment. Although each new curriculum prides itself on bringing something new to the subject matter, in truth, most curricula are composed of a common stock of content. And, as the influence of national and state standards increases, emphasis on common content is likely to increase. Thus we might expect innovative curricula to have one or perhaps even two constructs that do not overlap with typical curricula, but the remainder will form a fairly stable set of constructs that will be common across many curricula.

Construct maps are derived in part from research into the underlying cognitive structure of the domain and in part from professional judgments about what constitutes higher and lower levels of performance or competence. These maps are also informed by empirical research on how students respond to instruction or perform in practice (NRC, 2001). To more clearly understand what a construct map is, consider a construct map that focuses in particular on earth science knowledge in the area of Earth in the Solar System (ESS; Briggs, Alonzo, Schwab, & Wilson, 2006). The standards and benchmarks for ESS appear in Appendix A of the Briggs et al. (2006) article. According to these standards, by the eighth grade, students are expected to understand three different phenomena within the ESS domain—(1) the day/night cycle, (2) the phases of the Moon, and (3) the seasons—in terms of the motion of objects in the Solar System. A complete scientific understanding of these three phenomena is the top level of the construct map. In order to define the lower levels of the construct map, Briggs and his colleagues reviewed the literature on student misconceptions with respect to ESS. Documented student misconceptions with respect to the day/night cycle, the phases of the Moon, and the seasons are displayed in Appendix A of the Briggs et al. (2006) article.

In this work, the goal was to create a single continuum that could be used to describe typical students' understanding of the three phenomena within the ESS domain. In contrast, much of the existing literature documents students' understandings about a particular ESS phenomenon without making connections between understandings of related ESS phenomena. By examining student conceptions across the three phenomena and building on the progressions described by Vosniadou and Brewer (1994) and Baxter (1995), Briggs and his colleagues initially established a general outline of the construct map for student understanding of ESS. This general description helped them impose at least a partial order on the variety of student ideas represented in the literature. However, the levels were not fully defined until typical student thinking at each level could be specified. This typical student understanding is represented in the ESS construct map shown in Figure 2 (a) by general descriptions of what the student understands and (b) by limitations to that thinking in the form of misconceptions, labeled as "common errors," that help to clarify the difference between levels. Common errors in one level are resolved in the next level of the construct map. For example, students at Level 3 think that it gets dark at night because the Earth goes around the Sun once a day—a common error for Level 3—while students at Level 4 no longer believe that the Earth orbits the Sun daily but rather understand that this occurs on an annual basis.

The top level of the ESS construct map represents the understanding expected of eighth graders by national standards documents. Because students' understanding of ESS develops over several years of science instruction, it was important that the same continuum be used to describe the understandings of both fifth and eighth grade students. However, the top level is not expected of fifth graders; equally, we do not expect many eighth grade students to fall into the lowest levels of the continuum. Note that although the account of ESS given above might seem quite simple and straightforward, it is in fact the result of over two years of work, going well beyond the literature review mentioned above and including feedback from the next three building blocks, as described below, through several iterations as new waves of data were collected.

The main thrust of this chapter is to postulate a family of statistical models that can capture the relationship (or, rather, potential relationships) between multiple construct maps and an outcome progression. In order to help the reader see that the concept of a construct map (or its equivalent) is central to the successful development of assessments, I include below a description of the other parts of the BAS.

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: • the day/night cycle • the phases of the Moon (including the illumination of the Moon by the Sun) • the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that • the Earth is both orbiting the Sun and rotating on its axis • the Earth orbits the Sun once per year • the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky • the Moon orbits the Earth once every 28 days, producing the phases of the Moon COMMON ERROR: Seasons are caused by the changing distance between the Earth and Sun. COMMON ERROR: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the Moon.
3	Student knows that: • the Earth orbits the Sun • the Moon orbits the Earth • the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. COMMON ERROR: It gets dark at night because the Earth goes around the Sun once a day.
2	Student recognizes that: • the Sun appears to move across the sky every day • the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. COMMON ERROR: All motion in the sky is due to the Earth spinning on its axis. COMMON ERROR: The Sun travels around the Earth. COMMON ERROR: It gets dark at night because the Sun goes around the Earth once a day. COMMON ERROR: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. COMMON ERROR: It gets dark at night because something (e.g., clouds, the atmosphere, "darkness") covers the Sun. COMMON ERROR: The phases of the Moon are caused by clouds covering the Moon. COMMON ERROR: The Sun goes below the Earth at night.
0	No evidence or off-track

Figure 2. Construct map for student understanding of Earth in the Solar System. From "Diagnostic Assessment with Ordered Multiple-Choice Items," by D. C. Briggs, A. C. Alonzo, C. Schwab, and M. Wilson, 2006, Educational Assessment, 11, p. 42. Copyright 2006. Reprinted by permission of the publisher (Taylor & Francis Ltd), granted by Copyright Clearance Center Inc. on behalf of Taylor and Francis.

Principle 2: Match between Instruction and Assessment

So far, the main motivation for construct maps has been to provide a common framework for assessments and curricula and a method of making measurement possible. However, this second principle makes clear that the framework for assessments and the framework for curriculum and instruction must be one and the same since the connection must go beyond just the common basis and must extend to common contexts and styles that are involved in instruction.

Building block 2: The items design. The items design governs the match between instruction and various types of assessment. Note that the grain-size of "instruction" may vary from context to context, from specific classroom practices in a curriculum to broad specifications of local or state curricula in "standards" documents—the ESS example is more in line with the latter. The critical element to ensure the match between instruction and assessment in the BAS is that each assessment task and typical student responses are matched to levels of a construct map.

Returning to the ESS example, the items designed to test student understanding with respect to the construct map are distinctive, as they are Ordered Multiple-Choice (OMC) items, which attempt to use the cognitive differences built into the options to make for more valid and reliable measurement (Briggs et al., 2006). Following the BAS principles, OMC items were written as a function of the underlying construct map, which is central to both the design and interpretation of the OMC items. Item prompts were determined by both the domain, as defined in the construct map, and canonical questions (i.e., those that are cited in standards documents and commonly used in research and assessment contexts). In this instance, the instruction is somewhat distal, as it is mainly expressed through curriculum documents (i.e., standards) rather than instructional documents. The ESS construct map focuses on students' understanding of the motion of objects in the Solar System and explanations for observable phenomena (e.g., the day/night cycle, the phases of the Moon, and the seasons) in terms of this motion. Therefore, the ESS OMC item prompts focus on students' understanding of the motion of objects in the Solar System and the associated observable phenomena. Distractors were written to represent (a) different levels of the construct map, based upon the description of both understandings and common errors expected of a student at a given level, and (b) student responses that were observed from open-ended versions of the items. Two sample OMC items, showing the correspondence between response options and levels of the construct map, are shown in Figure 3. Each item response option is linked to a specific level of the construct map. Thus, instead of gathering information solely related to student understanding of the specific context described in the question, OMC items allow us to link student answers to the larger ESS domain represented in the construct map. Taken together, a student's responses to a set of OMC items permit an estimate of the student's level of understanding and provide diagnostic information about the student's understanding of a specific phenomenon.

Item appropriate for fifth graders:

It is most likely colder at night because

A. the Earth is at the furthest point in its orbit around the Sun.	Level 3
B. the Sun has traveled to the other side of the Earth.	Level 2
C. the Sun is below the Earth and the Moon does not emit as much heat as the Sun.	Level 1
D. the place where it is night on Earth is rotated away from the Sun.	Level 4

© WestEd, 2002

Item appropriate for eight graders:

Which is the best explanation for why we experience different seasons (winter, summer, etc.) on Earth?

A. The Earth's orbit around the Sun makes us closer to the Sun in summer and farther away in the winter.	Level 4
B. The Earth's orbit around the Sun makes us face the Sun in the summer and away from the Sun in the winter.	Level 3
C. The Earth's rotation on its axis makes us face the Sun in the summer and away from the Sun in the winter.	Level 3
D. The Earth's tilt causes the Sun to shine more directly in summer than in winter.	Level 5
E. The Earth's tilt makes us closer to the Sun in the summer than in the winter.	Level 4

Figure 3. Sample OMC items based upon the Earth in the Solar System construct map. From "Diagnostic Assessment with Ordered Multiple-Choice Items," by D. C. Briggs, A. C. Alonzo, C. Schwab, and M. Wilson, 2006, Educational Assessment, 11, p. 43. Copyright 2006. Reprinted by permission of the publisher (Taylor & Francis Ltd), granted by Copyright Clearance Center Inc. on behalf of Taylor and Francis [1]

Principle 3: Management by Teachers[2]

For information from the assessment tasks to be useful to instructors and students, it must be couched in terms that are directly related to the instructional goals behind the construct maps. Open-ended tasks, if used, must be quickly, readily, and reliably scorable.

Building block 3: The outcome space. The outcome space is the set of categorical outcomes into which student performances are categorized for all the items associated with a particular construct map. In practice, these are presented as scoring guides for student responses to assessment tasks. This is the primary means by which the essential element of teacher professional judgment is implemented in the BAS. The scoring guides are supplemented by "exemplars"—examples of student work at every scoring level of the construct map for every task—and "blueprints," which provide the teachers with a layout showing opportune times in the curriculum to assess the students on the different constructs. In the ESS example, exemplars were used to create the initial set of options for the OMC items (such as those shown in Figure 3). Empirical evidence was then used to revise these options during pilot- and field-testing.

Principle 4: Evidence of High Quality Assessment

Technical issues of reliability and validity, fairness, consistency, and bias can quickly sink any attempt to measure student performance with respect to a construct map, or even to develop a reasonable framework that can be supported by evidence. To ensure comparability of results across time and context, procedures are needed to: (a) examine the coherence of information gathered using different formats; (b) map student performances onto the construct map; (c) describe the elements of the accountability system—tasks and raters; and (d) establish uniform levels of system functioning, in terms of quality control indices such as reliability. Two ways that these procedures could be embodied in a measurement model are: (a) through a latent continuum (i.e., by placing cuts along the continuum to represent different levels of the construct) and (b) through an ordered set of latent classes. A latent continuum does not directly embody the ordering of the construct map levels explicitly as part of the parametric structure of the model—the ordering is developed in a second step, where segments of the continuum are labeled as indicating levels of the construct. This is the approach typically used in the BAS, which utilizes a Rasch-family item response modeling framework (Wilson, 2005) that parameterizes the item difficulty and the person ability in a way that allows one to report results using a visual representation known as the Wright map (see below). However, one might want to explicitly embody the levels in a model (NRC, 2001), and this could be realized through an ordered latent class approach (e.g., Vermunt & Magidson, 2008).

Building block 4: Visual representation of empirical results—Wright maps. Wright maps are one form of visual representation of the results of an estimation of item and person parameters; thus they represent this principle of evidence of high quality. Note that the maps represent this principle but do not themselves achieve it—there still remains the accumulation and documentation of validity and reliability evidence (as described, say, in Chapters 7 and 8 of Wilson's 2005 book). Wright maps can be used to show how the construct unfolds or evolves in terms of increasingly sophisticated student performances. For example, consider the hypothetical Wright Map shown in Figure 4. The Xs on the left side of the vertical line show the students' locations (as a histogram on its side). The numbers from 0 to 5 in the first five columns on the right side show the thresholds[3] for the multiple-choice options (labeled by their respective construct levels) for specific items. The Roman numerals on the far right show regions of the continuum relating to each construct level: these have been set in a post-estimation step to enclose the thresholds for the items.[4] These regions are thus a technical representation of the empirical results regarding the levels of the construct. The ordering, and even the ability to identify these regions, constitutes evidence for internal structural validity of the hypotheses set up in the construct map (Wilson, 2005, Chapter 7). The regions are not themselves parametrically represented in the statistical models used to estimate the person and item locations shown in the Wright maps.

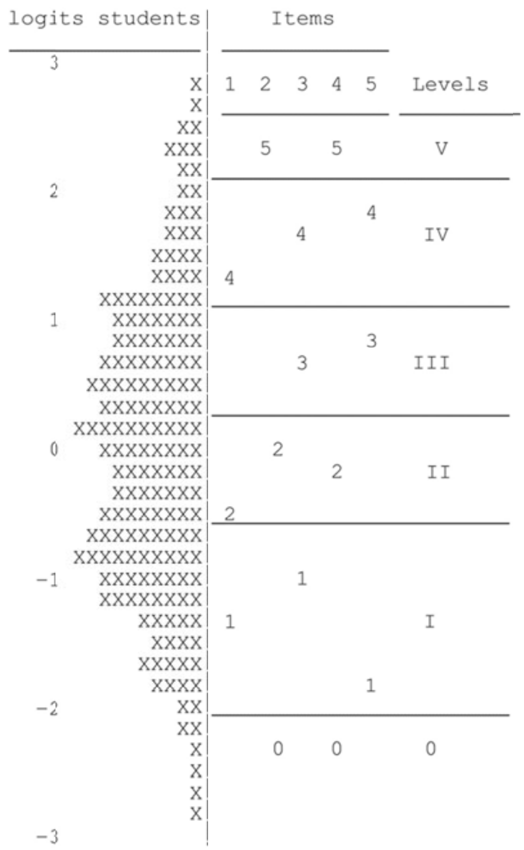


Figure 4. An illustrative Wright Map for the ESS construct.

We typically use a Rasch modeling approach to calibrate the maps for use in the BAS (see Adams, Wilson, and Wang, 1997, and Briggs and Wilson, 2003, for the specifics of this model). These maps have at least two advantages over the traditional method (i.e., classical test theory and/or norm-referenced approach) of reporting student performance as total scores or percentages. First, they allow teachers to interpret a student's proficiency in terms of average or typical performance on representative assessment activities without students having to take the same items (as each other or over time). Second, they estimate the relative difficulties of the tasks involved in assessing student proficiency, which are then used to generate the Wright map—an empirical version of the construct map. Later in this chapter, I will use an extension of this approach to modeling in order to integrate hypotheses about links among constructs within an outcome progression. This will alter some statistical aspects of the modeling, but many of the types of analyses and results will remain similar.

MAPPING OUT AN OUTCOME PROGRESSION USING CONSTRUCT MAPS

In the previous sections of this chapter, there was an implicit assumption that the relationship between the construct map and the outcome progression is one-to-one; in other words, the assumption that a single construct map adequately summarizes the outcome progression. However, the single construct map may have a certain degree of complexity, such as was illustrated by the ESS example (Figure 2), where levels of the construct map are divided into separate sub-categories for different misconceptions.

In addition, an outcome progression may be comprised of multiple construct maps, such that the relationship between construct maps and an outcome progression can be quite a bit more complex. (See Wilson, 2008, and Wilson, 2009, for examples of relationships between the construct maps and the outcome progression beyond those discussed here.) For instance, there could be an assumption that certain constructs lead to one construct rather than to another. This could be illustrated as in Figure 5. Here, the attainment of levels of a construct would be seen as dependent on the attainment of specific "precursor" constructs. An example of such thinking, this time in the case of the Molecular Theory of Matter for the middle school level—under development with Paul Black of King's College, London—is shown in Figure 6 (Wilson & Black, 2007). In this example, each box can be thought of as a construct map, and the relationship between them is specified by arrows. In particular, the Density and Measurement and Data Handling constructs provide important resources for the main series of constructs: Properties of Objects, Properties of Atoms and Molecules, Conservation and Change, and Molecular Theory of Macro Properties.

A more complicated way of seeing the relationship between construct maps is shown in Figure 7, where there are links hypothesized between specific levels of one construct and specific levels of other constructs (rather than the "top to bottom" relationships shown in Figure 5). An example of such a diagram, showing a finer grain of connections, is shown in Figure 8 for the Molecular Theory of Matter outcome progression. Note, for example, that there is a hypothesized link from the upper level of Properties of Atoms and Molecules to the middle level of Molecular Theory of Macro Properties. There is also a hypothesized link from the lower level of Properties of Objects to the upper level of Conservation and Change. These hypotheses about links are based on the research literature and also on professional judgment.

STRUCTURED CONSTRUCTS MODELS (SCMS)

With respect to the measurement models that one would use to model the data arising from assessments based on the construct map structures described above, a great deal will depend on the nature of the construct maps and the hypothesized links among them. Statistically speaking, the most common frameworks are essentially comprised of correlated dimensions, so that a multidimensional item response model (Adams et al., 1997) would be suitable. However, the approach in Figure 5 would constitute a variant of structural equation models (SEMs[5]: see

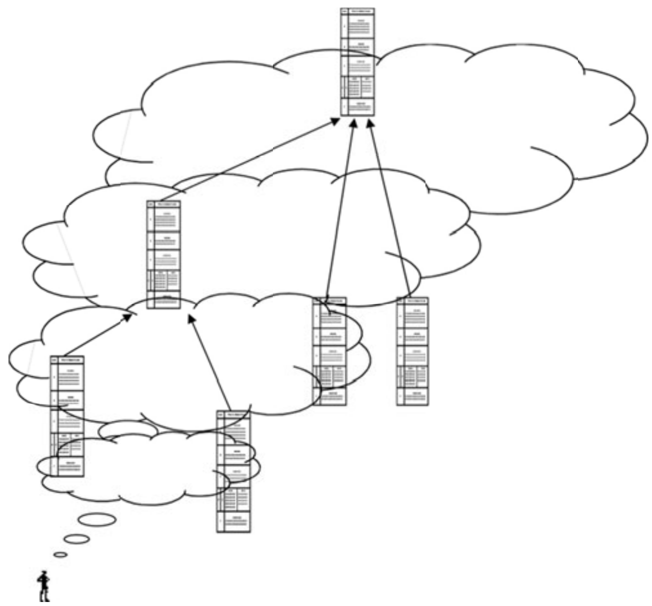


Figure 5. In this situation, there is a complicated dependency relationship between the construct maps in the outcome progression. From "Measuring Progressions: Assessment Structures Underlying a Learning Progression," by M. Wilson, 2009, Journal for Research in Science Teaching, 46, p. 727. Copyright 2009 by Wiley Periodicals, Inc. Reproduced with permission of Wiley Periodicals, Inc. via Copyright Clearance Center.

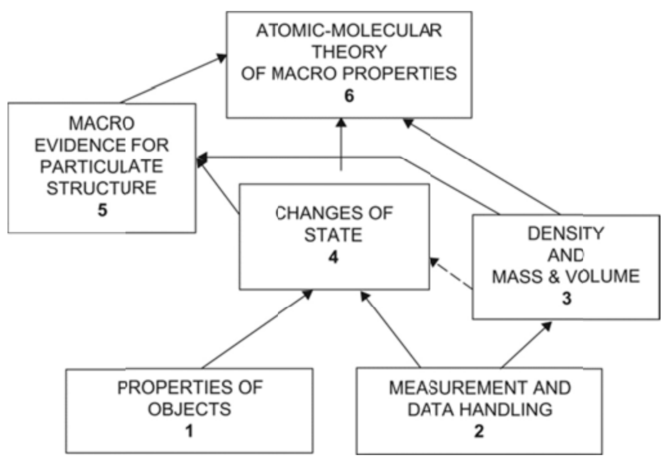


Figure 6. A set of constructs hypothesized to constitute a Molecular Theory of Matter. From "Roadmaps for Learning: A Guide to the Navigation of Learning Progressions," by P. Black, M. Wilson, and S.-Y. Yao, 2011, Measurement: Interdisciplinary Research and Perspectives, 9, p. 83. Copyright 2011 by Taylor & Francis Ltd. Reproduced with permission of Taylor & Francis via Copyright Clearance Center

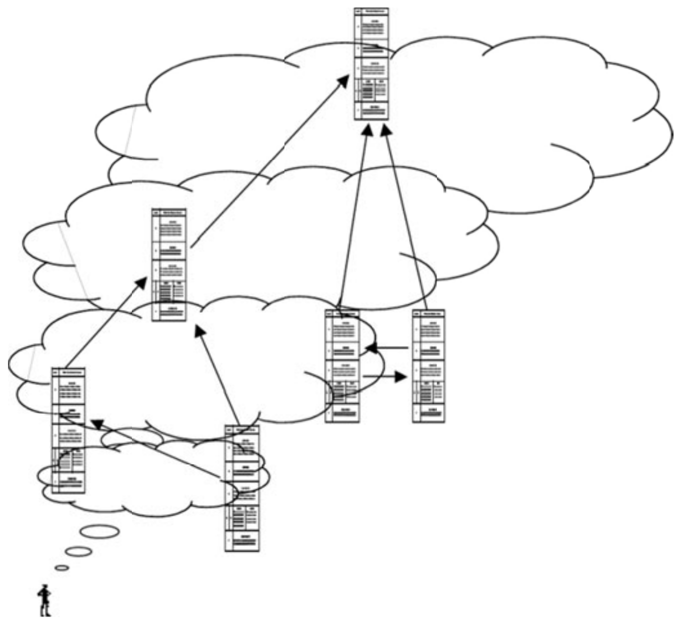


Figure 7. In this situation, the relationship between the construct maps is from level to level.

Bollen, 1989, for a comprehensive introduction). In this SEM-like view, each of the construct maps would be a single continuous SEM variable, and the arrows between construct maps would be the usual SEM paths, effectively representing regression relationships between the lower (predictor) variable and the higher (criterion) variable. This describes the relationships depicted in Figures 5 and 6; the detailed arrows running between particular levels (as in Figures 7 and 8) are not attended to in the SEM approach. In contrast, the SCM approach, in which the arrows arise from and point to levels inside the boxes, constitutes a more complicated model that goes beyond the usual SEM formulation. In this model, the paths run not between the constructs as a whole (i.e., between the SEM variables) but between specific levels within the constructs.

There are various ways that one might formally model what diagrams such as those in Figures 7 and 8 represent. In many circumstances, it can be reasonable to hypothesize that reaching a certain level on one construct map is a necessary (or perhaps a favored) prelude to reaching a certain level on a second construct map but not through a continuous relationship as for SEM; instead, the relationship may exist only for a particular pair of levels. For example, in the Molecular Theory of Matter example introduced above (Figure 8), there is a connection from the highest level of the Density construct map to the mixtures level of the Molecular Theory of Macro Properties (MTMP) construct map. This connection does not pertain to any other level of the Density construct; nor does it lead to any other level of the MTMP construct.

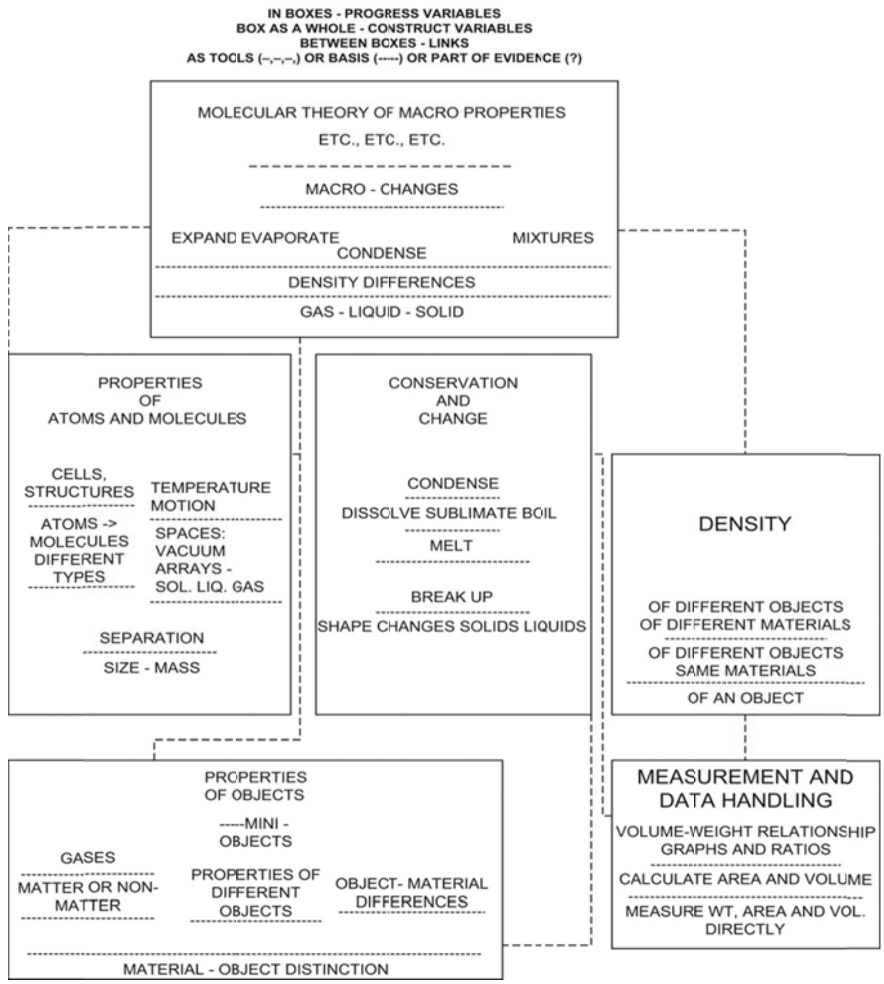


Figure 8. A more detailed version of the relationships shown in Figure 6.

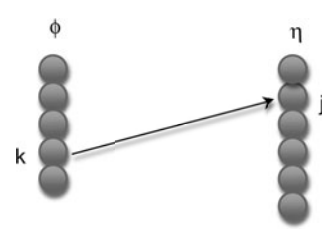


Figure 9. Diagram of a simple structured constructs model (SCM).

One way to mathematically represent this would be to see each of these two levels of the two different construct maps as one of a series of ordered latent classes (i.e., the levels) for each construct, where the regression relationship relates to the probability of a person being in the two specific latent classes—i.e., being in one level of the first construct (Density) makes it more likely that a person would be in a particular level of the second construct (MTMP). This is illustrated in Figure 9, where the small ovals represent the successive levels of each construct and the specific link is from level j of the first construct to level k of the second. (Of course, there could be more such links, but I will illustrate with just one.) I will call models, such as the example in Figure 9, with (a) more than one set of ordered latent classes (i.e., more than one construct map) and (b) hypothesized links between certain of the latent classes across those sets, structured constructs models (SCMs). See Appendix B for one way to specify a SCM using a statistical equation approach.

The first difference from the simple construct map formulation is that the underlying model for each construct is seen as a series of latent classes rather than as a continuum. The reason for the choice of a latent class approach in this case is that it allows qualitative a priori specification of hypothesized links between levels of different constructs. It is not clear to me how to achieve this in a latent continuum approach.

The estimation of a simple ordered latent class model for each of the constructs (Density and MTMP) without the link from *j* to *k* is quite straightforward and has been described in the literature already; it is simply a double case of ordered latent class models (e.g., Mokken, 1971; Mokken & Lewis, 1982). I will label this as the "BASE" model. However, the extra link adds complexity in that the link could be expressed in several different ways. For example, the probability of being at level *k* of MTMP could be modeled as a known function (i.e., a function of known form with certain parameters to be estimated) of the probability of being at level *j* of Density (as in Equation 1 of Appendix B). This relationship, once postulated, and, if estimated to be non-null, would, in turn, affect the estimation of the parameters of the standard latent class model estimated without the hypothesized link between *j* and *k*. This new model, the SCM, would fit the sample data to an extent different from that of the BASE model, and this difference in fit could then be used as the basis for an evaluation of the two models.[6] If there is no statistically significant difference, one could conclude that the *j*-to-*k* link was not implied by the data; conversely, a statistically significant difference would indicate that such a link is implied by the data. Of course, one would also need to develop an index of the effect-size of that link. This approach allows one to develop ways to test aspects of the fit of the postulated outcome progression to a data set that one has collected. However, if there are many hypothesized links, then testing each one as a direct model-to-model comparison can be time consuming and laborious; in addition, the order in which the models should be tested can be unclear (analogous to an issue that arises in testing effects in multiple linear regression). Other formulations are possible, such as postulating a relationship between the underlying latent class parameters for level *k* of MTMP and level *j* of Density. This would also facilitate the use of a simple test of these link parameters by seeing if their confidence intervals included 0.0, thus avoiding the problem of the ordering of the different models.

In the account so far, I have not specified how the observations of item responses would be related to the ordered latent classes. This would be described within a SCM by a function involving (at least one) parameter β_{is} that would control the probability of a person giving a certain response to item *i*, given membership in latent class *s* of a given construct, such as the Density or MTMP construct. Constraints among these parameters might also be used to describe the hypotheses that one might have about the relationship between the latent classes and the levels of response to the items, such as the possibility of "guessing," or "slippage" that might occur at levels above or below the person's estimated location. That is, a person at level *j* might still respond to an item as though he or she were at a higher level ("guessing") or at a lower level ("slippage"). When these relationships, along with the relationships linking levels of the construct maps, are fully specified, then the SCM can be written in a form that is a special case of a Generalized Linear Mixed Model (GLMM; Breslow & Clayton, 1993) and estimated in a number of ways (see Appendix B).

Note that the result from such an estimation will be a set of probabilities for each combination of person and ordered latent class (i.e., level). Even though the fundamental assumption is that each person is in just one level, we still get probabilities that express how much we believe they belong to each level. This will allow a visual representation that is analogous to the Wright Map mentioned above but that does not have the same metric properties as that representation since there is no analogy here of distance along the map to probability of answering an item correctly. The discrete nature of the latent classes do not allow such an interpretation. Instead, the classes are simply ordered as in the left side of Figure 9. Thus the SCM map will look far more like the construct map itself, so long as there was no evidence in the results that the construct map should be changed.

These results can be useful not only in identifying people who are clearly at a specific level but also people whose responses show them to be "misfitting" the model (in a way that is quite parallel to the use of "fit" and "misfit" in item response modeling). Thus there may be people who are roughly split between two levels, and there may be people who seem to have no strong tendency for a single level. Of course, when there are large numbers of such people who misfit, then we would consider that the misfit information was not just informative about those individuals but might also be informative about the specified model. This could lead to alterations in the construct map and/or the SCM, depending upon the specific patterns of misfit.

DISCUSSION AND CONCLUSION

In this chapter I have tried to outline a measurement response to the challenge posed by the complexity of the concept of an outcome progression. This has been done from a particular perspective, that of the construct map that forms the heart of the BAS. I make no excuses about this focus because, as discussed in Wilson (2009), even taking such a particularistic view, there are a great many ways that the construct map concept could be deployed to give structure and form to assessments to support an outcome progression. Other measurement approaches could also be used, but these would require separate development in separate papers. Laying out these possibilities is helpful to thinking about the possible issues and limitations of each approach.

One interesting question is whether the proposed SCM formulation is indeed a step forward from the venerable SEM approach. Indeed, it would be interesting to compare the results from the two modeling approaches. This could not be done in a technically precise manner within the classical SEM framework, in which the modeling is based on the variance-covariance matrix, but it could be carried out in an approximately similar way using a full item response modeling approach as is available in *gllamm* (Rabe-Hesketh, Pickles, & Skrondal, 2001) or in M-PLUS (Muthén, & Muthén, 1998-2007) software. Under these circumstances, the SEM solution would provide a robust "baseline" model for the SCM, allowing one to assess the value added by the preciseness of the level-to-level hypothesized links as compared to the more relaxed variable-to-variable links in the SEM approach. A SEM solution could also be seen as a precursor to a more fine-grained SCM solution, allowing for rapid estimations, and (perhaps) pilot work done on smaller samples. Thus the two approaches could have complementary purposes.

The ideas about SCMs that have been presented here can be extended to other complexities of the assessment context. Clearly, the simple SCM described above, involving a single connection between two constructs, can be extended to multiple connections and more constructs. Integration of SCM and SEM types of models would allow one to entertain the possibility that some constructs are composed of (ordered) latent classes while, at the same time, others are composed of continua. A more subtle extension would entertain the idea that some constructs could be better represented as simultaneously both class-like and continuum-like. One way to interpret that would be to see that the latent classes represented in Figure 9 could, in addition to the current conception, be arrayed along a continuum so that the "distance" between them becomes meaningful in a probabilistic sense. This goes beyond the current conception, in which the latent-class representation of the levels means there is no concept of relative distance between each of the successive levels. Similar possibilities have been considered in related domains, such as in the modeling of stage-like development, where specific models such as the Saltus model have been developed and applied in areas such as Piagetian studies (Wilson, 1989; Draney & Wilson, 2007). The development of the SCM, as shown in Appendix B (Equations 1 through 6), has been undertaken explicitly with the intention of making the extensions described in this paragraph.

Although the example SCM that is discussed in this chapter is (intentionally) a very simple one, it is possible to imagine a great many possible ways that construct maps could be deployed to support learning progressions with more complex outcome progression structures, including those that encompass multiple construct maps with detailed connections between levels. This flexibility is important since one would not want to have the potential usefulness of a learning progression constrained by the underlying conceptualization of construct maps.

It is also clear that there are some important decisions that will need to be made when designing a SCM appropriate to a given assessment structure (i.e., to the structure of a particular outcome progression). Being aware of the range of possibilities described here, as well as possibilities beyond these (some of which were briefly mentioned above), will help the developers of a learning progression consider the form they want their outcome progression to take, and, especially, how they will relate it to the assessments they will use. Considering whether one would prefer an assessment structure that is best thought of as a multidimensional item response model, as an SCM, or as something in-between, will be an important step in developing a new learning progression or in modifying an existing one.

These choices will also have important ramifications for the other building blocks, including the items design and the outcome space. This chapter has really just scratched the surface of an important aspect of the application of measurement ideas in science education in particular, and, potentially, across the whole range of educational achievement. Uni-dimensional and, more recently multidimensional, item response models have been a mainstay of measurement in educational achievement domains. Seeing how these models can be extended into the complex areas allowed by SEM-like approaches, and the more subtle SCM approaches described above, will be an interesting and challenging task in the future.

NOTES

1 The second item has been revised from the original for use in ongoing work on the ESS construct map (as reported in Briggs & Alonzo, 2009).

2 Other educational professionals may also use the assessments—we emphasize teachers' roles since we see them as the main users of the assessments.

3 The threshold for level k, say, is the point where a student is 50% likely to respond at level k or below (and equivalently, not to level k+1 or above).

4 I am skipping over quite a lot of detail here (e.g., how the segment boundaries are set——see, for instance, Kennedy and Wilson, 2007). In the process of setting the boundaries, many complications may arise—the levels may come out in a different order from that predicted, or the boundaries may be too confused to be readily "set"—but, for illustrative purposes, none of that is shown here.

5 For those not familiar with structural equation models, see a very brief conceptual introduction to this topic in Appendix A.

6 Specifically, because the two models differ only in that the SCM model has the parameters needed to formulate the *j*-to-*k* link, the BASE model is formally a special case of the SCM model; it is the SCM model with those parameters set to make the link non-active. Thus a likelihood ratio test could be performed.

REFERENCES

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Andersen, S. K., Olesen, K. G., Jensen, F. V., & Jensen, F. (1989). HUGIN - A shell for building Bayesian belief universes for expert systems. In N. S. Sridharan (Ed.), *Proceedings of the 11th International Joint Conference on Artificial Intelligence* (pp. 1080–1085). San Mateo, CA: Morgan Kaufmann.

Baxter, J. (1995). Children's understanding of astronomy and the earth sciences. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 155–177). Mahwah, NJ: Lawrence Erlbaum Associates.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I. Cognitive domain*. New York, NY: Longmans, Green.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Briggs, D., & Alonzo, A. (2009, April). *Building a learning progression as a cognitive model.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Briggs, D., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87–100.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Draney, K, & Wilson, M. (2007). Application of the Saltus model to stage-like data: Some applications and current developments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 119–130). New York, NY: Springer.

Haladyna, T. M. (1994). Cognitive taxonomies. In *Developing and validating multiple-choice test items* (pp. 104–110). Hillsdale, NJ: Lawrence Erlbaum Associates.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Kennedy, C. A., & Wilson, M. (2007). Using progress variables to map intellectual development. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in schools: Intellectual growth and standard setting.* Maple Grove, MN: JAM Press.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10,* 325–337.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212.

Masters, G. N., Adams, R. J., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies* (Supplementary vol. 2, pp. 628–634). Oxford, UK: Pergamon Press.

Masters, G., & Forster, M. (1996). *Progress maps: Assessment resource kit.* Victoria, Australia: Commonwealth of Australia.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Berlin, Germany: Mouton/DeGruyler.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

National Research Council. (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). *GLLAMM manual. Technical report 2001*. London, UK: Department of Biostatistics and Computing, Institute of Psychiatry, King's College.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The Fusion Model skills diagnostic system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge, UK: Cambridge University Press.

Vermunt J. K.,& Magidson J. (2008). *Latent Gold 4.5 user's guide*. Belmont, MA: Statistical Innovations Inc.

Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science*, *18*, 123–183.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2008, March). *Measuring progressions*. In A. C. Alonzo & A. W. Gotwals (Chairs), *Diverse perspectives on the development, assessment, and validation of learning progressions in science*. Symposium conducted at the annual meeting of the American Educational Research Association, New York, NY.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching, 46*, 716–730.

Wilson, M., & Black, P. (2007, March). The idea of a learning progression as a core for both instruction and assessment. In A. C. Alonzo (Chair), *Assessing understanding through the use of learning progressions*. Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181–208.

Wright, S. S. (1921). Correlation and causation. *Journal of Agricultural Research, 20*, 557–585.

APPENDIX A

A Very Brief Summary of Structural Equation Models (SEMs)

Structural equation modeling is a branch of statistical modeling for postulating and testing assumed causality. It was initially described by Wright (1921). A modern introduction has been offered by Bollen (1989). Pearl (2000) has given a new formulation using a calculus of counterfactuals.

Structural equation models (SEMs) can be used in both confirmatory and exploratory modeling, meaning that both theory testing and theory development are possible. Confirmatory modeling begins with a hypothesis that is summarized in a graphical causal model such as is shown in Figure A1. The boxes in Figure A1 represent variables that are connected causally by the hypothesis—the arrows indicate the hypothesized causes. In the example, variable C is hypothesized to be causally related to variables A and B via the causal links *a* and *b*, respectively. In the equivalent statistical model, these links, *a* and *b*, would be instantiated as parameters in the model. C is called a predictor or source variable, and A and B are called outcome or downstream variables. Items must be written to measure the variables in the model to allow testing of the relationships between the concepts in the model. The arrows are operationalized as regression equations connecting the predictor variables with the outcome variables (effectively, there will be one equation for each downstream variable). The causal assumptions of the model often have falsifiable implications that can be tested against data to determine model fit.

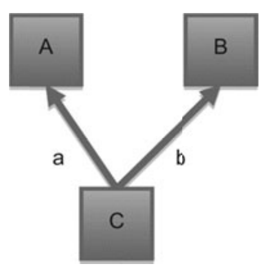


Figure A1. An example of a graphical representation of a structural equation model (SEM).

SEM allows for latent variables: that is, variables that are not measured directly but are estimated in the model from several measured variables (i.e., responses to items) that are assumed to tap into the latent variables. Thus SEM can incorporate the unreliability of measurement in the model, which in theory allows the structural relations between latent variables to be accurately estimated. Factor analysis, path analysis, and regression all represent special cases of SEM; item response modeling can be thought of as a version of SEM using non-linear relations.

In interpreting the graphical representation of a SEM, it is important to realize that it is the arrows that are not shown that represent the strongest assumptions—for those arrows, the assumption is that there is no causal relation between the variables.

Comparing Figure A1 with Figure 4, we can see that the arrows and boxes are being used in analogous ways. Note that, in order to simplify the diagrams, the figures in this chapter depict only the variables in the hypothesized models. The items that would be used to measure the variables are not depicted, although, indeed, in practice, they would be needed also.

APPENDIX B

Development of the Structured Constructs Model (SCM)

To write the situation illustrated in Figure 9 as an equation, consider each of the constructs as represented by an ordered set of latent classes (corresponding to the levels of the construct map for each). We will label the Density construct by ϕ and the MTMP construct by η. Then the ordered sets of latent classes are $\{\phi_1, \phi_2, \dots \phi_R\}$ (where the levels are indexed by *r* from 1 to *R*) for the first (predictor) construct, and $\{\eta_1, \eta_2, \dots \eta_D\}$ (where the levels are indexed by *d* from 1 to *D*) for the second (criterion) construct. The individual student *n* is characterized by an indicator vector for group membership in each construct $\underset{\sim}{\phi_n}$ and $\underset{\sim}{\eta_n}$. For the R potential person groups the indicator vector for the first construct is $\underset{\sim}{\phi_n} = (\phi_{nl}, \dots, \phi_{nR})$, where ϕ_{nr} takes the value of 1 if the person n is in group r and 0 if not, and similarly for the second construct. The model assumes that each person belongs to one and only one ϕ group and one and only one η group; thus, only one of the ϕ_{nr} and only one of the η_{nd} is theoretically nonzero. Consistent with the term "latent," the values of neither $\underset{\sim}{\phi_n}$ nor $\underset{\sim}{\eta_n}$ are observable.

Then the link between two specific levels, say *j* and *k*, of the first and second constructs, respectively, is given by an equation of the form:

$$\Pr(\eta_{nj} = 1) = f_{jk}(\Pr(\phi_{nk} = 1)). \tag{1}$$

The function f_{jk} may take a variety of forms, the simplest being a linear relationship. I will use *F* to designate the entire set of possible functions f_{jk}; in this simple case, only one is not null. This is what I call a structured constructs model (SCM). There could be many ways to hypothesize more complex versions of this type of model, but I will stick to the simple example in Figure 9 for now.

What is not yet clear is how the observations of item responses relate to these constructs. Suppose that we have polytomous item responses X_i and Y_h that relate to the first and second constructs respectively, where *i*=1, 2…I, and *h*=1, 2,…*H*. Note that, in terms of scoring, any one item (indexed *i* or *h*) may be sensitive to a subset of the levels for its respective construct (i.e., a subset of the levels 1 to *R* for ϕ and a subset of the levels 1 to *D* for η), but for simplicity, we assume that each item is sensitive to all possible latent classes of its construct. Of course, there could be more complex situations, where there are, for example, item bundles that relate to both constructs, but, again, I will stick to this simple formulation. (Note that I am not illustrating the relationship of the items to the constructs, as is commonly done in psychometric structural diagrams—this would make Figure 9 much more complicated.)

The probability of person n responding at level r of item i related to the first construct is given as:

$$P\left(X_{nir}=x_{nir} \mid \phi_{nk}=1, \underset{\sim}{\beta_i}\right)=\frac{\exp \sum_{s=1}^{r}\left(\beta_{is}\right)}{\sum_{t=1}^{H} \exp \sum_{s=1}^{t}\left(\beta_{is}\right)}, \tag{2}$$

where β_{ι} is a vector of latent class parameters, specific to item i, governing the allocation of persons to classes within the first construct ϕ. Similarly, the probability of person n responding at level d of item h for the second construct is given as:

$$P\left(Y_{nhd}=y_{nhd} \mid \eta_{nh}=1, \underset{\sim}{\delta_i}\right)=\frac{\exp \sum_{s=1}^{d}\left(\delta_{hs}\right)}{\sum_{d=1}^{D} \exp \sum_{s=1}^{d}\left(\delta_{hs}\right)}, \tag{3}$$

where δ_{h} is the corresponding set of parameters, specific to item h, for the second construct. In Equations 2 and 3, the index s is a dummy index, used just to count the classes.

Note that the parameters in the vector of latent class parameters $\underset{\sim}{\beta_i}=(\beta_{\mathrm{i1}},\ldots\ \beta_{\mathrm{is}},\ldots,\ \beta_{\mathrm{iR}})$ have specific interpretations. Assuming that $\phi_{\mathrm{nk}}=1$ (i.e., person n is indeed in level k), then, for $s = k$, β_{is} $(=\beta_{\mathrm{ik}})$ governs the probability that a person at level k will indeed give a response to item i that is also in level k. For $s>k$, β_{is} governs the probability that a person at level k will give a response to item i that is above level k. This might be interpreted as "guessing," which is how it is usually labeled in other latent class models—e.g., NIDA (Maris, 1999), DINA (Junker & Sijtsma, 2001), Fusion (Roussos et al., 2007). Similarly, for $s<k$, β_{is} governs the probability that a person at level k will give a response to item i that is below level k. This is called "slipping" in other latent class models.

Under some circumstances, it may be that the β_{is} parameters could be constrained in ways that correspond to the circumstances. For example, some researchers might postulate that there is no possibility for persons to respond at a higher level than k (i.e., "guessing" is absent). Then for $s>k$, one might set β_{is} = 0.0 or $\beta_{\mathrm{is}} = c_g$, where c_g is a value corresponding to a very small probability. This might then be modified such that this restriction applies for just the levels that are at least two above k, etc. Similarly for "slippage," one might set $\beta_{\mathrm{is}} = c_s$, where c_s is a value corresponding to a very small probability (setting β_{is} = 0.0 would seem to be less sensible here). Another circumstance might occur where the items are sufficiently similar that it would make sense to constrain the β_{is} parameters (other than β_{ik}) to be constant across items: i.e., $\beta_{\mathrm{is}} = \beta_{\mathrm{i}}$, for $s \neq k$. Under certain circumstances, it might make sense to constrain the parameters to be symmetric about $s=k$, although that seems unlikely in a cognitive context where "slippage" is seldom symmetric with "guessing."

Specification of a SCM such as that described above can be expressed under a Generalized Linear and Mixed Model (GLMM) framework. The probability that a person with group membership parameter $\underset{\sim}{\phi_n}$ will respond in category *r* to item *i* is given by:

$$P\left(X_{nir}=x_{nir}\mid\underset{\sim}{\phi_n},\underset{\sim}{\beta_i}\right)==\prod_g P\left(X_{nir}=x_{nir}\mid\phi_{ng}=1,\underset{\sim}{\beta_i}\right)^{\phi_{ng}} \tag{4}$$

Similarly, for the second construct, the probability that a person with group membership parameter $\underset{\sim}{\eta_n}$ will respond in category *d* to item *h* is given by:

$$P\left(Y_{nhd}=y_{nhd}\mid\underset{\sim}{\eta_n},\underset{\sim}{\delta_h}\right)=\prod_{g'} P\left(Y_{ndh}=y_{nhd}\mid\eta_{ng'}=1,\underset{\sim}{\delta_h}\right)^{\eta_{ng'}}. \tag{5}$$

As item responses are assumed to be independent given $\underset{\sim}{\phi_n}$, $\underset{\sim}{\eta_n}$, the item parameters and the link function parameters, *F*, the modeled probability of a response vector is:

$$\begin{aligned}P\left(\underset{\sim}{X_n}=\underset{\sim}{x_n},\underset{\sim}{Y_n}=\underset{\sim}{y_n}\mid\underset{\sim}{\phi_n},\underset{\sim}{\eta_n},\underset{\sim}{\beta},\underset{\sim}{\delta},\underset{\sim}{F}\right)&=\prod_g\prod_i\prod_r\left\{P(X_{nir}=x_{nir}\mid\phi_{ng}=1,\underset{\sim}{\beta_i})^{\phi_{ng}}\right\}^{x_{nij}}\\&\times\prod_{g'}\prod_h\prod_d\left\{P(Y_{nhd}=y_{nhd}\mid\eta_{ng'}=1,\underset{\sim}{\delta_h})^{\eta_{ng'}}\right\}^{y_{nhd}}\end{aligned} \tag{6}$$

subject to the constraints of Equation 1. Different programs, embodying different estimation approaches, can be used to implement this model—for example, LatentGOLD (Vermunt & Magidson, 2008), *gllamm* (Rabe-Hesketh, Pickles & Skrondal, 2001), HUGIN (Andersen, Olesen, Jensen, & Jensen, 1989) and WinBugs (Lunn, Thomas, Best & Spiegelhalter, 2000), with the last providing the greatest flexibility to estimate most complex versions of this family of models.

NARRATIVE STORYLINE

Initial Ideas

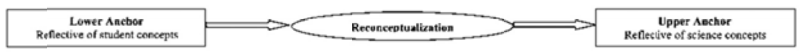
Before instruction children …

Conceptual Stepping Stones

Elementary students …
Middle school students …

Culminating Scientific Ideas

High school students …



CONCEPT & SKILL DETAILS

Initial Ideas	Conceptual Stepping Stones	Culminating Scientific Ideas
How do learners typically understand these concepts and what can they do (skills) *before instruction*? • Pre-, naïve-, or mis-conceptions • Common, experiential views of the world • These beliefs may often be incompatible with a scientific view (misconceptions), but be sure to articulate what students may know accurately (for example, many young children recognize that a bean seed planted in corn field will result in a bean plant, not a corn plant; they know that offspring look like their parents) • Be sure to try to include "accurate" understandings as well as misconceptions • Please include references for all citations at the end of this document. Reference pre-conception, misconception, naïve conception, and conceptual change research • Some pre-, naïve- or mis-conceptions will be included as initial beliefs in the stepping stone columns "Before instruction, students often believe and can…"	How do learners make the transition from initial beliefs to scientifically accurate understandings, *provided appropriate instruction*? • The stepping stones reflect a process of reorganizing and adding to a learner's conceptual network as they move toward understanding the central concepts and skills. Stepping stones are "descriptions of successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (NRC, 2007). These are not just adding elements onto the lower anchor, nor just a version of the upper anchor with fewer elements these are a rethinking or reconceptualization of what or how a student knows. They are 'intermediate' anchors that indicate key problems or conceptual shifts a student needs to make as a result of instruction • These should be goals phrased as student beliefs and abilities • Include as many grade spans (columns) as needed here (likely between 2 and 4) • Group concepts and skills by key or core ideas (rows) • Clearly note initial beliefs (preconceptions) students are likely to still have (that will be addressed in the next stepping stone/column) • Indicate key connections or linkages to other ideas or topics elsewhere • Reference learning progression, conceptual change, and curriculum research "Students who view the world in this way believe and can…"	How should learners understand these concepts and what should they be able to do (skills) *as a result of instruction*? • 'Scientifically accurate' conceptual understandings and skills • Societal goals agreed upon as important for scientific and technologic literacy, and/or college readiness • An initial articulation of these concepts will be provided by the Department. Feel free to adjust as you have evidence to do so • These should be goals phrased as student beliefs and abilities • Reference current state expectations, national documents (AAAS, NRC, NAEP, etc), and relevant research "Students who view the world in this way understand and can…"
Grades	**Grades**	**Grades**
Pre-instruction	Please indicate what grade(s) each stepping stone encompasses, or at least by what grade the stepping stone should be achieved. We will clearly explain that learning of these concepts and skills may not follow grades indicated depending on individual students and instruction	The central concepts and skills are assumed to typically be achieved in early high school, unless otherwise noted

Notes

- Add notes about key limits or boundaries; what should not be expected of students in each stepping stone
- Make explicit any connections to or alignment with other topics as needed, including any mathematics important to the progression
- Some progressions will be dependent upon the type of instruction provided. If necessary, indicate general instructional methods that this progression assumes or that elicit these conceptual changes and skill development
- It may help to continually ask "what do students have to reconceptualize?"
- Include any notes about vocabulary use
- Add notes about key limits or boundaries; what should not be expected of students
- Note separately any concepts or skills that go "beyond" introductory high school courses or scientific literacy/college readiness

Authors and Reviewers

References

Figure 1. Concept and skill progression template and guidance for Massachusetts Science and Technology/Engineering topics.

DEREK C. BRIGGS

MAKING PROGRESS IN THE MODELING OF LEARNING PROGRESSIONS

Summary of Exemplar Chapters

It can be easy to stereotype psychometrics as a field that encourages research with an aim of discovering "angels dancing on the head of a pin." This stereotype is not without merit: Many of the biggest advances in psychometric research are best understood as purely methodological improvements in the ways that the parameters of a given model can be estimated efficiently. And psychometricians are apt to develop new models even when the models are not a means to an end but the sole reason for the enterprise. While advances in psychometric modeling techniques are important, they may seem esoteric and far-removed from the more qualitative research and theorizing that has characterized much of the early work on learning progressions in science. The three exemplar chapters in the modeling strand of this book break from this stereotype. The authors emphasize not only the important reasons for engaging in the formal activity of modeling a learning progression (LP) but also the necessary synergy between the modeling strand and the underlying theory (defining strand), assessment (assessing strand), and planned use (using strand) of a LP. When applied to LPs, the modeling approaches described in these chapters are very clearly a means to an end.

A key objective of this book is to elucidate challenges in LP research. The authors of the three chapters in this strand motivate their modeling approaches as ways to overcome challenges that have been identified for LPs. This requires that the authors identify a specific challenge inherent in the work with a preexisting LP and then match this challenge with a particular modeling approach that can be used to overcome it. However, for any modeling approach taken, there are almost always one or more alternative approaches that could have been taken. For that matter, as I shall point out, even within a given approach, there are typically many decisions that must be made in the specification of the model.

West et al. (chapter 12) address the challenge of making model-based inferences about the knowledge, skills, and abilities (KSAs) of students as they are exposed to a comprehensive program of instruction. The specific LP context described by West et al. is the Cisco Networking Academy, where students learn to be computer network engineers through a blend of face-to-face and online courses. In this context, it is possible that a single LP may be identified for each KSA in the Cisco Networking Academy curriculum or that more complex LPs may be identified through the combination of these distinct KSAs. The Bayesian Network (BN)

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 345–355.

approach to modeling LPs has some immediate conceptual appeal because it is rooted in Bayes Theorem, which says, in a nutshell, that the probability of some event occurring (e.g., a student answering a particular item correctly) is proportional to the combination of a prior hypothesis about that event occurring (i.e., based on prior observation and/or on expert opinion) and new information gathered by observing whether related events actually occur or not. Hence the idea behind Bayes Theorem, and a BN that is derived through its repeated application, seems consistent with the ethos that a LP should represent a testable hypothesis.

Briggs and Alonzo (chapter 13) describe the challenge posed by the complexity of a novel item format. In previous research, these authors created the ordered multiple-choice (OMC) item format to facilitate diagnostic inferences when assessing students relative to a LP. However, because these items have structural features that could distort the inferences reached in the evaluation of the response of any student to any item (due to ceiling and floor effects and guessing), the challenge is to specify a model that takes these features into account. Briggs and Alonzo's solution to this challenge is to use an approach drawn from the literature on diagnostic classification models known as the attribute hierarchy method (AHM). The AHM is a potential solution to the OMC challenge because it gives the LP modeler considerable flexibility in making hypotheses about expected responses when a student with a certain level of understanding is presented with item response options reflecting conceptions that are above, below, or equal to this level of understanding. These expected responses could then be compared to the student's actual responses. The results of the comparison have two different uses: the probabilistic classification of students according to LP levels and the evaluation of the validity of LP hypotheses.

Wilson (chapter 14) addresses the challenge of the complexity of the hypothesis about students' developing understanding that underlies the LP (rather than the complexity of the item design). Wilson distinguishes carefully between a construct and a LP. In some cases, a LP may consist of a single construct. The Earth in the Solar System (ESS) LP described by Wilson and by Briggs and Alonzo (chapter 13) is an example. The IP_Addressing LP described by West et al. (chapter 12) is another. However, Wilson notes that a number of LPs could be much more complex through interactions between multiple constructs. As an example, he describes the Molecular Theory of Matter LP, which consists of six different, inter-connected constructs. He introduces a Structured Constructs Model (SCM) in anticipation of such LPs for which one may hypothesize that a student's mastery of a specific level on one construct (e.g., Changes of State) depends on that student's ability to master a specific level on a different construct (e.g., Properties of Objects).

A laudable feature of all three chapters is that their modeling approaches are intentionally couched within a broader research and development context. This is a context that necessarily includes the other LP strands addressed in this book: defining, assessing, and using. Briggs and Alonzo (chapter 13) and Wilson (chapter 14) tie this broader context to the principles and building blocks of the BEAR Assessment System (BAS). A central idea in the BAS is that the activity of modeling a LP is an iterative process that is performed in conjunction with the specification of one or more construct maps, assessment items, and rules for scoring these items. Similarly, West et al. (chapter 12) use the BN approach to represent the interpretation vertex of the "assessment triangle" proposed in the National Research Council's (2001) report *Knowing What Students Know.* As such, the BN approach is not useful without its connections to the cognition and observation vertices.

Participants in the modeling strand at the Learning Progressions in Science (LeaPS) conference expressed concerns that psychometricians are sometimes excluded from the early work on LPs, when theories of how learning progresses are debated and assessment items are developed. When a psychometrician is simply provided with a data matrix and is asked to work some inferential magic, the results are likely dissatisfactory to all involved. The three chapters in this strand suggest the reason for this dissatisfaction. Modeling processes rely on theories that can be abstracted and submitted for empirical scrutiny. Therefore, unless LP modelers have a solid understanding of the theory behind a LP and its evolution, it is unsurprising that there is a disconnect between the inferences about student understanding implied by a model and empirical observations of students.

In the next section I discuss three major decisions that require special attention in the process of modeling a LP. These decisions, which are implicit in the three exemplar chapters, were raised repeatedly by the participants in the modeling strand at the LeaPS conference. The first decision is on the "grain size" of the LP. The grain-size decision, which is usually made when a LP is first conceptualized, has important ramifications for modeling the LP reliably. A second decision is on the evaluation of model fit. It is important to set statistical criteria that can be used to evaluate the fidelity of a model: however, at present such criteria—at least in the context of newer modeling techniques—are not yet well-established. It is also important to recognize that at some point, no matter how carefully the statistical indicators have been computed, some degree of pragmatic judgment is required to decide how close is "close enough." Finally, a third—heretofore underappreciated—decision is on the assumptions that can (or cannot) be supported about the nature of the measurement construct. Are we measuring a construct that can be interpreted as a continuous quantity, or is our interest restricted to discrete classifications? While most LP hypotheses assume the latter, some of the earliest LP models have, implicitly at least, assumed the former. This decision has important ramifications for how growth in learning over time is modeled and interpreted.

In the last section of this chapter I conclude by emphasizing the importance of sensitivity analyses in the research on modeling of LPs. A sensitivity analysis asks whether certain decisions made in LP modeling lead to practically significant changes to the inferences derived from the model. Such analyses will give us a better sense of what works, when it works, and why it works when LP modeling decisions are made.

KEY DECISIONS IN CHOOSING A MODEL FOR A LEARNING PROGRESSION

The Concept of Grain Size

The term "grain size" is often used metaphorically in LP research. However, the term has a very specific meaning in geology, where grain size is an ordinal characterization of the average size of mineral particles composing rock or sediment. This meaning of grain size is illustrated in Figure 1: Eleven containers of sand are ordered from the finest to the coarsest grain size. In the LP context, it may be helpful to picture the construct of measurement (e.g., the understanding of Earth and the Solar System, the understanding of motion and forces) as one container in Figure 1. If the construct underlying a LP (e.g., the size of the container) is held constant, the grain size metaphor can be applied in two ways.

The first approach is to identify a number of unique KSAs as the elements used to define different levels of student understanding. Briggs and Alonzo (chapter 13) took this approach when specifying the attributes used to define the levels of the ESS LP. After identifying four attributes, they used the presence or absence of these attributes to define levels of the LP. In this case, the number of distinct KSAs present at one level is the same as the increase in sophistication expected between one level and the next.

The second approach is to imagine that each level of the LP consists of grains that become increasingly larger. The grains at one level are simply viewed as an aggregation of the finer grains in the preceding level. The number of grains is not the same as the expected increase in sophistication—in fact, the opposite may be true!

The concept of grain size is especially important from a modeling standpoint because it usually limits the reliability of student classifications into LP levels. From a psychometric standpoint, it is generally easier to make reliable inferences about a student's position at a specific level of a LP when KSAs are defined throughout the LP at a very fine grain size (as in the first approach described above), such that each KSA can be treated as a binary indicator—present or absent. In this way, assessments that contain items directly tied to these fine-grained KSAs can be scored as correct or incorrect.

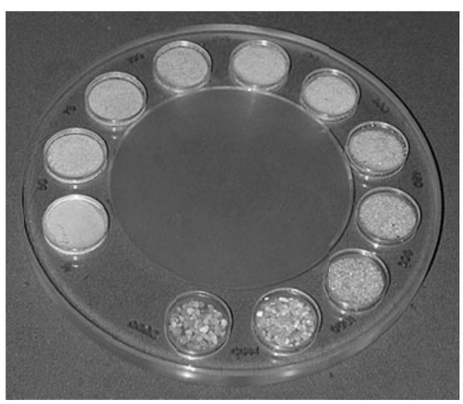


Figure 1. The concept of grain size. From http://en.wikipedia.org/wiki/File:Zandlineaal-schuin.jpg

However, two tensions result for LP developers. The first tension is philosophical. Many LPs are developed explicitly to oppose the idea that students should be assessed with respect to a simplistic dichotomy of whether they “get it” or they don’t. The second tension is more practical. If the LP is too fine-grained, it may not be useful for classroom instructional purposes; teachers may be overwhelmed with information about the many discrete characteristics of student understanding. Additionally, a LP that is too fine-grained may not be useful in large-scale assessments that produce a single summative conclusion about student achievement; too many discrete items would be required to fully characterize achievement with respect to the LP.

Thus the grain size “debate” depends strongly on the intended use of the LP. If the intended use is primarily formative, it may well be worth sacrificing reliability if the sacrifice leads to greater classroom utility and teacher (and student) buy-in. Teachers, in their day-to-day interactions with students, have many opportunities to gather additional information that can be used to correct a mistaken inference. But if the intended use is more summative (e.g., ranking students or placing them in classes), then the sacrifice is much more problematic. A mistaken inference may result in very negative high-stakes consequences.

For a particular LP, both views may be possible. The most complex computer programs can be broken down into binary operations; the same may be true of most LPs. There may be times when we wish to view grains of sand as though we were at the beach building a sand castle. At other times we may wish to view grains of sand from a distance in order to place them in the context of the surrounding landscape. Thus the challenge for the LP modeler is to understand the intended use of the LP before deciding which view is more relevant.

Evaluating Model Fit

There is a famous aphorism in statistics that “all models are wrong but some are useful” (Box, 1979, p. 202). The first part of this aphorism is clearly true. With careful scrutiny, “exceptions to the rule” can be identified for any model that makes formal assumptions about the mental processes of “typical” students as they attempt to make sense of scientific phenomena. The second part of the aphorism is more equivocal. Usefulness is a very subjective concept. Two models may be “useful” in a broadly defined sense, but that does not necessarily make them interchangeable. For example, two models might be deemed equally useful so long as both produce a numerical summary of a student’s level of understanding. Both models might be useful because, without this numerical summary, a teacher would require far more time to decide how to help that student learn. Here, utility is understood solely as an efficient allocation of a teacher’s time. In contrast, if utility is understood as the helpfulness of information resulting from a teacher’s diagnosis of a student’s understanding with one model versus another, an entirely different conclusion about utility is possible.

Box’s famous aphorism highlights a tension that arises between fit and utility when specifying a model: A model may “fit” the data better yet be so complex that it is no longer useful. All evaluations of the fit of a LP psychometric model involve variations on the same question: How closely do the things students say or do match with what the model predicts that they should say or do, if students are at known positions on the LP? Note that evaluations of fit represent proof by contradiction—one begins by assuming the model is true and then looks for evidence to the contrary. This difference between what is expected under the model and what is actually observed is aggregated for each student and/or each assessment task. If the discrepancy is large enough, this result may be taken as evidence that the model does not exhibit satisfactory fit to the data.

This "in a nutshell" presentation of the process of testing for model fit glosses over four key challenges to fit evaluation in the context of LPs. The first challenge arises because the finding that a LP model does not fit the data in the aggregate (when differences between what is expected and observed are averaged over items or respondents) is of limited value if this is the end of the fit evaluation. An aggregate model misfit represents the starting point for further detective work. Is model misfit a sign of a problem with the underlying LP hypothesis or with the way the hypothesis has been operationalized? Model misfit may result from the inability to distinguish between adjacent levels of a LP within the "messy middle" (Gotwals & Songer, 2010, p. 277). Or the problem may be how the assessment items are written and/or scored, not with the hypothesis that the levels are (a) hierarchical and (b) distinguishable. Or the problem may be that the students who took the assessment do not represent the target population in terms of their curricular exposure or other key characteristics. In order to explore plausible explanations for a mismatch between model and data, it is important to supplement aggregate fit statistics with more locally oriented statistics capable of flagging discrepancies at the level of specific items or students. For example, in the Rasch measurement tradition, careful attention is given to the fit or misfit of both items and respondents (cf. Bond & Fox, 2001).

A second challenge arises when there is evidence of misfit specific to certain items. When certain items show a lack of fit, an extreme solution is simply to remove or replace them. The problem with this solution, however, is that unless one takes the time to understand why a particular item does not elicit the expected student responses before removing or replacing it, the LP modeler is unlikely to gain any substantive insights about the underlying LP hypothesis. It is in this sense that the participants in the LeaPS modeling strand used the catchphrase, "misfit is your friend." Implicit in this catchphrase is the assumption that assessment items (and scoring rules for these items) have been thoughtfully designed to elicit responses according to the LP hypothesis. If items have been so designed, then item-level misfit should lead to targeted inferences about aspects of the LP hypothesis that need to be refined or reconsidered. Otherwise, the process of removing or replacing misfitting items is more akin to a statistical fishing expedition.

A third challenge arises when there is evidence of misfit specific to certain students' responses to assessment items. There are two reasons such evidence can be difficult to interpret. First, on any day, students' ability to think clearly about scientific concepts may be influenced by factors unrelated to their understanding of those concepts (e.g., because of a good/bad night of sleep or a fight with a sibling). Of course, such factors influence all students taking assessments, but it may be that these factors vary in severity, depending on chance. If this is true, then it is unsurprising to find that in any assessment, five to 10% of the students may answer easy questions incorrectly or hard items correctly—two types of response patterns apt to result in flags of student-level misfit. In other words, for every 100 students one expects that five to 10 will give responses that do not fit the model even if the model is accurate. Such a scenario is probable in the context of assessments with constrained choice items where a "distracted" student is more likely than a "focused" student to make guesses that are difficult to anticipate.

A second reason student-level misfit is difficult to interpret is that a LP developer will typically have much more information about the design factors used to create assessment items than about the cognitive characteristics of the students who respond to the items. When student responses do not fit the model, the next logical step is to compare the characteristics of the two groups of students—students whose responses fit the model and students whose responses do not. However, readily available variables such as gender, age, and race/ethnicity are seldom very informative in this step. The best way to diagnose possible causes of student-level misfit is to interview students in an exploratory manner after they have taken the assessment, using misfit statistics to explore unexpected responses. The problem is that such interviews, and the subsequent qualitative analyses of the results, can be very time-consuming. Again, if the evaluation of fit as part of the "modeling stage" is viewed as the goal of LP research, rather than as part of an iterative process, the computation of student fit statistics is unlikely to lead to new insights.

The fourth challenge concerns answers to the question, "How close is close enough?" When formal statistical tests are conducted to help decide whether differences between observations and model expectations are real or the result of chance, a decision must be made on whether the differences are practically insignificant—even if they are statistically significant. The point is that while it is crucial to establish statistical criteria for the evaluation of model fit, some degree of subjectivity always affects the decision to provisionally accept, reject, or revise a LP hypothesis after it has been modeled. Statistical tests of fit provide systematic evidence that can be used to inform such decisions.

While the authors of the three exemplar chapters acknowledge the importance of evaluating model fit, they offer little practical guidance for LP modelers. West et al. (chapter 12) and Briggs and Alonzo (chapter 13) describe the use of global fit statistics to evaluate the number of levels in a LP that can be supported by empirical data. However, because their chapters focus on approaches for modeling LPs, they cannot take the reader inside the more detailed detective work that is required once the models have been applied to empirical data. In addition, the fit statistics that have been developed for the models described in these chapters are relatively new—and for Wilson's (chapter 14) proposed SCM, non-existent—and as yet rarely applied in practice. For example, in proposing several innovative approaches for the evaluation of fit in the context of a BN, Sinharay (2006, p. 2) notes: "Model checking for BNs is not straightforward . . . the use of statistical diagnostic tools in this context is notably lacking." Research on fit evaluation for diagnostic cognitive modeling will be an active area for many years; there are many important issues to address. One issue is the likelihood that one may incorrectly reject, or fail to reject, the hypothesis that the model fits the data when it actually does, or does not (i.e., Type I and Type II errors). Another issue is that the use of misfit diagnoses to change the assessment, the model, or the underlying LP is not yet well understood.

The Construct of Measurement and the Representation of Growth

The three modeling approaches presented in these exemplar chapters tacitly assume that the construct(s) underlying a LP are discrete rather than continuous. Indeed, for Briggs and Alonzo (chapter 13), skepticism about the assumption of continuity for the construct underlying the ESS LP was one motivation for their interest in the AHM. Yet there are a number of examples in which LPs have been modeled under approaches that implicitly assume a continuous construct (cf. Gotwals & Songer, 2010; Kennedy & Wilson, 2007). Interestingly, in comparing the SCM approach to applications of the BAS that involve use of the Rasch family of item response theory (IRT) models, Wilson (chapter 14) writes:

> The first difference from the simple construct map formulation is that the underlying model for each construct is seen as a series of latent classes rather than as a continuum. The reason for the choice of a latent class approach in this case is that it allows qualitative a priori specification of the hypothesized links between levels of different constructs. It is not clear to me how to achieve this in a latent continuum approach. (p. 332)

Wilson also suggests it may be possible to specify latent continua within different latent classes, an idea he helped pioneer with his Saltus model (Wilson, 1989). Wilson (chapter 14) takes the sensible perspective that the assumptions invoked by any model must be sensitive to the hypotheses that LP developers wish to test (and not the other way around). However, in psychometric practice the assumption of whether a latent construct is discrete or continuous is very rarely acknowledged as equivocal, let alone submitted to empirical evaluation (Michell, 2000, 2008).

The question of whether a construct can plausibly be assumed to be either discrete or continuous is not just a philosophical one. Answers to this question may have important ramifications for how growth is modeled. In essence, all LPs are hypotheses about growth in student learning. Figure 2 depicts the intuition that most people have about growth. This is an example of a "progress variable" (which can be thought of as similar to a construct map) abstracted from a LP hypothesis about how a student's understanding of the concept of buoyancy (vertical axis) changes as the student is exposed to instruction over time (horizontal axis). In Figure 2, there is an implicit assumption that the underlying construct is continuous. If this assumption were falsified, then it would be a mistake to interpret the magnitudes along the vertical axis as a representation of the amount the student has learned from one time period to the next.

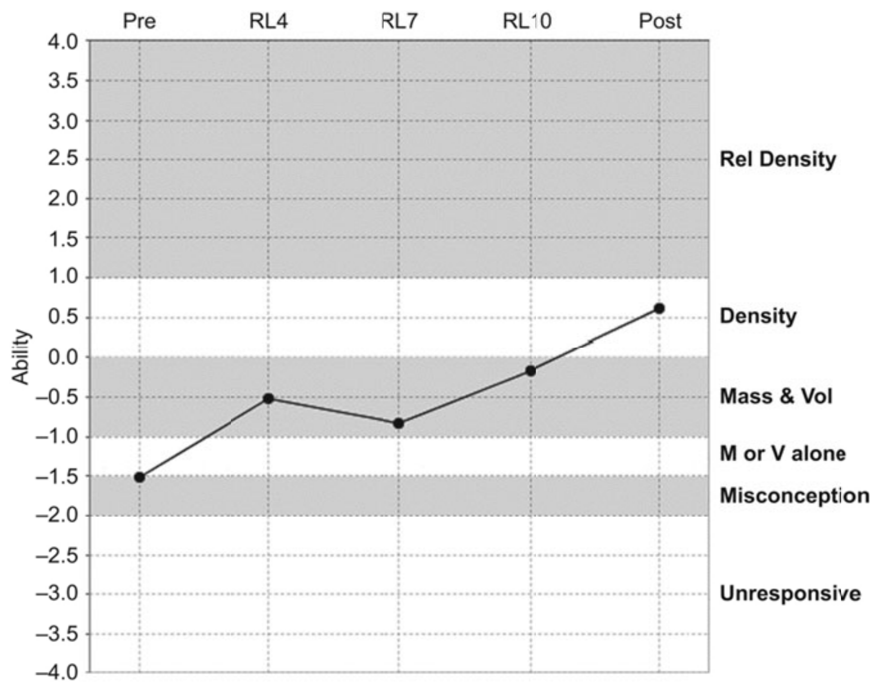


Figure 2. Growth along a progress variable.From Using Progress Variables to Interpret Student Achievement and Progress (BEAR Technical Report No. 2006-12-01) (p. 44), by C. A. Kennedy and M. Wilson, 2007. Reprinted with permission of the Berkeley Evaluation & Assessment Research (BEAR) Center.

If it is thought that the construct modeled consists of discrete latent classes with a known order, then a different approach must be taken to depict student growth along the LP. West et al. (chapter 12) give an example in their Figure 11 (p. 278). Here a student's growth is shown in terms of transitions from lower to higher levels of the LP over time.

However, there is a hidden challenge in West et al.'s (chapter 12) approach. If different tasks are administered at each time point, how can we untangle transitions between levels that represent growth from those that represent differences in the difficulty of the tasks administered? There are established approaches for this untangling in IRT that can be used so long as one is comfortable with the assumption that the construct is continuous. These approaches usually involve the use of overlapping items from one time period to the next in order to create a vertically linked score scale. I am unaware of any methods for establishing comparability across testing occasions when the latent construct is discrete.

Furthermore, when there are multiple discrete levels within a construct, there is no way to summarize growth using a relatively simple parametric function (e.g., a straight line or a polynomial equation). For example, if it is thought that a construct consists of four discrete levels, then in three time periods there will be up to 64 (4^3) possible "growth" transitions observable for a student. Should growth be characterized in terms of the percentage of students that maintain or increase their level of understanding across the three time periods? Or should growth only be characterized in terms of the first and third time periods? The modeling of change over time with discrete latent constructs represents something of a new frontier in psychometric research on cognitive diagnostic modeling (J. Templin, personal communication, August 18, 2010).

CONCLUSION: THE NEED FOR SENSITIVITY ANALYSES

In reading the exemplar chapters, it is reasonable to wonder whether and when one modeling approach is preferred over another. For example, the BN approach is probably the most general and most flexible approach presented—why not model all LPs in this way? Some psychometricians may agree that this is a sensible idea; others may argue that the generality and flexibility of BNs come at the price of very strong assumptions of conditional independence for each "node" in the network or that the complexity of parameter estimation for such models requires large amounts of student data.

One reason these sorts of debates do not often occur is that modelers tend to work on research that is a variation and/or application of the same overarching approach. In the process, they establish their own—often very specialized and esoteric—terminology and estimation procedures. Thus it is difficult for a psychometrician—let alone a LP developer with little or no training in psychometrics—to compare and contrast different approaches. However, such comparisons are starting to become possible as the many recent innovations in diagnostic classification models gain traction.[1] LP researchers must emphasize the importance of sensitivity analyses. These analyses should compare and contrast not only choices in model specification within a modeling approach (i.e., whether to specify three or four levels for a LP within a BN) but also choices of model specification across different approaches (i.e., whether the use of a BN or a Rasch model leads to substantively different diagnostic inferences about students' conceptual strengths and weaknesses).

At some level, all models may be wrong, but that does not mean that all decisions in the specification of a model are equally defensible. As research on LPs in science moves forward, it will be critical that the psychometricians involved are cognizant of, and transparent about, the strengths and weaknesses of the modeling approaches they use. All parties involved in research on LPs must play the devil's advocate with themselves and with their colleagues. It is in this spirit that the present chapter—and, for that matter, this book—is written.

NOTE

[1] For example, see the recent textbook by Rupp, Templin, and Henson (2010) that makes connections between many different types of diagnostic classification models in a very approachable manner.

REFERENCES

Bond., T., & Fox, C. (2001). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum Associates.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). New York, NY: Academic Press.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Kennedy, C. A., & Wilson, M. (2007, February). *Using progress variables to interpret student achievement and progress* (BEAR Technical Report No. 2006-12-01). Retrieved from the Berkeley Evaluation & Assessment Research Center website: http://bearcenter.berkeley.edu/publications/UsingProgVars_Kennedy_Wilson.pdf

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*, 639–667.

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, *6*, 7–24.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*, 1–34.

Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

USING LEARNING PROGRESSIONS

MARIANNE WISER, CAROL L. SMITH AND SUE DOUBLER

LEARNING PROGRESSIONS AS TOOLS FOR CURRICULUM DEVELOPMENT

Lessons from the Inquiry Project

Cognitively-based curricula may take into account research on students' difficulties with a particular topic (e.g., weight and density, inertia, the role of environment in natural selection) or domain-general learning principles (e.g., the importance of revisiting basic ideas across grades). Learning progressions (LPs) integrate and enrich those approaches by organizing students' beliefs around core ideas in that domain, giving a rich characterization of what makes students' initial ideas profoundly different from those of scientists, and specifying how to revisit those ideas within and across grades so that young children's ideas can be progressively elaborated on and reconceptualized toward genuine scientific understanding. Reconceptualization is central to a learning progression approach. To help students develop scientifically sound ideas, curricula need to go beyond enriching student knowledge and focus on deeply restructuring it. Evaluating learning progression curricula is an integral part of the learning progression approach to science education. The goals go beyond evaluating how well the curricula are "working"; they are to evaluate and revise the learning progressions themselves.

Currently, we in the Inquiry Project are engaged in such research. The Inquiry Project team consists of cognitive developmental researchers with expertise in science and math education, scientists, teachers, and curriculum developers. The team members have worked collaboratively to elaborate the Grades 3–5 portion of a larger K-8 LP developed previously by Smith, Wiser, Anderson, and Krajcik (2006) and to design a curriculum intended to nurture the development of students' knowledge in ways specified by the learning progression. The Inquiry Project Curriculum was implemented and assessed longitudinally in two schools over a 3-year span. The findings from this project will inform the evaluation and revisions of the Inquiry Project Curriculum and the Grades 3–5 portion of a LP for matter for the elementary grades. (We will use LPM to refer to the entire LP for matter at the elementary grades). Ultimately we hope to integrate LPM with middle and high school LPs on matter being developed by other research groups (e.g., Stevens, Delgado, & Krajcik, 2010).

This chapter describes three broad classes of challenges we encountered in designing and implementing the Inquiry Project Curriculum, how we dealt with

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 359–403.

those challenges, and what our implementation and assessments revealed about the merits of our solutions.

The bulk of our chapter addresses the first challenge: how to design curricula that promote reconceptualization. Specifically, we address the challenges inherent in our three major design goals: specifying the knowledge targeted in each grade; structuring the curriculum for coherence and continuity; and devising learning experiences that promote reconceptualizations. We discuss how four theoretical constructs—*stepping stones, core concepts, lever concepts, and key representations*—structure LPM and provide needed guidance for the Inquiry Project Curriculum. The chapter concludes with a discussion of the second and third challenges in our work: coordinating research and curriculum development and designing LP-based curricula that can fit with the realities of classrooms.

The K-8 LP (Smith et al., 2006) began with conceptual analyses of middle school and high school students' difficulties with the tenets of the atomic molecular theory (AMT). In particular, it linked those difficulties to incomplete mastery of core macroscopic concepts—weight, volume, density, material, matter, states of matter, and phase changes—as well as to epistemological obstacles. Wiser and Smith (2008) offer rich evidence of the role of macroscopic and epistemological knowledge in understanding versus misunderstanding AMT, as well as of the interrelationships among students' physical, mathematical, epistemological, and symbolic knowledge of matter at the macroscopic level. For example, many students believe that if a chunk of material is repeatedly halved, the pieces eventually will weigh nothing and then will disappear. This makes atoms problematic: what is their ontological status and how can they be the sole components of matter? The idea that very small pieces of matter have no weight can, in turn, be traced to students' belief that weight is reliably assessed by hefting and more broadly to an epistemology that includes the idea that our (unaided) senses tell the truth about the world (Snir, Smith, & Raz, 2003; Wiser & Smith, 2008).

The idea that AMT curricula should take into account the difficulties posed by atoms and molecules and the ideas students bring to the classroom is part of any constructivist approach, but the LPM approach ups the ante in two ways. First, the LPM approach views learning AMT as a broad and deep *reconceptualization* of ideas about matter. Second, it makes the reconceptualization of macroscopic level concepts central to this reorganization of ideas. Students' difficulties with AMT relate perhaps not so much to atoms and molecules per se but rather to the lack of fit between AMT as a model of matter and their macroscopic ideas about matter, and, more generally, between their and scientists' macroscopic ideas about matter.

LPM is a proposal for how young children's knowledge can be transformed into an understanding of matter that supports learning AMT in productive ways. Before launching into our main topic—the challenges involved in designing and implementing a curriculum that makes that transformation possible—we examine more closely our characterization of the knowledge that is transformed and what we mean by reconceptualization.

WHAT DO WE MEAN BY RECONCEPTUALIZATION?

LPM is based on a theoretical view of conceptual change that assumes that concepts and beliefs develop from universal core knowledge that heavily constrains knowledge acquisition in infancy and early childhood (Carey, 2008; Pinker, 2007; Spelke, 2000). The universality of core knowledge and early learning experiences in the domain of matter (all children encounter objects made of different materials—solids, liquids, etc.) ensures the universal development of the concepts of object, aggregate (liquid or powder), material, heaviness, and size, although languages, cultures, and specific environments emphasize some concepts more than others (e.g., the material and object levels of description have different emphasis in English, Japanese, and Quechu'a) and produce different knowledge about specific objects and specific materials (Imai & Mazuka, 2007; Lucy & Gaskins, 2003).

These concepts are, of course, far less sophisticated than corresponding scientific concepts. However, and more importantly, these concepts differ from scientific concepts in very fundamental ways, making young children's and scientists' understanding of objects, materials, and their transformations incommensurable (at least in a weak sense).[1] Therefore, achieving an understanding of matter commensurate with the scientific view requires major conceptual changes; those changes, as a whole, are the reconceptualization we referred to in the Introduction. This reconceptualization is not inevitable; the majority of American high school graduates have learned statements, models, and equations by rote although they have not acquired an understanding of matter much more advanced than that of kindergartners. LPM describes how the reconceptualization could take place. LPM is virtual in the sense that it cannot unfold without a curriculum that supports the restructuring of knowledge and sustains its meaningfulness in instruction. LPM is also hypothetical since we will not know whether the course of knowledge development it postulates is an effective path to scientifically sound understanding of matter until we observe its effect in classrooms. We also believe that the complexity of the reconceptualization places such numerous and interrelated constraints on curricula that there may be only a few ways to design successful curricula.

In any constructivist account, learning links new information with pre-existing knowledge and is productive only if the new links provide the basis for new interpretations of observations, which are both meaningful to students as well as compatible with the scientific knowledge targeted by educators. (We later address the meaning of "compatible.") Students do not learn when new information does not relate to their current knowledge; such new information is thus either quickly forgotten or is remembered only as rote-learned statements. Students also do not learn when linking new information to their current knowledge results in irresolvable conflicts or leads to modifying that knowledge in unproductive ways. For example, the information "gases are matter," as typically presented to young students, can be meaningless if students do not understand gases and/or matter. New information may create a conflict if students think that matter is what can be held and touched. This conflict may lead students to revise their prior belief about matter in a way not intended by their teacher: since air and other gases cannot be touched or held, "matter" is not what students thought. They may further conclude that "matter" must refer to anything physical. Thus heat, light, forces, and energy must be matter, too.

In a different scenario,[2] students first learn that any solid or liquid (i.e., matter as they think of it) has weight,[3] and that weight is proportional to the amount of material.[4] This information makes cogent a range of experiments about dissolving, evaporation, and gases in which weight can be used as indirect evidence for the presence of material. Students discover that when salt is added to water, the salt is no longer visible, but the weight of the salt water is the sum of the weight of the salt and the weight of the water. They also discover, using a "two bottle system"[5] and a heat lamp, that although the amount of water in the first bottle decreases and water does not (yet) appear in the second bottle, the weight of the system does not change. Because they have learned to associate weight with amount of material (for solids and liquids), students conclude that the salt (in the first experiment) and the water (in the second experiment) must exist although both are invisible. The conclusion that "the salt and the water still exist" is not yet a belief but rather a statement that is both implausible and worthy of further consideration. Students are caught between the logic of the conclusion and its incompatibility with their deeply entrenched belief that matter, including water, can be seen and touched. However, they notice drops forming on the wall of the second bottle. In time, more and more water appears in the second bottle until there is no more water in the first bottle. These observations strengthen the plausibility of the idea that water can travel through air in an invisible form.

Other investigations provide further reasons for considering that, in some contexts, matter can exist in invisible form while retaining familiar properties such as having weight and exerting force. A basketball weighs more after being filled with air than before and syringes cannot be pushed all the way in; one can feel the resistance of the air inside. Air must be "something"—it has weight and occupies space, like solid and liquid samples. Again, these are conclusions, not yet beliefs. Students might then watch iodine sublimate—the space above the solid iodine turns purple. ("Some gases are visible!") Students have started to entertain the possibility that gases are material just as solids and liquids are material: they have weight, occupy space, and are (sometimes) visible. Again, this hypothesis is "hard to believe." Children need an *explanation.* ("How is this possible given everything I *know* about matter?") The situation is ripe for introducing a particulate model that provides a unitary (and compelling) way of explaining these diverse observations using a small set of assumptions about the structure of matter. To help students commit to a particulate model of matter, however, it is important to reconcile it explicitly with perception.

A visual demonstration helps students understand how something invisible becomes visible. For example, when tiny dots are spread equally across the page, they are very difficult to see from beyond a distance of a few feet; when those same dots are clustered, it is much easier to see them. This demonstration may help students understand why we see salt in its solid form but not after it has been dissolved, or why we see water in its liquid form but not in its gas form. It is not the amount of salt or water that changes; the change is the distance between the tiny particles that make up the salt and the water.

Students then explore a particulate model of matter. Applying the model in various contexts, they realize that it is a "good" model of matter because it explains a number of their observations. For example, the model explains differences in materials and why material identity and amount are conserved and weight does not change when a sample of material melts or freezes. Faced with the "explanatory success" of the particulate model, students are more willing to give up their old beliefs that matter is continuous (it only appears to be) and that gases are not matter (they are matter distributed very sparsely). Students' view of the nature, structure, and transformations of matter is now well on its way to being *reconceptualized.*[6] The differences between the students' knowledge systems before and after reconceptualization are deep and broad; the two systems are incommensurate. Yet each step described in the previous paragraphs creates new knowledge that is closer to scientific understanding while remaining meaningful. How is that possible?

We view the ongoing reconceptualization as involving the creation of a succession of small "local" networks of beliefs that are first entertained as credible, then reinforced by further experiences, and finally, given the epistemic status of principles. In deeply constructivist fashion, beliefs consistent with the scientific theory are progressively adopted by students, as carefully ordered inquiries reduce the cognitive distance between students' initial beliefs and scientifically compatible ones. In contrast observing the water level drop in a bottle and being told "the water evaporates" does not lead anywhere because the cognitive distance between what students know and what they hear is too great ("What is evaporation? I cannot see water in the air.") Weighing the two-bottle system every day and observing that its weight remains constant provide evidence that the amount of stuff in the closed system remains the same. However, without further scaffolding, the experiment may confuse students. On the other hand, helping them to notice that drops have formed on the bottle wall strongly suggests that water moved from one place to the other. Relying on a belief, rooted in infancy, that if an object moves from A to B, it must follow/have followed a continuous path (even if the path is partially visually obstructed), students now have a very good reason to believe that at some point, there was water in the air although they could not see it. When this idea is linked with the idea of unchanging weight, the hypothesis that water can exist in an invisible form is now credible enough to apply to other learning experiences, increasing the credibility of the hypothesis. For example, students are now more likely to entertain the possibility that air has weight as an explanation of why pumping air into a basketball makes a pan balance scale tip because they are already entertaining the possibility that solids and liquids can be invisible, like air, and still have weight. In other words, the more numerous the phenomena pointing to a single explanation, the more acceptable the explanation.

We describe our pedagogical approach as follows. First, we *propose* a hypothesis within a context that makes it plausible (rather than taking for granted that a statement has to be accepted right away); then we increase its plausibility by presenting more and more empirical evidence for it, making it more widely applicable, and fostering more links between the hypothesis and existing knowledge (e.g., the iodine experiment would create a link with "things that are visible are matter").[7] A set of increasingly plausible and coherent hypotheses can form a new local network of knowledge—a liquid material becomes invisible and yet exists; air is a gas, is invisible, and has weight; a gas (iodine) may be visible, proving it exists; all matter has weight and all that has weight is matter. This local network is more and more salient as the plausibility of its components increases. Other local networks may develop in parallel or in succession (e.g., a network about a particulate model of matter) and then become linked to each other. Local networks give each other credibility in the same way that individual, plausible hypotheses do.

At the same time, beliefs that were part of the initial network and inconsistent with the new hypotheses begin to fade away or begin to be re-interpreted or qualified—"air is nothing" becomes "air is invisible"; "water disappears when the level goes down" becomes "water turns into a gas state"; "matter is what can be touched and seen" becomes "matter in solid and liquid form can be touched and seen (if one has large enough samples)."

However our "success story," as described above, depends on the following conditions. Students believe that small pieces of solid and liquid material have weight; students are skilled at making direct and indirect measurements of weight and using weight as evidence for claims about matter; and students can distinguish between what things look like and what they are. Developing those beliefs and skills is part of a prior and lengthy reconceptualization process.

We hope we have convinced our readers that, in order not to remain a fairy tale, the reconceptualization of matter must be carefully prepared for and given ample time to take place. Numerous items of new information are needed, with each item introduced at the right time and relationships between new information and current knowledge must be established in a certain order. Learning experiences must be selected with care because specific contexts require different elements of knowledge. Given the time frame and the complexity of the reconceptualization, a curriculum needs to be planned at multiple time scales and grain sizes. In the Inquiry Project, we asked: What is our target knowledge at the end of fifth grade? Given that, what is our target knowledge at the end of third and fourth grades? What should we start with in third grade? In what contexts should third graders explore materials? When should we start using the word "matter"? What specific experiences could help students develop a concept of volume? What key representations can be introduced to support the reconceptualization of weight?

CHALLENGE 1: HOW CAN CURRICULA BE DESIGNED TO FOSTER RECONCEPTUALIZATION?

The original K-8 LP developed by Smith et al. (2006) was used as our starting point. However, this LP was not ready for translation into a curriculum. The Inquiry Project team members elaborated on the Grades 3–5 segment of the original LP as they concurrently developed LPM and the Inquiry Project Curriculum. The team had three goals:

(a) Given that the Inquiry Project Curriculum prepares students for learning AMT but does not teach it, we needed another end goal—that is, we had to specify a coherent and meaningful state of knowledge about matter, compatible with the scientific view and useful for further learning, that would be the target for the whole curriculum. We also needed to specify similar end goals for the third and fourth grades.
(b) Taking for granted that LP-based curricula should be broadly organized around networks of interrelated concepts, we had to decide which concepts were important to include, when to include them, and in what contexts to introduce them.
(c) Given that reconceptualizations are complex and long-drawn, we needed to design and sequence instructional activities in each grade so that each activity would be meaningful and, at the same time, get students closer to a scientific understanding of matter. In particular, we needed to specify key experiences that promote the progressive integration and generalization of new pieces of knowledge.

In the sub-sections that follow, we discuss the challenges posed by our three goals and how we dealt with these challenges in the Inquiry Project (using the theoretical constructs of stepping stones, core concepts, lever concepts, and key representations). We also address how applicable these new constructs might be for LPs that focus on other grade bands or content areas. While the specific decisions taken in the Inquiry Project Curriculum, as reported in this chapter, should be viewed as “first generation” informed guesses, as all decisions are open to revision given new empirical findings, we think that the usefulness of our theoretical constructs may be enduring.

Table 1 describes the four theoretical constructs and how they are related. Table 2 uses these constructs to summarize the ways we elaborated LPM.

Table 1. An Overview of Key Theoretical Constructs for Structuring a Learning Progression.

Theoretical Construct	Brief Characterization	Examples	Relation to Other Constructs
Core Concepts	Concepts that are central to (a) the scientific theory itself, or (b) learning the scientific theory Core concepts provide vertical continuity and horizontal coherence to the LP. The content of core concepts and the relations among them change over time. Individual rows in Table 2 correspond to different core concepts targeted by the Inquiry Curriculum.	Of (a): matter, mass, volume, density, states of matter (for later grades: pure substance, mixture, element, compound, atom, molecule, bond) Of (b): material, amount of material, weight, size, particle	Concepts are mental entities constituted of many kinds of beliefs (e.g., beliefs about what invariants they refer to in the world and how they are related to other mental entities in laws, explanations, and generalizations about phenomena.) Important concepts are often symbolized in language with single words so that beliefs involving those concepts can be expressed with sentences. Relations among scientific concepts are also expressed with the language of mathematics.
Lever Concepts	A subset of core concepts that (a) students already have at the beginning of instruction and are connected to other concepts in capturing generalizations about the world, (b) need to be revised, and (c) promote the revision or introduction of other new concepts. They function to provide guidance about productive entry points at a given grade level.	Material, amount of material, weight, and size (for grades 3-4) Solid and liquid materials, particle, and heavy for size (for grade 5) Note: How long a core concept remains a lever depends on how long the reconceptualization(s) of which it is part take(s).	Lever concepts are a subset of core concepts. They are typically precursors of scientific concepts, some more clearly so than others. For this reason, they are often overlooked as important targets of instruction by science educators, who identify learning goals primarily through decomposition of the expert knowledge. They drive the reconceptualizations necessary to reach stepping stones.

Table 1. An Overview of Key Theoretical Constructs for Structuring a Learning Progression (continued)

Theoretical Construct	Brief Characterization	Examples	Relation to Other Constructs
Key Representations	Mental models or external symbolic representations that serve to integrate existing knowledge elements and support new inferences about properties of the physical world. They are powerful tools for reconceptualization. New kinds of mental representations are typically linked with new forms of external symbolization.	Weight and volume measure lines linked to thought experiments about repeated divisions of chunk of material. "Compositional model" of materials Particulate model of materials (In later grades: periodic table of elements linked to an atomic-molecular model of matter)	Key representations are needed to support changes in core concepts, which constitute the achievement of a given stepping stone. The importance of threading instruction around the construction of these key representations is often overlooked by traditional curricula, which emphasize factual or procedural knowledge.
Stepping Stone	A new equilibrium state of students' knowledge network which is *conceptually* closer to the upper anchor of the LP (e.g., AMT) than the lower anchor and allows students to move forward on the LP with further instruction. A stepping stone is reached when a subset of core concepts has undergone significant reconceptualization so that students have a new set of beliefs, principles, models, numerical and mathematical understandings, and representational tools that provide them with coherent interpretations of a broad range of phenomena.	A stepping stone organized around a compositional (mental) model of materials (see Table 2) enables students: to distinguish properties of objects from properties of materials; to associate weight and volume with any piece of solid and liquid, however small; and to understand that material identity, amount of material, and weight stay the across reshaping/grinding. A stepping stone organized around a particulate model of matter (see Table 2) allows children to understand gases as matter, to account for material identity and its conservation during phase change, to explain changes of state, and to develop packing schemes to explain differences in density.	Stepping stones are multi-faceted and complex. Achieving a stepping stone involves the reconceptualization of a significant subset of core concepts (e.g., the lever concepts that were the target of instruction) as well as the introduction of new core concepts (e.g., volume, matter) into the network. This in turn requires the development of a variety of new beliefs relating those concepts to each other and situations in the world as well as new explanatory models, epistemological understandings and representational tools.

Table 2. An Elaborated Learning Progression for Matter (Grades 3 to 5)

	Lower Anchor (Grade 2)	Grade 3	Grade 4	Grade 5
Core Concepts (Lever concepts are in italics)	**Materials**: • Differentiates **object** from **material** it is made from • Knows the names of some liquids and solid materials. • For solids, object level description is more salient than material (e.g., it is a boat rather than that it is wood) • Associates some intensive properties (smell, taste, hardness, brittleness) with some materials	***Material:*** • Knows broader range of **solid materials** (organizes names hierarchically: e.g., pine and oak are wood; copper and steel are metal) • Material as an explicit concept: **made of"** differentiated from **"made from"**—and means constituted of little pieces of X • Materials have specific (intensive) properties	***Material:*** • Knows greater range of **solid, granular,** and **liquid** materials • Extends knowledge of **properties** to include **heaviness for size** • Understands the **same material** can be in a **different form** (e.g., rock, crushed rock, etc.)	***Material*:** • Extends notion of materials to include **gaseous materials** (air, water vapor) • Understands that some materials can be in ***solid, liquid,*** or **gaseous form** (e.g., ice, water, water vapor)
	Weight: • Measured by hefting • Some objects are heavy; others are light. Not all objects have weight (e.g., tiny things, Styrofoam, balloons) • Changes when shape changes (because object identity changes) & when adding and removing material • Liquids and powders do not (necessarily) have weight	***Weight:*** • Solid objects have weight • Weight is measured with a scale; differentiated from heft • Weight is extensive (i.e., proportional to amount of material) • Any piece of solid material, however small, has weight • Weight is invariant across shape change	***Weight*:** • Granular materials and liquids as well as solids have weight • Weight of liquids can be measured indirectly • Any granular or liquid sample, however small, has weight • "Heavy object" differentiated from "heavy kind of material" • Weight invariant across grinding	**Weight**: • Gases have weight; measures weight of gases indirectly • Any piece of material, however small, **visible** or **not**, has weight • Invariant during phase change (freezing, melting, evaporating, condensation) or in dissolving • Uses invariance of weight across phase change as evidence that no material has been gained or lost
	Amount of Material • Shape changes do not change amount of material. • May not quantify amount of material	***Amount of Material*** • Quantifies amount of material • Associates weight with amount of material	***Amount of Material*** • Relates volume to amount of material • Invariant across grinding	***Amount of Matter (Mass)*** • Associates weight with amount of matter and quantifies amount across different materials • Begins to differentiate amount of matter from volume

Table 2. An Elaborated Learning Progression for Matter (Grades 3 to 5) (continued)

Core Concepts (Lever Concepts are in Italics)	Lower Anchor (Grade 2)	Grade 3	Grade 4	Grade 5
	Size/Occupying Space • Lacks geometric concepts of length, area, volume • "Bigness" is perceptually based and not explicitly differentiated into spatial dimensions. Big objects do not fit into small places and are hard to carry. • No two solids objects can occupy the same space at the same time	***Size/Occupying Space*** • Solid objects take up 3-D space • Occupied 3-D space differentiated from length and area. • Begins to measure and compare occupied space of objects using centimeter cubes	**Volume** • Granular and liquid materials occupy space • Measures volume of liquid and granular materials • Understands water displacement depends upon volume of submerged object, not weight • Measures volume of solid objects using water displacement	**Volume** • Differentiates different senses of volume (volume of material, space occupied by object, amount of empty space) • Gases are compressible (i.e., volume depends on size of container)
	Heavy and light materials • Objects made of some materials are heavier than objects made of other materials (knowledge specific to some materials)	**Weight associated with Material** • Objects made of some materials are heavier than objects made of other materials	**Heavy for Size** • Differentiates heavy object from heavy material (solids, liquids) • Solid and liquid materials are (more or less) heavy for size	***Heavy for Size*** • Gases are much less dense than solids or liquids
	"Materiality" • Solid objects and liquids can be touched and seen	**"Materiality"** • Any piece of solid material, however tiny, has weight and occupies space • Begins to think materials might still be there when dissolved	**Matter** • Any amount of liquid, however tiny, has weight and occupies space • Begins to see deeper similarities among solids, liquids, and granular material	**Matter** • Gases are material in the same sense as solids and liquids. • Matter has weight and occupies space • Solids, liquids, and gases are forms of matter; some materials can exist in all three forms (e.g., water) • Beginning particulate model of matter (material X consists of small X particles, spaces between particles, motion of particles increases with temperature)

Table 2. An Elaborated Learning Progression for Matter (Grades 3 to 5) (continued)

	Lower Anchor (Grade 2)	**Grade 3**	**Grade 4**	**Grade 5**
Epistemological Beliefs	• Believes the "senses" tell the truth • Has difficulty with the appearance-reality distinction	• Understands weight measurement is more precise, reliable, and valid than felt weight • Understands scales vary in sensitivity • Knows principles of good length and weight measurement • Understands things may be different from how they look • Generates competing explanations for an observation (e.g., why an aluminum cube is heavier than a wood cube)	• Extends principles of good measurement to volume • Uses mathematical reasoning (i.e., subtraction, division) to infer weights and volumes • Understands idea of a fair test in experimentation (i.e., need to control for other variables) • Uses weight and volume data to make claims; does experiments to distinguish rival hypotheses (i.e., weight versus volume as relevant to water displacement)	• Understands idea of a "system" • Uses models to represent unseen, invisible events in a system • Makes predictions based on models • Compares different models and evaluates models by how well they explain data • Uses weight and volume data to make claims about unseen entities and the existence of invisible material
Key Representations		**Measure Line**: Uses measure lines to represent weight as dimension **Imagined Small Pieces**: Uses thought experiments of repeated division to form representations of tiny pieces of material	**Measure Line**: Uses and begins to coordinate measure lines for weight and volume **Compositional Model of Materials**: Uses thought experiments to imagine whole as sum of parts	**Beginning Particle Model** is constructed and used to explain: • Why gases are matter yet invisible and compressible • What happens in freezing, melting, and evaporation including why weight is same
Anchors and Stepping Stones	**The Lower Anchor** Children organize their thinking primarily in terms of the behavior and properties of **Whole Objects,** with little explanation of why objects have the properties they do. These properties are judged perceptually and hence change with context or with changes in the whole object. There is no mathematical quantification or measurement.	**The Grade 3 and 4 Stepping stone** Coordinates all of the above understandings (Grade 3 and 4) in a **Compositional (Mental) Model of Materials:** Central to this model are: (a) making an ontological distinction between **objects** and **materials**; (b) understanding that objects are constituted of materials that have characteristic properties that are maintained in decomposition; (c) coming to see that pieces of material, **no matter how tiny,** have weight and take up space; (d) understanding that both solids and liquids have weight and take up space; and (e) developing measures of weight and volume.		**Grade 5 Stepping Stone** Coordinates all of the above understandings in a **Beginning Particulate Model:** Central to this model is being able to imagine: (a) that there can be little tiny pieces of matter too small to see that have weight and take up space and are in motion; (b) that gases are matter, but invisible, because the pieces are so tiny and widely separated—in contrast, in solids and liquids they are more tightly packed; (c) that heat increases the motion of particles; and (d) that phase change does not change the weight or identity of the particle, but does change spatial arrangement.

Challenges in Specifying Productive Learning Goals

LPM is structured around core concepts and stepping stones. We define core concepts as those that (a) are central to scientists' understanding of a domain *or* (b) play a central role in how students learn scientists' concepts (and hence are important concepts to include in a long-term LP account although they are not necessarily core concepts in the scientific theory) (see Wiser & Smith, 2009). Stepping stones are intermediate, coherent states of knowledge that can be targeted as learning goals for different grades.

Core concepts in LPM. What are the core concepts of LPM? Concepts that are central to scientists' understanding of a domain are easily identified. Among these, matter, mass, volume, density, and states of matter are clear choices. (Atoms and molecules, bonds, elements, compounds, pure substances, and mixtures commonly appear in middle school and high school science standards.) One challenge for any LP is to identify *other* concepts that may be involved in developing the concepts that are in the scientific domain, including some concepts from early childhood whose relevance to the LP may not be apparent at first. The importance of those concepts is less obvious and only emerges from an extended analysis of studies of children's learning and reconceptualization in a given domain. For example, in earlier work with middle school students who were learning about matter, Smith and colleagues found there were strong correlations between believing that "any small piece of stuff has weight" and developing an understanding of density (Smith, Grosslight, Davis, Unger, & Snir, 1994; Smith, Maclin, Grosslight, & Davis, 1997). In our current longitudinal study, we also found that this understanding strongly relates to the emergence of precursors for the concept of density in Grade 4 (Smith, Wiser, & Carraher, 2010). We are currently examining its relationship to other important ideas in students' matter knowledge network, such as the emergence of a general concept of matter in Grade 5. Thus LPM also includes the concept of weight as well as the concepts of material, amount of material, solid, liquid, and particle that are crucial in bridging young children's ideas about matter to the scientific theory (see Table 2).

Another challenge for any LP is specifying the precursors and "intermediate" forms that core concepts take as they interact with a curriculum designed to foster their transformation into scientific concepts. These specifications require making inferences from the difficulties students experience when taught the scientific theory. Such inferences require a large database, which is not always available at first, and often extensive and complex conceptual analyses. Many aspects of student knowledge are implicit, poorly articulated by students, or often extremely context-dependent. For example, young children may answer conservation-of-amount-of-material questions differently depending upon the kind of object and the material involved (e.g., clay balls, metal cubes, wire coils, plastic wire) and the kind of transformation process (reshaping, dividing into little pieces) (Uzgiris, 1964). Similarly, in sorting different size objects covered with white tape into those made of steel or aluminum, young children are more likely to attend to *heaviness for size* when they are handed objects one at a time rather when they have two objects to compare (Smith, Carey, & Wiser, 1985). Moreover, students may give the same word multiple meanings. For example, when students say "Wax becomes water when it melts," do they use "water" to mean "liquid" or does it really mean "water"?

Starting point in the Inquiry Project. Once core concepts are established, another challenge is richly characterizing what children may know at the beginning of our intervention (the lower anchor). This is a challenge because concepts in the matter knowledge network are complex. Such concepts consist of a large set of beliefs about objects and materials (some can be expressed verbally, some not). These beliefs are closely interrelated and include the situations (or contexts) to which they apply, the invariants students pick up about those situations, and the ways the situations are symbolized (Vergnaud, 1996).

Nonetheless, cognitive developmental researchers have (fairly well) established some aspects of young children's knowledge about matter (Au, 1994; Baillargeon, 2002; Carey, 2008; Dickinson, 1987; Piaget & Inhelder, 1974). Because of the limited attention science instruction receives in most American schools at the K-2 grade level, with its lack of focus on promoting conceptual change, we assumed (and confirmed in our pre-test interviews) that our third graders' knowledge about matter would not be qualitatively different from that of kindergartners (see Wiser & Smith, 2008, for a discussion of this issue).[8]

Table 2 summarizes many of these initial ideas. Children make some distinction between objects and materials (Au, 1994; Dickinson, 1987) and realize that reshaping a clay ball does not change "the amount of clay" in the ball (Piaget & Inhelder, 1974). However, their knowledge of materials is limited. They focus on perceptual rather than objective (and measurable) properties of materials, do not yet clearly differentiate between *made of* and *made from*, and are likely to think the name of a solid material refers to the object itself (e.g., "This is wood"). Consequently, material is not yet an ontological category. Children's ontology still rests on object-level descriptions. Matter is not yet an ontological category either. Liquids and solids are (at best) seen as similar in the sense that they can be seen, touched and held, while gases cannot. Children think weight is reliably assessed by hefting. Their ideas about the conservation of material identity across transformations and the distinction between material kind and state are sporadic and context-dependent (Stavy & Stachel, 1985). They think weight changes across transformations (Piaget & Inhelder, 1974). Explicit concepts of density, volume, and matter are absent (Smith et al., 1985).

One challenge of LPs in science education is to chart how scientifically compatible concepts are developed and integrated. We assume that the path to the concept of matter starts with children's ideas about solid and liquid materials and that amount of material is a precursor to mass. As sensible as these assumptions seem, they require empirical validation. Will a curriculum based on these assumptions advance children's understanding of matter and mass in a scientifically compatible way? Moreover, in other LPs, young children's ideas relevant to each core concept may be far less transparent.

Stepping stones. What is the targeted knowledge for Grades 3, 4 and 5 in the Inquiry Project Curriculum? We propose that what differentiates two successive states of the knowledge network in LPM is not that one *contains more elements* of the expert theory or that it *resembles* the expert theory more closely. Rather, the difference is in the *structure* and *content* of each state. The structure and content of the targeted intermediate states are such that, with the support of carefully crafted curricula and effective teaching students can move from one to the next and, eventually understand a basic version of AMT. We think of each state as a *stepping stone* along the path to understanding AMT.

We hypothesize that the states of the knowledge network qualify as stepping stones only if they are sets of core concepts, beliefs, principles, models, numerical and mathematical understandings, and representational tools that provide students with relatively *coherent* interpretations of a broad range of phenomena, while allowing them to progress toward understanding AMT with further instruction. When the LPM reaches a stepping stone, the network has undergone significant reconceptualization; that is, the network has reached a new state of equilibrium and is *conceptually* closer to AMT than the lower anchor. (This does not imply that the whole network has changed, as we explain later. We hypothesize that the interconnections between elements of the network are such that sub-networks can be reconceptualized successfully while other sections are unaffected.)

The Grades 3–4 stepping stone. Given their initial understandings, what relevant stepping stone might children be able to achieve by the end of Grade 4 that would better prepare them to progress in Grade 5 toward a scientific understanding of matter?

We decided that an important first step for students is to develop a number of mutually supportive beliefs about solid and liquid materials. Samples of solids and liquids are similar because they can be seen and touched, they occupy space and have weight, and they may exist in pieces so small they cannot be detected by the senses alone. Solids and liquids made of certain materials may be heavier for their size than objects made of other materials. When a solid chunk of material is deformed (e.g., ground up), the identity of the material and the weight and volume of the object do not change. Weight is scale weight, not heft, and is a measurable objective quantity. Volume refers to the 3-dimensional space occupied by an object, not its area or length, and therefore volume can be measured by water displacement.

These key beliefs, which constitute the Grades 3–4 stepping stone, were the target of the Grade 3 and 4 Inquiry Project Curriculum: together, these beliefs provide a new, coherent understanding of matter (or rather, of solid and liquid materials) that is conceptually closer to AMT than the Grade 2 lower anchor. This understanding supports further reconceptualization in Grade 5 toward a new stepping stone that includes the belief that gases are matter. The *compositional model of material* is the center of the Grades 3–4 stepping stone. This new representational tool, which allows students to focus on the parts and the whole simultaneously, supports the belief that material identity, amount, weight, and volume do not change when the chunk of material is deformed, cut, or ground. The model also supports students' learning and understanding of the measurement of weight and volume.

The *compositional model* is a representational tool, not a model of the structure of matter. Unlike the particulate model and AMT, this model (a) is fundamentally about materials (not matter) and does not apply to gases, and (b) represents materials as continuous and homogeneous (not as discretely spaced particles). The development of this robust model of material is a crucial step toward a scientific understanding of matter. However, the model is currently overlooked as an important target of elementary science instruction. Perhaps the reason is that the model has obvious differences from the expert model (curricula generally target "pieces" of the expert theory that are "scientifically accurate" even though they may not cohere for students, unlike coherent stepping stones). Other reasons may be that curriculum developers mistakenly think young children already have these beliefs or think that children can easily develop such beliefs themselves—something our findings explicitly disprove.

The Grade 5 stepping stone We now consider what coherent sub-set of concepts and principles, representational tools, and epistemological points could constitute the next stepping stone. How should students think about matter at the end of elementary school (e.g., Grade 5) if they are to make sense of AMT as taught in middle school and appreciate its explanatory value?

Since AMT is a model of matter, the Grade 5 stepping stone could include the knowledge that matter exists in three states. The stepping stone could articulate that gases are matter for the same reason that solids and liquids are matter–they have weight and volume, materials keep their identity in physical transformations,[9] and amount of material, and therefore weight, does not change in physical transformations. Density and thermal dilation are also important components of a scientific account of matter; they can be used reliably to identify materials. Moreover, AMT provides a powerful explanation of the dependency of specific volume on temperature. Therefore density could be the nexus of the Grade 5 stepping stone. Finally, the Grade 5 stepping stone could include a particulate model of matter (in which no distinction is made between atoms and molecules) as a scaffold to understanding AMT.

We found ourselves at a crossroads in our development of the fifth grade curriculum. Given the limited number of lessons in the curriculum, we could not do justice to all the components of the Grade 5 stepping stone. Hence, we decided we could either aim for: (1) a fifth grade curriculum organized around density, quantifying the notion of heavy for size developed in the fourth grade and developing new representational tools that support this explicit quantification or (2) a fifth grade curriculum organized around an ontological reconceptualization of the concept of matter. It would include the notions that material objects include gases, weight and volume are inherent properties of matter, and amount of matter and weight are conserved across phase changes.

Approach 1. This fifth grade curriculum recognizes that fifth grade students are (potentially) ready, conceptually, to understand density. They can move from the belief that objects made of certain materials are heavier for their size than objects made of other materials to the belief that some materials are denser than others. Dot models (Smith, Snir, & Grosslight, 1992) could be used to illustrate that matter is “packed more tightly” in denser materials. In addition, to enrich children’s understanding of density, it is possible to use weight and volume lines or to use methods developed by Lehrer, Schauble, Strom, and Pligge (2001) to develop children’s understandings of coordinate graphical representations. A unit on sinking and floating might be meaningful at this point. By the end of this version of the Grade 5 curriculum, the stepping stone would be a solid understanding, at the macroscopic level, of weight, volume, material, mass, and density, and their interrelationships.

Approach 2. This fifth grade curriculum might include the concept of *matter* as superordinate to *material*. With this perspective, the curriculum could develop the understanding that matter exists in three states and that some materials melt, freeze, evaporate, condense, and dissolve in liquids. By the beginning of fifth grade, students are (potentially) ready to consider these transformations and to discover which properties of matter are constant (weight, amount of material, material identity) and which properties of matter are not (e.g., volume). Using the conservation of matter as a central theme, the curriculum might introduce a basic particulate model of matter (e.g., a model that assumes matter consists of different kinds of discretely spaced particles that have characteristic weights, are held together by bonds, and are in motion). Introducing a particulate model of matter at this juncture makes sense because it may be the only satisfactory way, reconciling theory with perception, to explain changes of state, the conservation of stuff across them, and why gases can be material and yet invisible. A particulate model better explains the expression “made of” and why objects of different materials differ in heaviness for size. Finally, the fifth grade curriculum would be a suitable context to begin developing an explicit understanding of modeling.

We decided to take Approach 2. A focus on density (Approach 1) would have enriched and entrenched the student understanding of the relationship between weight and volume, and the idea that gas has much lower density than solids and liquids. Overall, the density approach would have offered a more complete set of the core concepts relevant to understanding matter at the macroscopic level. However, we thought an ontological reconceptualization of matter as particulate and including gases (Approach 2) would be more immediately accessible to students and would also provide the most pay-off or “legs” from a scientific perspective. The understanding that matter is fundamentally particulate, gases are matter, and mass is conserved, combined with the development of the epistemology of modeling, has extremely wide-ranging implications. This approach introduces students to a productive new framework for thinking about matter. Additionally, the phenomena students explore in this approach might be “newer,” more surprising and interesting, and more in line with some expectations in their existing science curriculum. Last but not least, the approach prepares students for later study of density. If they have used the particulate model to account for the compressibility and behavior of gases, they will have visual representations that make it easier to understand density when it is studied formally.

Summary of the stepping stones. We are on new and exploratory ground in introducing the particulate model at the fifth grade level. We continue to ask: (1) What prior experience do students need in developing, using, and analyzing models?; (2) What set of elements should be included in a particulate model?; (3) How should these elements be introduced in a way that deepens students' epistemological understanding of models (e.g., their revisable nature and their use as tools of inquiry) and of phenomena too small to see? Our observations in Grade 5 classrooms suggest that the 3-year emphasis on matter as decomposable into arbitrarily small pieces lays the groundwork many students need to understand matter as made of particles.

Table 2 shows the Grades 3–4 and the Grade 5 stepping stones that were the targets of the Inquiry Project Curriculum. We imagine that, when the Inquiry Project Curriculum includes a pre K-2 section, some third grade and fourth grade content may move to earlier grades. As a result, there will be time to teach the whole Grade 5 stepping stone (i.e., to include density). In this revised curriculum, while the particulate model may precede density for the reasons stated above, both the model and the density concept could be taught together. Whether one approach is better than another is an empirical issue.

We conclude this section by noting that, although it may seem there are numerous curricular options for each grade, this is not so in our theoretical framework. Once the constraints imposed by the cognitive mechanisms of conceptual development are taken into account, the number of options decreases dramatically. In the next section we argue that the order in which core concepts should be introduced (and therefore which core concepts should be focused on in each grade) is also constrained by the mechanisms of conceptual development.

Challenges in Deciding in What Order to Introduce Core Concepts: Lever Concepts

Our descriptions of the Grades 3–4 and Grade 5 stepping stones, as we structured them around core concepts, only established broad learning goals. Before writing the curriculum, we had to decide on the order of the introduction and revision to those core concepts. How would we make this decision?

Challenges arise from the multiplicity of the conceptual changes that must be coordinated. The reconceptualization of matter from continuous to discontinuous, as explained in the Introduction, gives a flavor of the complexity involved. First, concepts change in small steps; the ordering of these steps must be optimized. Second, concepts cannot be revised in isolation; by nature, conceptual change is a change in *beliefs* (i.e., in relationships among concepts). For example, the change from "air has no weight" to "air has weight" is a change in belief that requires and amounts to the reconceptualization of both air and weight. Third, conceptual changes are interdependent. We have mentioned that understanding that gases share significant similarities with solids and liquids requires that student have made some important revisions in their concept of weight. Further revisions to the concept of weight, however, depend on a more advanced concept of matter. For example, reconceptualizing weight as a gravitational force is only possible when the concept of mass is understood or perhaps as that concept is being developed.

Therefore, curricular units should foreground relationships among concepts (rather than focus on one concept at a time, completing work on it before moving to the next, as current curricula often do.) Given that each curricular activity should focus on one conceptual relationship at a time, and that those relationships are not yet in their scientific form, core concepts have to be revisited several times across a grade range, and sometimes within one grade, in different combinations and in different contexts. The order of these "visits" is important, with some conceptual relationships more amenable to change early on than others.

Which conceptual relationships are third graders more likely to revise productively? We propose that these are the relationships between core concepts that are already "mental units" (i.e., concepts with rich content in the lower anchor) that allow students to make sense of a variety of experiences in a relatively consistent way. These core concepts, which offer many contact points with instructional materials, help students "break into" the new system of ideas (the stepping stone) in several ways. Thus they offer multiple sources for conceptual changes. Relationships between these core concepts can be brought significantly closer to their form in the stepping stone in a relatively short period of time; at the same time, they constitute important *revision*s of the core concepts. For example, linking weight to amount of material is an important revision of the concepts of weight and material. Once revised, these core concepts may enter into the productive revision or the construction of other core concepts (e.g., revised concepts of weight, size, and amount of material enable the differentiation of weight of objects, heaviness of materials, and volume). These core concepts provide the "most bang for the buck." We call them *lever concepts*.

A lever concept has a temporary and shifting status, defined relative to its role in the knowledge progression from one stepping stone to another. Hence, the concepts that function as lever concepts in moving from the lower anchor to the Grades 3–4 stepping stone are different from those central to moving from the Grades 3–4 stepping stone to the Grade 5 stepping stone. There are three criteria a core concept must meet in order to qualify as a lever concept for a particular transition: (a) it is already present as a distinct mental unit at the beginning of instruction where it is involved in relevant generalizations about the world; (b) it requires reconceptualization in order to be integrated coherently in the target stepping stone; and (c) its reconceptualization supports the revision or construction of other concepts in the target stepping stone. In the next section, we explain why we chose *size, weight, material,* and *amount of material* as lever concepts for Grades 3–4, and why we chose *solid and liquid materials, particle,* and *heavy for size* as lever concepts for Grade 5.

Ultimately, whether a core concept functions as a lever concept is an empirical matter. If sufficient research exists, it is possible to determine which concepts are rich in pedagogical possibilities and require reconceptualization. However, it is only possible to hypothesize whether these concepts can be revised successfully in a reasonable amount of time and whether they will drive the revision or construction of other concepts. For example, the literature on young children's concepts of weight and some specific materials is sufficient to justify tackling those concepts first. Yet there is no way to predict how long it takes third graders to develop an objective, extensive concept of weight and a concept of material such that these concepts can contribute to constructing other concepts. Therefore, a challenge in designing the Inquiry Project Curriculum, was that we had to be ready to change our plans for one grade based on classroom experiences in the preceding grade.

Lever concepts in Grades 3–4. In many elementary and middle schools, instruction in solids, liquids, and air begins with definitions. ("Air is a gas. Solids, liquids, and gases are matter.") We explained in the Introduction why we think this approach fails (see also Johnson and Papageorgiou, 2010, for another critique of this approach). Even when taking a more step-by-step approach, we argue instruction should not begin with solids, liquids, and air (i.e., they are not lever concepts for the third grade). Children's conceptual understanding of solids, liquids, and air is too distant from the content of the stepping stone; many other concepts need to be in place before solids, liquids, and air can act as precursors to matter. Children know a lot about different solid objects and some liquids but little about solids and liquids as categories. Moreover, they think of solids and liquids as very different (some think liquids do not have weight) and think air is "nothing." Relating these concepts to each other too early in the curriculum does not help children learn new, productive relationships among the concepts because they have nothing to build on. In particular, as we have argued in the chapter, until students believe that any amount of liquid or solid material has weight, understanding that air is material is very difficult.

Weight, size, material, and amount of material, on the other hand, have a much richer initial set of connections that children can build on–connections to each other and to the physical world. Many two-year-olds know the word "heavy" (Hart & Risley, 1999; Wiser, Citrin, Drosos, & Hosek, 2008). By age three, many children can compare weights qualitatively by hefting (Wiser et al., 2008). Many four-year-olds have established empirical relationships between weight and size (bigger things are heavier) and between weight and some materials (steel objects are heavier than wood objects) (Wiser et al., 2008). Heft generalizes to "pushing on objects." For example, if A feels heavier than B, young children expect that A will tilt a pan balance scale and that A is more likely than B to break a small foam bridge (Metz, 1993; Smith et al., 1995; Wiser et al., 2008). Kindergartners' knowledge of specific solid materials (e.g., wood, plastic, metal) includes properties that are characteristic of materials in a scientific theory (e.g., hardness). Therefore this knowledge lends itself to productive generalizations (e.g., some materials are harder than others) (Wiser & Fox, 2010). Second graders know that the amount of solid or liquid material remains constant as a sample changes shape and nothing is added or removed. They also link weight and amount in the sense that adding clay (e.g., to a clay ball) makes it "more clay." Finally, they know about size: they can qualitatively judge relative lengths and areas, and whether an object will fit into a particular box or go through a particular hole. However, they are not always clear about the difference between length and area and do not have a concept of occupied 3-dimensional space.

We assumed that with an appropriate set of learning experiences, the third graders in our study could advance from measuring weight by heft to measuring it quantitatively. We also assumed they could relate weight to amount of material and could develop a concept of material as an ontological category (kind of stuff things are *made of*) and as related to weight (when objects of the same size are compared, those made of some materials are heavier than those made of other materials). These new conceptual relationships bring the concepts of weight, material, and amount of material closer to the concepts in the Grades 3–4 stepping stone. We also assumed that drawing on their informal understanding that physical objects occupy space and vary in both their size and weight, fourth graders would understand that water displacement depends on the size, rather than the weight, of the object, and would begin to construct a measurable objective concept of volume as occupied space. Thus size, weight, material, and amount of material are the lever concepts in reaching the Grades 3–4 stepping stone. In contrast, the concepts of volume and heaviness for size (a precursor of density) are new concepts constructed as part of achieving the stepping stone, but they are not lever concepts (because they were not present as distinct concepts in the lower anchor).

Lever concepts in Grade 5. We assumed that, by the end of Grade 4, students' knowledge about solids and liquids would have been reconceptualized in major ways, and include the general concept material. Objects are made of (in the sense of constituted of) certain types of materials. These materials have specific properties, not all of them perceptual; liquids are specific kinds of materials. We also expected that students would have developed a concept of particle (from repeatedly dividing chunks of material physically and mentally) that included the belief that any piece of material has weight and occupies space. Implicit in such knowledge about solids and liquids is a new concept of matter – "what has weight and occupies space" (distinguished from "what can be touched and seen") and "exists in the form of different materials."

These concepts of material and particle fit the criteria of lever concepts for the fifth grade: they are rich and salient and require (further) revision. As they develop a particulate model of matter, students realize that the same material can sometimes be in solid, liquid, and gaseous form, and that particles of material are not merely emergent (when you grind things), but are pre-existing microscopic pieces that have characteristics not observable at a macroscopic level (e.g., spaces between particles, motion). The revised concepts of material and particle enter into the construction of the new explicit concepts of matter and mass that are part of the Grade 5 stepping stone. Heavy for size is also a lever concept in the fifth grade. We assumed that the fourth grade curriculum would make this concept more explicitly differentiated from weight in the context of solids and liquids. In the fifth grade, this lever concept, which may be generalized to samples of gases and explained by a particulate model of matter, prepares students for the study of density in middle school.

We believe the construct of lever concepts can be applied effectively to other grade ranges. For example, in the pre K-2 grade range, amount of material and solid materials are good candidates for lever concepts (they continue as lever concepts in Grade 3). In the Grades 6–8 range, matter and mass become lever concepts in addition to those in Grade 5.

Challenges in Designing Learning Experiences That are Meaningful and Productive

As we noted in the Introduction, there are challenges in designing meaningful and productive learning experiences. How can one present meaningful information to students, i.e, that can be interpreted in terms of their existing concepts and beliefs, and at the same time change those concepts and beliefs to make them more compatible with scientific knowledge? Recall that lever concepts such as weight, size, and material exist in the lower anchor in "recognizable form" but are incommensurate with their scientific counterparts. Constructing a scientific meaning for one term (e.g., weight) depends on constructing scientific meanings for other terms (e.g., material). This in turn implies revising many of the relationships among the initial concepts as well as between them and other concepts (e.g., hardness, heft, liquid, and solid). It also implies constructing (almost) entirely new concepts (e.g., volume).

Our overall strategy is to reduce the incommensurability gap between students' and scientists' knowledge networks progressively and coherently. Stepping stones use this strategy for coordinating curricula across grade ranges by identifying useful "intermediate" goals between the lower (kindergarten) and upper (Grade 8) anchors. These stepping stones are relatively coherent and hence provide new equilibrium points in the knowledge network. We now focus on the challenge of using the "small-step but coherent strategy" within a grade. Small steps are not enough. Given the heavily interconnected nature of knowledge about matter, the new small pieces of knowledge must be integrated. Moreover, ideally these small pieces should be integrated as they accumulate and develop so that their integration does not result in a "big conceptual surprise" (e.g., "What do you mean hefting is not a good way to measure weight?").

We propose that achieving this progressive integration and transformation within a grade involves engaging students in progressive cycles of model construction and revision, and in threading instructional experiences around *key representations.* These key representations are mental models and external symbolic tools that support reasoning about physical quantities and conceptual change. A central challenge to our approach involves identifying these key representations and then using them in the curriculum. We highlight them as distinct components in LPM because we think they are critical components in conceptual change that are frequently overlooked as important targets of instruction in traditional science curricula.

For example, in the Inquiry Project, we identified the *compositional model of material* as key to the overall reconceptualization of materials as underlying constituents of objects. Applying a compositional model to a chunk of material involves mentally decomposing it into identical pieces and imagining that each piece retains the identity and properties of the material while, as a group, the pieces still constitute the original whole. We argue elsewhere that a compositional model may contribute to enriching the belief in the conservation of amount of material and may support the conservation of material identity and the reconceptualization of the concepts of weight and volume (Wiser, Smith, Asbell-Clarke, & Doubler, 2009). The compositional model also allows imagining repeated divisions and concluding that pieces too small to see exist and retain the properties of the chunk. The model may also prepare students for taking the next step in further model development and revision—a *beginning particulate model of matter*. As central as the compositional model is likely to be for reconceptualization, it is generally overlooked as an important target of elementary school instruction because it is a *mental model* constructed by *thought experiments*.

In the Inquiry Project, we also introduced *measure line representations*, first for weight and later for volume. These representations are used to reconceptualize these physical quantities as measured additive quantities and as (continuous) dimensions. Used in combination with a (mental) compositional model of material, *measure line representations* can effectively support the belief that small pieces have weight and occupy space. Yet the *conceptual* importance of measure line representations has also not been generally acknowledged by traditional science curricula.

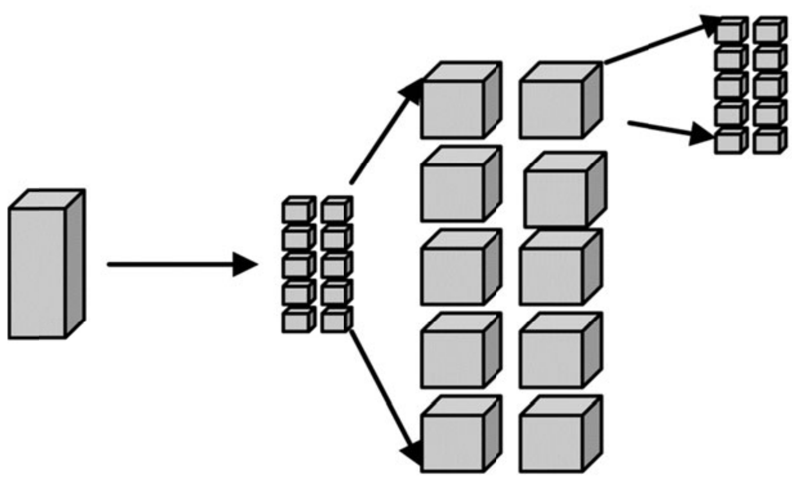


Figure 1. Students can start cutting a chunk physically, then turn the operation into a thought experiment. The picture represents mentally enlarging the chunk in order to cut each piece into smaller and smaller pieces.

In this section, we begin by examining two cases in some detail—one involving weight where we were reasonably successful in using measure lines as key representations and one involving volume where, for a variety of reasons, we were less successful. Our initial step may have been too large in the second case. The new representations we introduced may not have been as meaningful for students as we thought, and we may have overlooked other new representations that should have been included. We also discuss the potential for generalizing this approach to other age groups.

As part of this discussion, we include some results from a three-year longitudinal study of two cohorts of students at two schools. The traditional school curriculum is used with one group (the Control students); the sequence of units in the Inquiry Project Curriculum is used with the second group (the Treatment students who are a year behind the Control students). A subset of students in each group participated in the same intensive two-hour, individual interviews that probed their understanding of a number of concepts (e.g., matter, material, weight, density, length, area, volume, ratio, proportion, number) in diverse ways at multiple time points. The Treatment students were interviewed before the start of the Inquiry Project Curriculum and at the end of third, fourth, and fifth grades. The Control students were interviewed at the end of third, fourth, and fifth grades.

Example 1: Threading instruction around the weight measure line to facilitate the reconceptualization of weight as an extensive quantity in the third grade. Two main topics of our third grade curriculum were the extensivity of weight and the nature of good measurement. Weight is an important aspect of children's physical world. By the time they enter third grade, students have learned many relationships between weight and other physical concepts that can be used in the curriculum. However, for the majority of third graders, hefting is privileged as a way to measure weight—how heavy an object feels is how heavy it is. Thus weight is not extensive in the lower anchor. Twice the amount of material does not (reliably) feel twice as heavy, Styrofoam weighs nothing, and tiny things weigh nothing so adding small amounts to something does not change its weight. Although some third graders may realize that balance scales can be used to compare the weight of objects, few of them understand how to use those scales to "measure weight," or the underlying logic of weight measurement.

It is a mistake to think that learning about the extensivity of weight is achieved simply by adding or combining new knowledge (e.g., about measuring weight with a balance scale) with the heft sense of weight. "Combining" suggests enriching existing knowledge. In the case of lever concepts, existing knowledge needs to be transformed as well as enriched. Simply showing that heavier objects "make one side of the balance scale go down" adds little to students' knowledge of weight. If a conflict is introduced (e.g., using a small piece of steel that feels heavier than a big piece of wood, but actually is lighter), some students may simply ignore the conflict—the object that feels heavier is heavier and "I don't know how a scale works"; others may try to explain why the "lighter" object makes the scale go down (e.g., by invoking the idea that size affects the scale as well). Clearly, learning the steps involved in using a scale does not bring the students closer to a scientific concept of weight. Identifying the difference between scale weight and felt weight is only a first step in the complex process of reconceptualizing weight as an extensive quantity. The scale is one tool for reconceptualizing weight. The weight measure line is another. Carefully sequenced activities using both tools in interaction allow students to coordinate multiple discoveries and insights and to integrate the new with the old.

Measure lines scaffold the reconceptualization of weight (and later, volume) in the Inquiry Project Curriculum. They are representations, not concepts. Although they are not part of the lower anchor, once introduced, they make (some) sense. They are constituted during classroom activities from knowledge elements in different domains and are sources of thought experiments, inferences, and discoveries about the physical variables they represent. Through guided inquiry, they come to progressively express structural aspects of concepts and/or relationships between concepts in the upper anchor. Without them, hands-on experiments and observations would enrich the knowledge network in the lower anchor but would not move it toward the upper anchor.

In the sub-sections below, we identify the elements that we recruited from the lower anchor in constructing an understanding of a weight measure line. Then we analyze the individual steps in this process and how the use of the weight measure line facilitated the integration of these steps as well as the discovery of something new.

Overview: Elements recruited from students' lower anchor to construct, use, and learn from using a scale and a weight measure line. Hefting and using the balance scale are two productive entry points for revising children's concept of weight. Using a pan balance scale to establish whether two things have the same weight, and if not, which is heavier is intuitively obvious to third graders. They understand that objects push on other things as they push on their own bodies. It is also obvious to third graders that two identical things are heavier than one, and that adding stuff to an object increases its weight (as long as the piece you add is big enough).

Third graders also know about integers. They have an implicit understanding of cardinality—the last count word represents the cardinality of a set (i.e., it tells you "how many objects there are"). They know that if you combine two sets of objects, you can compute how many objects there are by adding the numbers of objects in each set. They can also represent integers on a number line.

All these pieces of knowledge in the lower anchor can be organized and coordinated by involving students in measuring weight using a pan balance scale and representing their measurements on a weight line. The weight line can then be used to support computations and inferences. Coordination and inferences are sources of genuine new knowledge about weight and its reconceptualization.

Steps 1 and 2: Linking felt weight and scale weight by constructing a weight line. In the third grade Inquiry Project Curriculum, the weight line begins as an actual linear array of objects (density cubes)[10] arranged according to increasing felt weight (Step 1 in Figure 2). Uncertainties about some pairs of cubes motivate the use of a pan balance scale. Students readily relate the function of the scale to their own hefting, a similarity that helps establish the pan balance scale as a good way to compare weights of objects. As they realize that the pan balance scale is both more sensitive and more reliable than hefting, their focus switches easily to arraying objects according to comparisons with the pan balance scale. The concept of weight, at this point, is enriched because "The scale hefts more accurately than I do." The weight line displays weights ordinally, without numbers (see Step 2 in Figure 2).

The next investigation ups the ante—children move from the question of which cube is heavier to the question of how much heavier. This question cannot be answered without quantifying weight. Very problematically, the question is relatively vague or metaphorical until weight is measured in terms of units and the meaning of "no weight" is related to those units. The lower anchor concept of weight lends itself to qualitative comparisons only—A is heavier than B, and B is heavier than C—and is judged holistically—how heavy an object feels—rather than as the sum of identical weight units. This is in part because weight is not yet a property of amount of material—it is not extensive. Thus the quandary arises: the question is meant to lead to quantification of weight, which will give it extensivity, but the question does not make much sense unless extensivity is in place.

How can classroom activities around a question that students cannot fully grasp be productive? It is the joint use of the pan balance scale and the weight measure line that shapes the meaning of the question in a series of progressive steps outlined in the next section. Shaping the question goes "step in step" with reconceptualizing weight.

Steps 3 and 4: Measuring weights with weight units and using the weight measure line. We start with activities using the pan balance scale and non-standard units—plastic bears, paper clips, and washers (Step 3 in Figure 2). These activities help third grade students discover the need for a uniform and shared system of weight units (Step 4 in Figure 2).

Measuring the weight of a density cube with a pan balance scale and non-standard units can be understood with the lower anchor concept. How many plastic bears have the same weight as the cube? (i.e., "How many bears does it take to balance the cube?") is a meaningful question for a student who has the lower anchor weight concept. The "weight" part of the question (making the weights equal) is qualitative. The answer—"7 bears"— is quantitative, but it is not weight that is quantified (yet). Rather, it is the number of bears with a weight equal to the weight of the cube.

The number of bears can be displayed on a number line, a process familiar to young students. But the line is already a weight line because it has been used to represent the qualitative ordering of the weight of the cubes. Without knowing it explicitly, students are blending, in Fauconnier and Turner's (2003) sense, a number line and the qualitative weight line. That is, they are applying the properties of numbers (at first, of integers) to weight.

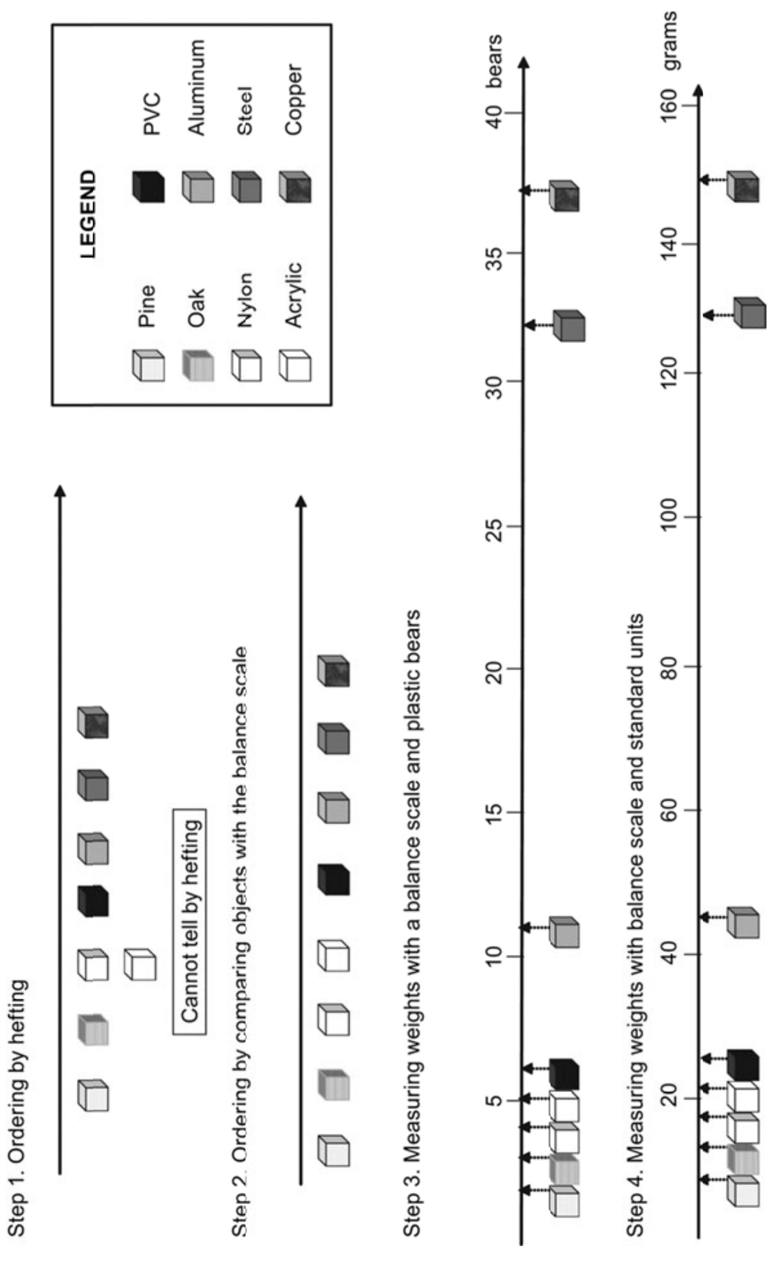


Figure 2. Using the weight measure line to represent qualitative and quantitative ordering of the weights of the density cubes.

A first reconceptualization involves the weight line and language. The teacher shifts the emphasis from "How many bears have the same weight as the cube?" to "What is the weight of the cube in bears?" and then to "What is the weight of the cube? "Seven bears." This shifts "bear" as an object with a certain weight to "bear" as a weight unit. As the meaning of "bear" shifts, students are developing a sub-concept of weight—the beliefs that weight can be measured with a pan balance scale and that it can be assigned numbers give weight an extensive and quantifiable aspect. This sub-concept is compatible with the concept of weight in the lower anchor (the blue cube feels heavier than the green cube, therefore it makes its side of the balance scale go down) and enriches it (the weight of the blue cube in bears is greater than the weight of the green cube).

Once this first reconceptualization occurs, using a pan balance scale to measure weight in grams is a meaningful activity. (Without scaffolding using non-standard units first and without the weight line, the same activity might been a routine with shallow meaning.) Students use a set of gram weights to measure the weights of the density cubes. Students then place the density cubes along the weight line, according to their weights in grams (Step 4 in Figure 2). They can now discuss the question "How much heavier (or lighter) is one object than another?" much more meaningfully.

Step 5: Drawing inferences from the weight measure line. Students can start exploring new questions. Are there objects that are just one-gram difference in weight? Can objects be less than a gram apart? Are there objects that have a weight "between" the weights of the objects on the weight line? The fact that there are other weights "between" the weights of the density cubes becomes more apparent as children imagine cubes made of other materials (e.g., clay, soap, stone, glass, brick, concrete) and speculate where their imaginary cubes might be on the weight line and why. These activities enrich the new sub-concept—the extensivity and quantifiability of weight—and the meaning of weight units. They also help develop the notion that weight is a continuous variable (an issue not addressed in this chapter) by helping students see that there can be weights between any two points on the line.

To help students appreciate the additive nature of weight, teachers can ask students, "How much weight would need to be added to the pine cube to have it be the same weight as the oak cube?" Although this question is similar to the one asked earlier ("How much heavier is the oak cube than the pine cube?"), the question might be answered directly and empirically using the pan balance scale. Place the pine cube in one pan, the oak cube in the other pan and see how many grams need to be added to the pine cube to balance the scale. It can also be answered by using the weight line, although how to do so is not obvious and is challenging for many students. The isomorphism between (a) physical actions with cubes, grams, and the pan balance scale and (b) counting line segments and reading marks on the weight line gives the weight line its "measuring" meaning. In other words, students develop the implicit understanding that the conclusions one reaches by reasoning with the weight line are true of the real world. They learn to validate the weight line as a model of weight.

More generally, students will internalize the structure of the weight line as the quantified structure of weight. One weight unit is represented by one line segment on the weight line. The weight of an object is represented by as many contiguous line segments as there are grams balancing the object on the scale. When the student reads, "The weight of this object is 15g" from the mark on the line, the student realizes it is the sum of 15 weight units.

Step 6: More inferences from the weight measure line and the discovery of the relationship between weight and material. As students become accustomed to thinking of weight along the weight line, interesting questions arise about smaller and smaller pieces of material. For instance, many students do not immediately understand that there can be values on the weight line between, for example, 3 and 4 grams, or more interestingly, between 0 and 1 gram. The visual representation of weight makes this idea graspable, especially if it is linked to the idea that a more sensitive scale would discriminate between these different weights.

Can one make pieces of stuff small enough that they weigh nothing? Students break a 4-gram piece of playdough into two pieces, place one piece on the 2-gram mark, then divide it again, place a piece on the 1-gram mark, and so on, closer and closer to the 0 mark. This leads to questions for discussion: "Will you ever reach 0?" Do you think a tiny piece could weigh nothing at all?" The weight line is part of the discussion. Representing 1 gram with a larger line segment allows students to keep dividing the playdough further. The weight line is also linked to the real world—a very sensitive scale would detect a very small piece. Many students conclude from these (and other) activities that indeed, any tiny piece must have weight.

Weight line discussions also help give meaning to the zero point. For example, we found it is not initially obvious to students that weight lines begin at 0. Some students think light objects weigh less than 0. Furthermore, when students consider what happens with repeated division, some students, who confuse repeated division with repeated subtraction, reach negative numbers. Part of the confusion is with students' understanding of fractions, numbers, and division. These weight line discussions are an important context for grounding the meaning of fractions and division operations. In addition, if weight is tied to amount of matter, the connection motivates student thinking that something could have no weight only if there was no matter. (See also Carraher, Smith, Wiser, Schliemann, and Cayton-Hodges (2009) for additional discussion of issues about how children conceptualize weight, length, and number as dimensions, including the specific role of number and measure line representations.)

Interim conclusions: Weight line as a tool for reconceptualization. Both classroom observations and post-interview data in our study suggest that the sequence of activities and the discussions were productive in helping children think about weight in new ways. The weight line linked felt weight and scale weight while privileging scale weight and linking elements of weight measurement to knowledge about numbers and counting. The weight line also linked dividing an amount of stuff and dividing a line segment on the weight measure line, thus supporting the idea that even tiny pieces weigh something and that the weight line starts at 0 (when there is no stuff). Indeed, by the third grade post-test, 60% of the Treatment students now confidently argued even tiny pieces of clay must weigh something and must take up space (versus 11% of the Control students and 7% of the Treatment students at pre-test). This strongly indicates that weight was now more centrally linked to amount of material than hefting. Furthermore, this reconceptualization of weight was stable and contributed to other productive changes, such as the beginning of a strong precursor concept of density by the end of Grade 4 (Smith et al., 2010).

Example 2: Using cubing as a tool for constructing a 3-D understanding of volume. While students successfully integrated their learning experiences about weight, developing their understanding of volume proved more difficult. There was ample evidence in our pre-interviews that third graders lacked an understanding of the concept of volume. They either did not know the word "volume" or did not know it referred to occupied, 3-dimensional space. In asking the following question, "Which (of two rectangular prisms) takes up more space?" the interviewer gestured with her hands, encompassing a 3-dimensional region of space. The question was interpreted as "occupying more area on the table." When asked to use a set of cubes to measure which prism took up more space, most children used the cubes to measure lengths or areas. They certainly had a concept of "big," but it had multiple, non-differentiated meanings—tall, long, wide, occupies a large area, is difficult to carry, and presumably "won't fit in a small (physically bounded) space such as a trunk." But they did not seem to understand that an object occupies an unbounded region of 3-dimensional space. When questioned about water displacement, almost all the third graders thought that the heavier object would make the water level rise more (although the two objects were identical sizes, i.e., had the same height and diameter). Almost all students said that when a clay ball was flattened into a pancake, it now took up/filled up more space, pointing to the areas (Smith et al., 2010). Thus we propose that volume, unlike weight, does not exist in the lower anchor. The challenge here is to build on existing concepts (weight, amount of material, various meanings of "big") in order to construct an entirely new concept.

Our first attempt at teaching volume was at the end of our third grade unit. Students compared how much space objects with visibly different volumes took in a paper bag and placed the objects along a volume measure line. Then we invited the students to construct little replicas of four regular solids from small cubes in order to determine "how much space" each took up. (The four regular solids had different dimensions and shapes but were not visibly different in volume). One hypothesis was that "cubing" the solids by building little replicas would show students they were not measuring length or area. Rather, they were stacking cubes along three dimensions, and the number of cubes was not related to any single dimension. Moreover, comparing the number of cubes would give them a sense that volume is an attribute of objects quantified by a number of equal sized, 3-dimensional objects, thereby increasing their understanding of the three-dimensionality of volume.

Unfortunately, both classroom observations and post-interview data showed that we greatly overestimated how meaningful these activities would be for the third graders. In class, the students used the cubes in many, very imaginative ways other than building replicas of the regular solids. Typically they placed the cubes along one dimension or one face, or “encased” the object with cubes. Although teachers worked with students and showed them how to use the “correct” strategy for making replicas, it was not clear that the students understood why this was a better strategy or what they were really trying to measure. In retrospect, cubing was nothing more than building a copy of an object with cubes. Unsurprisingly, our third grade post-interviews revealed that the students showed limited progress in using cubes to measure the volume of two regular solids. There was no (statistically significant) difference between the Treatment group and the Control group.

These results occurred despite our use of similar sequence of experiences for weight in which we first drew on comparison of objects with visibly different volumes and then built a case for measurement when the difference between two objects, with similar volumes, cannot be determined by sight. However, this ignored some profound differences between volume and weight: weight was already a clear “unit of thought” ready to be linked with the scale because a scale works as hefting does—heavier objects push on it more. This in turn provided meaning to weight units and to scale “reading”: the number of weight units could be linked to students’ concept of weight and helped make it more scientific. In contrast, there was no distinct “unit of thought” corresponding to volume to link cubing to; unfortunately, cubes can be used to measure length, height, perimeter, and area, which are all conflated in “big.” Our “step” was too large; we were unable to connect with what children already knew in a productive way[11] because the number of cubic centimeter (cc) cubes did not make contact with a precursor of “occupies 3-dimensional space.” Measure line representations are effective in highlighting the extensivity (and dimensional nature) of a quantity children are initially treating categorically or ordinally; thus, they do not address the fundamental problem the children had with volume—understanding the quantity referred to in the first place.

We suspect that our cases for volume and weight had different results because of the nature of children’s experiences with objects. First, weight is relevant in many familiar activities—wondering, “Am I strong enough to move this from here to there?” or in hearing parents’ admonitions, “This is heavy, don’t drop it!” or in observing that heavy adults squash toys when they accidentally step on them. Second, heft is an explicit aspect of human haptic experiences whereas volume is not; we are explicitly aware of the width and cross section of objects because we take them into account when we grasp and hold objects. However, grasping and holding objects usually does not require information about their volume.

We revisited volume in our Grade 4 unit on Earth materials. In this unit we provided students with a richer, broader, and more scaffolded range of experiences aimed at developing a sense of occupied 3-dimensional space in interaction with measuring amounts of liquid, granular, and solid materials. Students first compared the weights of samples of different granular and liquid materials that reached the same height in containers of the same size and shape. In this exercise they learned to distinguish the weights of materials from their amounts (this activity generalized to samples of liquid and granular materials the distinction between weight and size students had learned in third grade in the context of the density cubes). Students then compared the amounts of water in different shaped containers by creating a "fair test"—they poured each sample into the same shaped container. They then learned to measure amounts of water in cc's in a variety of activities. One activity was constructing their own graduated cylinder, marked in units of 20 cc's, by filling a 4 x 5 cc rectangular container first with cc cubes to verify its volume and then with an equal volume of water, and marking the height of each 20 cc's of water that was poured into the larger cylinder (see Figure 3).

In these activities students were learning to measure amounts of liquid and to relate the size of regular solids (cc cubes) to a specific amount of liquid. We then built on students' knowledge that two objects cannot occupy the same location at the same time. They first established that water displacement depends on size rather than weight by using same size and shape cylinders that differed only in weight and material, and a plasticine cube paired with a smaller aluminum cube of the same weight. Students could explain their experiences meaningfully in terms of their new extensive concept "scale weight" (a concept necessary to establish that the two cubes were the same weight) and of their concept of "big" (because the cubes differed in all three dimensions). Weight does not matter; only "size/bigness" matters. Students then measured the amount of water displaced by irregularly shaped rocks of visibly different sizes. We suspect that they could conceptualize this activity as "The level rises because water and rock cannot be in the same place at the same time," in part

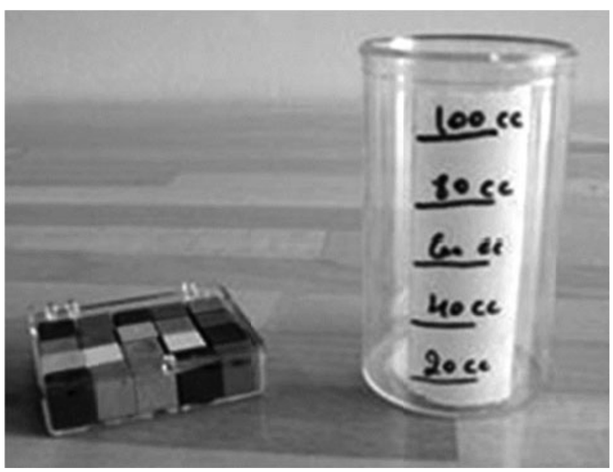


Figure 3. Students were asked to construct their own graduated cylinder for measuring the volume of liquids in the following way. First, they determined that a rectangular container had a volume of 20 cc's by seeing that it held 20 cc cubes. Then they emptied the container, filled it with water, and poured the water into the larger cylinder container to mark off units of 20 cc.

because other curricular units showed them that liquids are importantly similar to solids (liquids have quantifiable weight and amount and can be "broken down" into small pieces). Understanding the greater similarity of solids and liquids therefore allowed students to think of a water sample as an object and to extend their childhood belief that "two (solid) objects cannot be in the same place at the same time" to "an object in the beaker cannot be in the same place as the water." They could also establish the correlation, "Bigger rocks make the water rise higher." Integrating these two conclusions leads to the understanding that "bigness" refers both to the size of the rock and the space the rock occupies in the water. Students learned to measure the space occupied by the objects they placed in water by the number of cc's of water displaced. The number of cc's of water displaced became a measure of an attribute of the objects placed in the water—their volume.

By the end of the fourth grade, students in the Treatment group performed significantly better on the volume measure tasks and water displacement tasks in our clinical interviews than the Control students. Treatment students scored 40% and 58% for volume measure tasks and water displacment tasks, respectively. Control students scored 11% and 15% for volume measure tasks and water displacement tasks, respectively (Smith et al., 2010).

Of course, by the end of the fourth grade, many Treatment students still confused volume measurement with other spatial measures (especially the area of the face of a cube or its total surface area). This confusion suggests the curriculum design can be improved. To improve the design, we are considering revisions to the sequencing of activities. We could start with displacement activities before we introduce formal measurement of volume. We could then link these activities to the gradual development of new forms of symbolization (e.g., having children construct and discuss 2- and 3-dimensional drawings, including drawings linked to water displacement). We think that involving students in constructing their own drawings, although a challenging task for them, may encourage higher-level analysis and reconceptualization than simply working with cubes.

Other valuable key representations at earlier and later grades. Identifying key representations (including new forms of symbolization) that are pedagogically effective for earlier and later grades presents its own challenges. Exactly what aspects of reconceptualization does a key representation address? How do different forms of representation interact in the process of reconceptualization, or build on one another over time?

Representations that support reconceptualization may not be obvious, especially in the early elementary school years. See Acher and Arcà (2006) for interesting examples of early forms of representation in very young children, including gesture, body movement, and drawings. Richard Lehrer (and colleagues), David Carraher, and Analúcia Schliemann have presented other ideas about key representations that may be important to develop in young children (e.g., Carraher, Smith, Wiser, Schliemann, & Cayton-Hodges, 2009; Lehrer & Pritchard, 2002; Lehrer & Schauble, 2000; Lehrer et al., 2001; Schliemann & Carraher, 2002). We share their view that mathematics and science need to be closely linked from the beginning of schooling. A major difference between the third graders' concepts and scientific concepts is that scientific concepts (mass, weight, force, density) are measured quantities or include quantification (e.g., materials are characterized by their melting points). The "small-step" approach almost dictates that students work on quantification very early in their schooling. One challenge therefore is to find ways to make quantification meaningful in the early grades and to specify the role visual representations play in that respect.

For older students, particulate models of matter are relatively obvious representations (e.g., Johnson, 1998; Margel, Eylon, & Scherz, 2008; Merritt & Krajcik, 2009; Nussbaum, 1997; Snir et al., 2003). Johnson's longitudinal studies illustrate the importance and challenge of specifying the elements that should be included in an initial model. These studies also show how to make the model progressively more complex. Another important representation is the periodic table of elements.[12] It is one thing to display the table in its final form and another to see it as a tool to be developed and co-constructed progressively. IQWST[13] starts linking discussions of the periodic table with beginning particulate models as early as Grade 6 and then deepens children's understandings in the next two years (Shwartz, Weizman, Fortus, Krajcik, & Reiser, 2008). How the progressive construction of particulate models of matter and representations of the periodic table can productively interact in the middle-school years is an important topic for conceptual change researchers to study.

CHALLENGE 2: HOW CAN LP RESEARCH AND CURRICULUM DEVELOPMENT WORK HAND IN HAND?

Most curricula are evaluated (at best) based on how well they "work" in large-scale field-testing. The criteria for "working" might include that the curricula were sufficiently well-specified that teachers could use them as intended, that both teachers and students liked using them, and that students made measurable learning gains when using them.

In contrast, the criteria for evaluating both LPs and their curricula are tied to ongoing (basic) research on children's learning. Thus a LP approach to curriculum development calls for marrying sometimes conflicting goals and time scales for research and development. LPs are ultimately hypotheses about how knowledge develops and changes over time. They are generated from empirical research on students' learning (or lack thereof), which depends on the curricula and other learning experiences. LPs are revised as follows: a LP is translated into a curriculum; if the curriculum is more successful than traditional curricula or other curricula based on other LPs, then that LP is supported as a viable characterization of how knowledge in a domain can evolve toward an understanding of scientific concepts. The curriculum is then supported as a viable tool to bring about that evolution.

We expect the findings from the Inquiry Project (such as those presented above) will inform the evaluation of LPM and the Inquiry Project Curriculum in three complementary ways. First, we will compare the pattern of relationships among different aspects of developing concepts in the Treatment group to those hypothesized in the current version of LPM. Second, by comparing those patterns in the Treatment and Control groups, we will test the hypothesis that the Inquiry Project Curriculum is more effective in promoting reconceptualization in children's knowledge of matter network. Third, the comparison of those patterns will inform an important issue about LPs—how many are there? In other words, are there multiple ways to reach the upper anchor? We suspect that students' initial concepts strongly constrain knowledge acquisition in the same way across curricula with different learning goals. We predict we will find that if knowledge progresses toward scientific understanding, it does so in the same way in both groups. However, we also predict that progress will be more limited for Control group students and they may develop counterproductive ideas not found among the Treatment group students. In other words, we expect to find evidence that supports our view that the number of ways to link the lower and the upper anchors may be very limited. See Carraher et al. (2009) and Smith et al. (2010) for a description of some findings from the first two years of the study that relate to these issues.

Several challenges relate to the development of LPs and related curricula. One challenge is to limit the number of cycles of development/testing/revision this process requires. LPs, when constrained by a strong research base, are more likely to produce effective curricula at their first introduction. Iteration cycles are fewer if one can evaluate students' knowledge after each grade in time to modify the curriculum for the next grade. We noted earlier that we overestimated our third graders' initial knowledge about volume. Therefore, we took these limitations into account in developing the Grade 4 curriculum. Specifically, we realized that we had to foreground volume in Grade 4 and explore multiple aspects of volume in a series of interrelated investigations.

Another challenge in the development of LPs and related curricula relates to the time it takes to test and revise them for the whole K-12 grade span. One solution is to develop LPs for limited grade ranges (as in the Inquiry Project and the IQWST project) and to coordinate them later. In the domain of matter, this strategy has the additional advantage of testing some specific hypotheses embodied in the LPs. In the case of matter, integrating LPM with the middle school and high school matter LPs would allow testing the major tenet of LPM—building stronger macroscopic understandings of matter in elementary school facilitates the acquisition of AMT in middle school and high school. The problem this approach creates, however, is that the lower anchor of the higher-grade range of the LP is based initially on research among students who have not benefited from LP-based curricula in the previous grade range. Thus the curriculum for the later grade range, designed at first on the basis of that lower anchor, needs to be revised before being administered to students who have had the benefit of a LP-based curriculum in the earlier grades. We predict that such revision will be necessary in the Inquiry Project Curriculum after we have developed its pre-K-2 "prequel." Moreover, if different researchers develop the different segments of a LP and curriculum in a domain, theoretical constructs and pedagogical methods may require considerable alignment.

Other important questions are: If the curriculum is successful, to what do we attribute the success? Conversely, if the curriculum is unsuccessful, should the LP or the curriculum be revised? LPs are complex, multi-faceted hypotheses, and the curricula they inspire orchestrate many key elements for the promotion of learning. For this reason, it is difficult to isolate or test the factors that contribute most to their success. Indeed, it is highly probable that the orchestration of the elements, not the elements themselves, is the critical success factor.

Moreover, how do we measure success? Compared to what? Existing curricula? Alternative LPs? On what time scale? Most curricula produce some measurable gains but may involve long-term trade-offs. An important aspect of determining which approach gives "the most bang for the buck" is that one consider both short-term and downstream consequences. Currently, we are assessing the effectiveness of the Inquiry Project Curriculum compared to the current curriculum (or standard of care) in the schools of our study. We hope we can do better than that standard of care. However, we recognize it may be some time before we can test any alternative LPs against LPM.

Of course, the criteria for evaluating LPs are not just empirical. Another way to evaluate our proposed LP framework is in terms of its fruitfulness or generativity. For example, do core concepts, stepping stones, and lever concepts yield important insights when applied to LPs for other age spans or other domains? Do LP developers find them productive frameworks? We are currently developing a LP for energy for the elementary grades; this work suggests that the answers may be a qualified "yes." We found it very useful to organize our work around stepping stones working backwards from "What should eighth graders know about energy so that they benefit from high school science courses" to "What should fifth graders know about energy that will allow them to reach the eighth grade stepping stone?" to "What do young children know about…anything (!) that we can use to begin shaping concepts relevant for understanding physical phenomena in terms of energy?"

Those questions are more challenging for energy than for matter. The lower anchor in particular is an unknown. Young students know the word "energy," but what they know may be difficult to build on or may even interfere with learning. For example, having been bombarded with messages about the need to conserve energy may get in the way of learning one of the most important principles about energy—it is conserved, no matter what. Similarly, lay expressions such as "food is fuel" and "burning calories" are, at best, unhelpful. One can explicitly address and attempt to dispel such beliefs, of course, but not until late in the curriculum. Re-interpreting "we need to conserve energy" in a scientific way requires understanding energy dissipation, stored energy, energy conversion, etc. Similarly, understanding the difference between burning fossil fuels and processing food requires some knowledge of chemistry. How does one deal with those initial beliefs in the meantime? (In contrast, one can build a scientific understanding of solids and liquids without using the word "matter" and therefore with less interference from nonscientific beliefs about matter.)

Like anchors and stepping stones, core concepts and lever concepts are very useful constructs for energy. However, it is much more challenging to characterize them than it is for matter. Some core ideas are controversial among scientists—energy conservation may be on everybody's list, but should one teach about energy transfer or energy transformation? We suspect there may not be any lever concepts in the lower anchor because any knowledge in the lower anchor that we recruit and build on is not a direct, or even a recognizable, precursor of the “big ideas” about energy that make up the stepping stones.

Given that everything about energy is abstract, we suspect symbolic representations will play a large role in our curriculum. Again, it is one thing to borrow the idea from LPM, but another thing to implement it. Energy is a domain in which research and development will have to work together for a long time. Many iterations are needed. For example, we envision making educated guesses about core ideas (system, transfer, equilibrium, the relationship between energy transfer and change in a system, tracking energy in a system) and the sub-domains with which to start (perhaps heat and motion). We will begin our energy curriculum with developing curricular activities that will help us *find* lever concepts rather than beginning with a very good idea of what the lever concepts are and organizing the curriculum around them. Finally, it is still unknown how accessible we can make the big ideas to children in a different grade range.

The lesson we draw from the matter-energy comparison is that our theoretical approach and some of our specific solutions will be helpful in guiding our work with other LPs. However, we understand that core concepts, lever concepts, stepping stones and key representations take very different forms in different domains and therefore require much domain-specific research work.

CHALLENGE 3: HOW CAN LP-BASED CURRICULA FIT WITH CLASSROOM REALITIES?

The primary aim of the Inquiry Project is to examine a segment of LPM in which students strengthen their macroscopic understanding of matter and begin to develop a particulate understanding of matter (Grades 3–5). To support our study, we developed curriculum materials, formative assessments, and teacher professional development aligned with LPM.

Besides being the means to assess and then revise the curriculum and LPM, our longitudinal study provided a context in which to explore an important pragmatic question: *Is LP-based learning doable within the realities of schools?* We have learned that the answer depends mainly on three factors: teachers, time, and mathematical understanding.

Current science curricula are commonly topic-based. For example, students study weather, trees, and circuits. In contrast, a LP-based curriculum is concept-based. A few critically important concepts are at the heart of such study. This conceptual focus seemed foreign and less tangible to the teachers involved in our longitudinal study who struggled to understand the concepts and to fully appreciate their importance. They often began with the same understanding the students had. Teachers thought, for example, that a copper cube would displace more water than an aluminum cube of the same volume because it was heavier. The perception that their understanding was at times fragile made the teachers tentative in their teaching. In the professional development meetings, we focused on the core concepts, on how students encountered concepts in their learning, and on the difficulties students might have as they developed deeper understanding of these concepts. Yet, more support was needed.

For the LPM curriculum to be effective, teachers needed to think beyond the unit they were teaching. They needed to understand the full scope of LPM, the learning in the earlier grades, and the learning in the next grades. Knowing the broader landscape would help teachers connect to students' prior experiences and foreshadow understandings that would develop in later grades. Throughout the project, third, fourth, and fifth grade teachers met and worked together to understand all three curricular units.

Teachers also needed to understand the nature of scientific inquiry: science is a discipline in which multiple ideas are explored and evaluated against evidence. For many teachers in our study, this understanding required them to shift from teaching to the right answer to helping students increase their knowledge using investigation, analysis of data, and discussion. We realized that our curriculum was better positioned for developing conceptual understanding than curricula focused on discrete topics; however, with our curriculum, whole class discussions frequently lost focus. Teachers reported that they were "fishing" for how to make the discussion coherent and productive. More support for facilitating productive discussions was needed to ensure that the learning opportunities added up.

We met monthly with teachers from our study schools. Together, we tried out learning experiences from the curriculum, discussed the core aims of the curriculum, and considered issues of assessment and students' thinking about the core concepts. We joined the teachers in their classrooms as co-teachers of the curriculum. As well as informing the revision of the curriculum, this joint teaching experience helped us understand what we needed to do in order to implement the curriculum on a broader scale. Teachers needed additional help in understanding the learning progression and how to support student learning.

In response to teachers' learning needs, we began to develop grade-specific, Web-based professional development materials for the Inquiry Project Curriculum. These resources were designed to be available "just-in-time" to teachers as they teach the curriculum. They include videos in which scientists investigate many of the questions students ask and allow teachers to see the scientists engaged in scientific thinking. The aim of these professional development materials is to help teachers better understand the nature of science and the core concepts. These materials also include classroom cases designed to help teachers facilitate more productive discussions. Each classroom video case focuses on a particular learning experience from the Inquiry Project Curriculum. Thus teachers can see the learning experience and classroom discussion in another classroom before they teach the lesson. We hope these materials will help teachers incorporate new ways of teaching that support deeper conceptual learning.

Perhaps the hardest, unresolved challenge in teaching our curriculum relates to classroom time. In the typical elementary classroom, science is taught in 45-minute lessons, two or three times per week. Conventional science units are 12 to 18 lessons. We were determined to meet this convention. However, in the classroom, learning experiences often felt rushed. More time was needed for students to own the question, to become familiar with their own ideas, to work with data, to build explanations based on evidence, and to make meaning of their experiences through discussion. Often teachers, who had time-flexibility, extended a lesson to an hour or held the discussion on another day. We have worked hard to streamline the experiences, but continue to find limited instructional time an extremely difficult challenge.

Finally, we frequently overestimated the students' mathematical skills. Our curriculum simply could not get ahead of or ignore the realities of the school's mathematics program. For example, we anticipated that students would be able to make quantitative comparisons for weight and volume using the measure line. We also thought they would be able to describe how much heavier one material is than another and how much greater volume one object has than another. We found it was challenging for third graders to make these quantitative comparisons; therefore, we had to slow the pace of the curriculum to allow time for discussions of these issues. Furthermore, consistent with Lehrer, Jaslow, and Curtis's (2003) research, we found that the students began with little metaconceptual understanding of measurement. They lacked even a basic understanding of the role of units in length measurement. If the students had had mathematics instruction that developed stronger understandings of length and area measures, we would have been in a better position to develop their understanding of the measurement of volume.

Our hope is that standards and the structure of schools will begin to shift based on the field's collective insights from learning progressions research. However, we cannot count on this. Instead, we must find a way to reconcile two realities: the potential of the LPs and the reality of today's schools. Since school structure does not change easily, if LPs, as well LPM, are to be effective, at least in the short-term, they must work within that structure.

CONCLUSIONS

As they are inherently constructivist, LP approaches to curriculum design assume not only that students' initial ideas are meaningful but also that they are the only basis on which to build further understanding. Learning progression approaches are based on research that relates to some aspects of earlier research on misconceptions (e.g., Driver, Guesne, & Tiberghien, 1985). The misconception research brought to science educators' attention the shallowness of most students' scientific knowledge and the deep incompatibilities between students' and scientists' interpretations of the same phenomena. But that research mostly focused on older students and the ideas they developed in formal science instruction. Such research tended to be diagnostic; when it offered alternative teaching methods, it was often for a specific topic, for a specific grade, and with the goal of "undoing" particular misconceptions. The goal of a LP approach to science instruction, in contrast, is to avoid many of those student misconceptions by starting to teach science early in school (ideally in kindergarten) and by developing curricula that are coherent within and across grades. Learning progression approaches acknowledge that young children's ideas about the physical world are very different from scientists' ideas. However, these approaches also emphasize *transforming* students' ideas to bring them progressively into closer alignment with scientists' ideas rather than replacing *mis*conceptions with scientifically correct ideas.

Learning progression approaches also assume that the knowledge networks of both students and scientists in a domain are complex and can be characterized using similar units of analysis (e.g., concepts, beliefs, models, and epistemological and ontological assumptions). LPs focus on interrelationships among those units and the constraints those interrelationships place on the mutual development of curricular units. Of course, students' knowledge networks differ in content from scientists' and support different kinds of explanations. Hence students should engage in reconceptualization to bring their knowledge networks into alignment with those of scientists'—relationships among elements of the network change as new elements are introduced and coordinated with existing ones and with each other. This reconceptualization occurs in conjunction with new epistemological and ontological beliefs and with the support of new symbolic representations.

Such reconceptualizations take time. They require the support of curricula with a long time span (e.g., K-12) that give ample time for revisiting core concepts across grades in new, broader, and more challenging contexts in relation to more sophisticated epistemological knowledge and in interaction with new symbolic tools. Concepts are enriched, differentiated, and recruited to account for different phenomena; the relationships among and between them change; new concepts emerge; and explanatory accounts shift. Learning progression approaches highlight the many small, coordinated steps needed to bridge children's initial understandings in a domain to scientific accounts. In doing so, they attempt to demystify the process of reconceptualization as well as make it translatable into curricula.

Designing LP-based curricula that promote reconceptualization is a challenging process that requires making choices. In this chapter, we addressed the following three sub-questions:

(a) Given that one cannot bridge the lower and upper anchors (i.e., young children's and science knowledge) in one step, what are reasonably meaningful and coherent "intermediate" knowledge states? What content and structure can students' knowledge network have at different points along the learning progression that can form the end goal for learning at a particular grade?
(b) Given that one organizes curricula around deepening students' understanding of (inter-related) core concepts, where should one begin? Which concepts (and relations) should one address first to ensure meaningful learning and leverage future change?
(c) Given that reconceptualizations are complex and long-drawn, how can we design and sequence instructional activities within each grade in ways that promote the progressive integration and generalization of new pieces of knowledge?

We discussed how grappling with these questions led us to define new theoretical constructs (core concepts, lever concepts, stepping stones, and key representations such as measure lines and other models as tools for restructuring) that may have lasting value for LPs at different grade spans and in different domains.

In addition, we outlined the challenges in coordinating research and curriculum development (e.g., marrying the sometimes conflicting goals and time scales of research and development, deciding if the curriculum or the LP should be revised, etc.). We also described the challenges of adapting LP-based curricula to current classroom reality (e.g., limited classroom time, teacher support, and students' mathematics knowledge) while still promoting change.

In concluding, we mention another challenge. Why is it often so difficult for curriculum developers to (fully) recognize the extent of the reconceptualizations involved in science learning? Indeed, in our work, we often found ourselves *underestimating* children's difficulties and *overestimating* how meaningful particular activities would be for them. We think this challenge reflects a deep problem: namely, as human beings, we see the world (automatically) in terms of our own frameworks (rather than others' frameworks) and use ours to solve problems. Thus when we, as curriculum developers, think of ways to introduce children to a new concept meaningfully (e.g., volume), we (naturally) evaluate strategies based on how meaningful they are to us. "Cubing" is judged a meaningful strategy because it is a basic, transparent way to quantify volume. The problem is that it is only transparent to someone who already has a concept of volume..

Similarly, when we hear children say that a tiny piece of a material weighs nothing, it is natural to gloss their expression and assume that what they really mean is that the piece of material is light or it weighs very little. We do not consider that their whole network of ideas (i.e., their understanding of weight, nothing, heavy/light, material) may be different from ours. Or when we hear children say, "That's heavy because it's big," we may assume they think that weight and volume are directly proportional. We may not recognize that "heavy" and "big" may have very different meanings for them than for us. We understand that "weight" can have an entirely different meaning only when we consider many other things children say—for instance, that small things do not weigh anything because "they are too tiny to weigh," that two objects of different size made of the same material can have the same weight, that there are only one or two weights between the weight of a large and small ball, and so on. One advantage of systematic research on children's understandings is that it allows us to more fully reconstruct their knowledge network and thus to imagine how they see the world. This research will always be a difficult and "mindful" process, but unless the researchers engage in this process, they will constantly overlook key assumptions that need to be discussed with children in the classroom and overestimate how much children learn in activities and lessons. Designing curricula that are both meaningful for students and productive is challenging indeed.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under Grant No. DRK-12 0628245. Susan Doubler (TERC) was the lead PI of the project. Roger Tobin (Tufts University), David Carraher (TERC) and Jodi Abell-Clarke (TERC) were co-PIs. Any opinions, findings, conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This chapter presents the work of a team of a large number of researchers, curriculum designers, teachers, graduate students and research assistants (listed at http://inquiryproject.terc.edu/about/team.cfm). We would like to give particular thanks to the teachers and students at the Forestdale School and the Mason School for their participation.

NOTES

1 "Incommensurability" refers to the lack of communication between people who hold different theories because the same terms (e.g., "matter") have different meanings and different referents (Kuhn, 1962). However, terms that preserve their meaning across a theory change can provide a basis for coherent discussion of differences in the theories and can help make sense of incommensurate terms.

2 This scenario is partially fictional. It does not describe the Inquiry Project Curriculum in specific detail. However, this account conveys the general strategy and sequence in the Inquiry Project Curriculum across Grades 3-5, establishes an understanding that solids and liquid materials have weight and take up space in Grades 3-4, and uses weight as indirect evidence for invisible (gaseous) materials in Grade 5. A complete description of the Grades 3-5 Inquiry Project Curriculum is available at the Inquiry Project website: http://Inquiryproject.terc.edu

3 This is true only in a gravitational field. However, the more general statement—all matter has mass—is meaningless to students in this grade range, given that the concept of mass is a difficult and late achievement. We argue in the Stepping Stones section below that the belief that all matter has weight is actually essential to developing a scientific concept of matter and to constructing a concept of mass later on.

4 This is not the same idea as weight is proportional to mass although it is a precursor of the idea. Fourth graders learn that the weight of various quantities of a particular solid, aggregate, or liquid is proportional to those quantities (as measured by their volumes).

5 A "two-bottle system" consists of two plastic liter bottles sealed together at their openings and tipped at an angle with a puddle of water in the lower bottle. A heat lamp warms the system, and the water moves from the lower bottle to the upper bottle.

6 Fully reconceptualizing matter as made of particles and then atoms and molecules will take several years.

7 This is, of course, not entirely correct (e.g., shadows serve as a counter-example). However, this is a case where students' current beliefs can be used even if they are not entirely correct. We return to this issue in the Stepping Stones section of the chapter.

8 If we introduced the LPM in kindergarten or pre-kindergarten rather than in Grade 3, some aspects of the Grades 3-4 stepping stone would be incorporated in a K-2 stepping stone, leaving more time for full development of our current Grade 5 stepping stone in Grades 3-5.

9 We have argued elsewhere (Wiser & Smith, 2008) that chemical transformations and the notion of pure substance should not be introduced before AMT.

10 Density cubes are a commercially available sets of equal volume cubes made from different materials of different densities (e.g., pine, oak, nylon, PVC, acrylic, aluminum, copper, steel). Consequently, each cube has a different weight.

[11] Lehrer, Jaslow, and Curtiss (2003) beautifully document how the concept of volume can be constructed using length and area as levers. They involve students in thought experiments in which they "pull an area through a length". However, our students' knowledge of length and area was insufficient to act as levers. Time constraints did not allow us to work on those concepts first.

[12] See DeBarger, Ayala, Minstrell, Krauss, and Stanford (2009) for further discussion of this issue.

[13] IQWST (Investigating and Questioning our World through Science and Technology) is an innovative 3-year middle science curriculum that promotes understandings of big ideas in science.

REFERENCES

Acher, A., & Arcà, M. (2006). Children's representations in modeling scientific knowledge construction. In C. Andersen, M. N. Scheuer, M. P. Perez Echeveerria, & E. Teubal (Eds.), *Representational systems and practices as learning tools in different fields of knowledge*. Rotterdam, The Netherlands: Sense Publishers.

Au, T. K-F. (1994). Developing an intuitive understanding of substance kinds. *Cognitive Psychology*, *27,* 71-111.

Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development*. London, UK: Blackwell Publishers.

Carey, S. (2008). *The* origin *of concepts*. Oxford, UK: Oxford University Press.

Carraher, D., Schliemann, A., Brizuela, B., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics Education*, *37*, 87-115.

Carraher, D., Smith, C., Wiser, M., Schliemann, A., & Cayton-Hodges, G. (2009, June). *Assessing students' evolving understandings about matter*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

DeBarger, A., Ayala, C., Minstrell, J., Krauss, P., & Stanford, T. (2009, April). *Facet-based progressions of student understanding of chemistry*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Dickinson, D. K. (1987). The development of material kind. *Science Education*, *71*, 615-628.

Driver, R., Guesne, E., & Tiberghien, A. (1985). *Children's ideas in science*. Milton Keynes, UK: Open University Press.

Fauconnier, G., & Turner, M. (2003). *The way we think: Conceptual blending and the mind's hidden complexities*. New York, NY: Basic Books.

Hart, B., & Risley, T. R. (1999). *The social world of children: Learning to talk*. Baltimore, MD: P. Brooks Publishing.

Imai, M.,& Mazuka, R. (2007). Revisiting language universals and linguistic relativity: Language relative construal of individuation constrained byuniversal ontology. *Cognitive Science, 31*, 385-413.

Johnson, P. (1998). Progression in children's understanding of a "basic" particle theory: A longitudinal study. *International Journal of Science Education*, *20*, 393-412.

Johnson, P., & Papageorgiou, G. (2010). Rethinking the introduction of particle theory: A substance-based framework. *Journal of Research in Science Teaching*, *47*, 130-150.

Kuhn, T.S. (1962). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Lehrer, R., Jaslow, L., & Curtis, C. (2003). Developing understanding of measurement in elementary grades. In D. Clements & G. Bright (Eds.), *The National Council of Teachers of Mathematics: Yearbook on learning and teaching measurement* (pp. 100-121). Reston, VA: The National Council of Teachers of Mathematics.

Lehrer, R., & Pritchard, C. (2002). Symbolizing space into being. In K. Gravemeijer, R. Lehrer, B. van Oers, & L. Verschaffel (Eds.), *Symbolization, modeling and tool use in mathematics education* (pp. 59-86). Dordrecht, Netherlands: Kluwer Academic Press.

Lehrer, R., & Schauble, L. (2000). Developing model-based reasoning in mathematics and science. *Journal of Applied Developmental Psychology*, *21*, 39-48.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 39-74). Mahwah, NJ: Erlbaum.

Lucy, J. A., & Gaskins, S. (2003). Interaction of language type and referent type in the development of nonverbal classification preferences. In D. Gentner & S. Goldin-Meadow, (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 465-492). Cambridge, MA: MIT Press.

Margel, H., Eylon, B.-S., & Scherz, Z. (2008). A longitudinal study of junior high school students' conceptions of the structure of materials. *Journal of Research in Science Teaching*, *45*, 132-152.

Merritt, J., & Krajcik, J. (2009, June). *Developing a calibrated progress variable for the particle nature of matter.* Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Metz, K. E. (1993). Preschoolers' developing knowledge of the pan balance: From new representation to transformed problem solving. *Cognition and Instruction*, *11*, 31-93.

Nussbaum, J. (1997). History and philosophy of science and the preparation for constructivist teaching: The case of particle theory. In J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Teaching science for understanding* (pp. 165-194). Boston, MA: Academic Press.

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities: Conservation and atomism*. London, UK: Routledge and Kegan Paul.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* New York, NY:Penguin Group.

Schliemann, A., & Carraher, D. (2002). The evolution of mathematical understanding: Everyday vs. idealized reasoning. *Developmental Review*, *22*, 242-266.

Shwartz, Y., Weizman, A., Fortus, D., Krajcik, J., & Reiser, B. (2008). The IQWST experience: Using coherence as a design principle for a middle school science curriculum. *The Elementary School Journal*, *109*, 199-219.

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight and density. *Cognition*, *21*, 177-237.

Smith, C., Grosslight, L., Davis, H., Unger, C., & Snir, J. (1994). *Using conceptual models to teach inner city students about density: The promise and the prerequisites*. Final report to the McDonnell Foundation.

Smith, C., Maclin, D., Grosslight, L., & Davis, H. (1997). Teaching for understanding: A comparison of two approaches to teaching students about matter and density. *Cognition and Instruction*, *15,* 317-393.

Smith, C., Snir, J., & Grosslight, L. (1992). Using conceptual models to facilitate conceptual change: The case of weight/density differentiation. *Cognition and Instruction, 9,* 221-283.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*, 1-98.

Smith, C., Wiser, M., & Carraher, D. (2010, March). *Using a comparative, longitudinal study with upper elementary school students to test some assumptions of a* learning *progression for matter.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Snir, J., Smith, C. L., & Raz, G. (2003). Linking phenomena with competing underlying models: A software tool for introducing students to the particulate model of matter. *Science Education*, *87*, 794-830.

Spelke, E.S. (2000). Core knowledge*. American Psychologist, 55*, 1233–1243.

Stavy, R., & Stachel, D. (1985). Children's conceptions of changes in the state of matter: From solid to liquid. *Archives de Psychologie*, *53*, 331-344.

Stevens, S. Y., Delgado, C., & Krajcik, J.S. (2010). Developing a hypothetical multidimensional learning progression for the nature of matter. *Journal of Research in Science Teaching, 47*, 687-715.

Uzgiris, I. (1964). Situational generality of conservation. *Child Development*, *35*, 831-841.

Vergnaud, G. (1996). The theory of conceptual fields. In L. P. Steffe, P. Nesher, P. Cobb, G. Goldin, & B. Greer (Eds.), *Theories of mathematical learning* (pp. 219-239). Hillsdale, NJ: Erlbaum.

Wiser, M., Citrin, K., Drosos, A., & Hosek, S. (2008, June). *Young children's understanding of the relation between weight and material kind.* Poster presented at the annual Meeting of the Jean Piaget Society, Quebec City, Canada.

Wiser, M., & Fox, V. (2010, November). *Quantifying amount of material: A teaching intervention in pre-K and kindergarten.* Paper presented at the International Science Education Symposium on the Particulate and Structural Concepts of Matter, University of Athens, Greece.

Wiser, M., & Smith, C. L. (2008). Teaching about matter in grades K-8: When should the atomic-molecular theory be introduced? In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 205-239). New York, NY: Routledge.

Wiser, M., & Smith, C. L. (2009, August). *How does cognitive development inform the choice of core ideas in the physical sciences?* Commissioned paper presented at the National Research Council Conference on Core Ideas in Science, Washington, DC. Retrieved from the National Academies website: http://www7.nationalacademies.org/bose/Wiser_Smith_CommissionedPaper.pdf

Wiser, M., Smith, C., Asbell-Clarke, J., & Doubler, S. (2009, April). Developing *and refining a learning progression for matter: The inquiry project* (grades 3-5). In C. Smith (Chair), Developing and refining a learning progression for matter from pre-K to grade 12: Commonalities and contrasts among four current projects. Symposium conducted at the annual meeting of American Educational Research Association, San Diego, CA.

ERIN MARIE FURTAK, JESSICA THOMPSON, MELISSA BRAATEN AND MARK WINDSCHITL

LEARNING PROGRESSIONS TO SUPPORT AMBITIOUS TEACHING PRACTICES

One challenge faced by teachers, especially novice teachers, is navigating the messy and confusing landscape of science teaching reforms. In reform-based classrooms, students may be moving around and talking as they share ideas. Part of developing expertise as a teacher is learning which aspects of the classroom environment can be ignored and which ones can be pursued to fruitful ends. Teachers must learn to separate the signal from the noise, so to speak, during the act of teaching. Goodwin (1994) identified this ability as professional vision; namely, the ability to survey a complex landscape, identify important elements in that landscape, and connect those elements to a larger framework of understanding that is shared by a profession. The field of science education is only beginning to develop effective supports for helping teachers develop professional vision (McDonald, 2008).

In this chapter we describe ways in which teachers' professional visions of the content and ambitious teaching practices promoted by the *National Science Education Standards* (National Research Council [NRC], 1996) can be built using tools called learning progressions (LPs). LPs as currently designed use multiple approaches to identify developmental sequences of performances that map the terrain from naïve ideas and misunderstandings to scientifically accepted explanations and practices (NRC, 2007).

LPs are defined as "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (NRC, 2007, p. 219). LPs have great potential as teacher preparation and professional development tools since they contain knowledge of student ideas and how students learn, and – in some cases – suggest strategies or actions to help students learn. In order to achieve that potential, however, we must attend to how teachers develop as practitioners, not just how students learn.

While LPs for students help map the terrain between students' prior knowledge and experiences and the scientific practices and understandings students should develop through instruction, they neither describe how teachers can meet these student needs nor illustrate how teachers develop along a trajectory of expertise in their own practices. Thus LPs designed to scaffold improvement in teaching practice must include a system of tools that supplements the LP. In our experience, teachers know that they should engage their students by using ambitious science

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 405–433.

teaching practices, but they do not always know what this kind of teaching looks or sounds like, nor how to enact it in their classrooms. To address this need, we have developed LP tools that support teachers in developing more sophisticated pedagogical performance. We have then used those tools with professional learning communities to structure the process of teacher development.

Our assertions for the development of LPs and a system of tools and routines are based on the following logic. As most practitioners acquire experience, they often use little more than informal observations of students to assess their own instructional efficacy and competence. They depend upon a kind of untested folk wisdom to make changes in practice (Goodlad, 1990; Huberman, 1995; Little & Horn, 2007; Lortie, 1975; McGlaughlin & Talbert, 2001) that is characterized as "bricolage" or tinkering. In this view, teachers develop as independent artisans, picking up a new activity here and a new technique there, choosing these to fit their individual teaching styles and work settings. In light of this, it is important to the field that expert performances in science teaching be systematically described in ways that are useful to educators. Currently, "what counts" as advanced pedagogical practice is underspecified. Reform-based teaching practices are often abstract and, in many cases, have been defined by what teachers should not do rather than constructively suggesting what they should do. To date, few resources in science education have been developed that help teachers recognize where elements of their practice are on such a continuum and what their next level of performance might be.

As we explored how LPs might support teachers' development, we encountered various challenges. The first challenge, creating a format for a LP that has instructional utility, involves presenting the content of the LP in ways that will be useful to teachers as they develop new teaching practices. The second challenge, developing a vision tool for all teachers, involves creating tools that have utility for teachers from different backgrounds and with different experiences. The third challenge, developing multiple tools to fill vision-to-practice gaps, involves creating tools that scaffold teachers' development of ambitious teaching practices.

This chapter presents the cases of two ongoing research projects in which these challenges were salient. In these cases, we highlight the challenges associated with developing and using LPs created for the purpose of scaffolding teachers' adoption of ambitious teaching practices. The first project explores a LP representing teaching practices designed to support model-based inquiry; the second project explores how a LP representing student understanding can support teachers' adoption of everyday assessment practices. Both cases involved the creation of a LP and an accompanying suite of tools that could be used to interact with and influence teacher development. We had the same hypothesis for both cases. The hypothesis is that, from the sociocultural standpoint, a LP for teachers, with accompanying tools, can form a common language useful in professional learning communities for collaboratively critiquing and improving practice. The chapter concludes with suggestions for other researchers considering the design of LPs for teacher development.

CASE 1: A MODEL-BASED INQUIRY LEARNING PROGRESSION FOR TEACHERS

In this study, we are engaged in ongoing empirical and theoretical work on the development of a LP that describes increasingly sophisticated ways that teachers plan, enact, and assess various components of reform-based teaching—specifically how they support students' Model-Based Inquiry (MBI; Windschitl, Thompson, & Braaten, 2008). In the past four years we have conducted longitudinal case studies of 11 secondary science teachers, tracking their teaching performances and using these data to design a LP for early career educators. Our data sources include written materials from their methods coursework, 81 classroom observations during student teaching and first-year teaching, videos of teaching, and artifacts from four rounds of analysis of student work and subsequent Critical Friends Group meetings (Windschitl, Thompson & Braaten, 2011a). We have additional data from several participants that extends into their second year of teaching, as well as data from two additional cohorts we are just beginning to track. The teachers' struggles and successes in taking up ambitious practices informed our design for a beginning teacher's repertoire of practices and also provided us with a system of tools and socio-professional routines that could foster such teaching over time.

The MBI Learning Progression

Our LP encompasses a continuum of pedagogical sophistication along 11 criteria of reform-based teaching that supports MBI. The design of our LP is informed by three areas of scholarship: (1) the nature of authentic disciplinary practices in science, including areas of epistemology, science studies, and sociocultural perspectives on the development of scientific knowledge;[1] (2) the way in which students learn science, including areas of meta-cognition, conceptual change, and model-based reasoning;[2] and (3) novice teacher development, including novice-expert and teacher learning literature.[3]

This LP tool includes criteria such as "planning and designing lessons with attention to students' engagement with models in an inquiry context" and "identifying full scientific explanations of phenomena and students' approximations of these explanations" (see Figure 1). Some criteria are multifaceted and include sub-criteria similar to those that guide LP-development for student performances in various content domains (Corcoran, Mosher, & Rogat, 2009; Smith, Wiser, Anderson, & Krajcik, 2006). However, the LP was not based on teacher knowledge but rather on teacher performances in the broad areas of planning, enacting, assessing, and reflecting on instruction. The lower anchors of pedagogical performance are based on empirical analyses of how novice secondary teachers typically engage students in inquiry (Crawford & Cullin, 2004; Windschitl & Thompson, 2006; Windschitl et al., 2008). The upper anchors (advanced levels of supporting MBI) are based on empirical research and in some cases hypothesized practices of how expert teachers foster more sophisticated scientific practice and discourse in the classroom around developing models, applying evidence, and

MBI Learning Progression

	Criteria	Increasingly Sophisticated Facilitation of Model Based Inquiry→		
Pre-planning for MBI: Identifying Models and Explanations	**1.** Identifying inquiry-worthy ideas (big ideas in science, potential data sources, and interest to students)	**1.1.1.** T. has a topic oriented approach to inquiry. A "thing" rather than a process or theory is object of study (i.e., cells). **1.2.1.** T. considers only first-hand experimental studies to generate data and evidence for inquiry. **1.3.1.** Relevance to students is not a lens T. uses when assessing the worthiness of study OR relevance not incorporated into the curriculum.	**1.1.2.** T. selects processes (i.e., osmosis) to be the focus for an inquiry. **1.2.2.** T. recognizes various types of inquiries and data but does not use these types in practice nor teaches Ss. how each explicitly relates to model development. **1.3.2.** T. uses relevance as initial hook to interest students in topic of inquiry.	**1.1.3.** T. selects a theory as focus for inquiry (i.e., evolution, cell theory, environmental interrelationships). **1.2.3.** T. recognizes secondary data as sources of evidence (using maps of earthquakes, fault lines) or combines both 1st hand data collection/ analysis and secondary data analysis to enrich inquiry (i.e., examining data from a stream collected by a local organization and building a stream table in the classroom). Selects comparative or correlative studies (with primary or secondary sources) OR uses a descriptive study as a pre-cursor to a comparative or correlative study. Plans to make explicit to students various types of inquiry and variations in the way each relates to models. **1.3.3.** T. elicits Ss. ideas to determine how science ideas could be relevant to Ss.—personally, locally and/or culturally. Relevance interwoven throughout inquiry.
	2. Identifying how an inquiry can be about the big idea/model(s)	**2.1.** Inquiry is about a thing (i.e. batteries & bulbs) without being about a big idea (i.e., energy) or a model (i.e., diagram of energy transfers).	**2.2.** Inquiry is about unproblematic science processes (not as a set of models) or about methods used to do the inquiry. T. cannot re-envision existing inquiry or curriculum in terms of models.	**2.3.** Inquiry is about a big idea as a model or set of models. T. identifies "this is model of X" in terms of big ideas (i.e., sees difference between model of battery vs. model of energy, or, model of a pond vs. model of nutrient cycling). T. envisions how big ideas can be represented as models and examines curriculum to identify which models help students understand scientific processes. T. plans to begin unit by introducing a phenomenon on which instruction, models, and inquiry can be based.
	3. Assessing the nature and function of models being used	**3.1.1.** T. selects a model that is a pictorial or physical replication of a "thing" considered to be real. **3.2.1.** T. uses models to simplify, illustrate, or show something.	**3.1.2.** T. selects models that portray observable processes and systems or a mathematical representation or set of rules. T. portrays models as having multiple representations – there is no singularly "correct" model. **3.2.2.** T. uses model to facilitate understanding, help Ss understand what an expert knows, and/or to separate out effects and variables of a complicated phenomenon.	**3.1.3.** T. selects model that portrays theoretical processes and relationships; the model represents ideas rather than "things." T. portrays model as fallible because of the creative nature, logical limits, and underlying assumptions. T. considers the model (and/or components of the model) as investigate-able, revisable. **3.2.3.** T. uses model for two kinds of working value: a) It is generalizable and can serve to make predictions. b) It facilitates new insights into relationships, generates novel hypotheses and questions for inquiry.
	4. Planning and designing lessons with attention to Ss' engagement with models in an inquiry context	**4.1.1.** T. elicits Ss' ideas about a topic, which are left as a brainstormed list of things, not like thoughtful, tentative explanations of scientific processes. **4.2.1.** T. does not intentionally plan to test ideas as models (i.e., T gives Ss Earth/Sun models to test the influence of the angle of the earth on the seasons). **4.3.1.** T. uses inquiries over a unit/ academic year, which tend to be confirmatory. **4.4.1.** T. plans to use scientific forms of talk/reasoning but does not provide scaffolds for Ss to do the same	**4.1.2.** T. elicits Ss' ideas about a scientific process but not about science ideas as models. **4.2.2.** T. plans to instruct use only existing models of value to a scientific domain OR focuses only on student-constructed models without engaging Ss in models relevant to the discipline. **4.3.2.** T. does not provide enough guidance such that the focus of the inquiry is on material activity rather than targeted learning about a topic. OR T. does not gradually turn over more responsibility to Ss over the course of an academic year. **4.4.2.** T. plans to teach everyday ways of talking about science but do not scaffold transitions between these and scientific ways of talking and reasoning	**4.1.3.** T. elicits all Ss' ideas about scientific models and/or elicits ideas to create a Ss-generated initial model. T. adapts instruction based on these ideas and chooses to emphasize some initial ideas as leverage points for advancing content within the inquiry. **4.2.3.** T. plans to use both existing scientific and student-constructed models together, merging the two. There is a balance between teacher- or domain-generated and student-generated models (i.e., T. plans to have Ss first hypothesize about interaction in a pond system, then incorporate existing models of nutrient cycles, then perform inquiry based on class model; T. plans to have Ss investigate 3 models of the causes of the seasons and weigh evidence for each rather than just testing one parameter like angle of earth). **4.3.3.** T. makes decisions about level of guidance students need based on Ss' background experiences and sequencing during the year. **4.4.3.** T. assesses scientific language demands in the lesson and plans classroom conversations about the nature of models, model-based explanations and how these forms of scientific talk/reasoning differ from conversational/everyday talk
Guiding Classroom Discourse: Pressing for Explanation	**5.** Identifying full scientific explanations and Ss' approximations of these explanations	**5.1.1.** In planning, T. does not identify full explanation(s) for phenomena being investigated. T. only plans for Ss to discuss what happened during the inquiry. **5.2.1.** T. does not access Ss' initial explanations in ways that inform upcoming instruction.	**5.1.2.** T. identifies explanations for scientific phenomena ahead of time but unobservable/theoretical components are tangential. T. only plans for and has Ss discuss how something happened during the inquiry. **5.2.2.** T. accesses Ss prior knowledge, common pre-conceptions and naïve ways of understanding phenomena but these are not used to adapt instruction.	**5.1.3** T. envisions possible models and explanations for why something happened. T. orients inquiry toward particular scientific models so that guided instruction about key concepts ensured. T. imagines levels of sophistication of possible explanations and guides Ss discourse toward more coherent, evidence-based explanations. T. constantly assesses student explanations before, during, and after inquiry for students of all achievement levels. **5.2.3.** T. elicits Ss' ideas about tentative hypotheses and adapts instruction to explicitly unpack Ss' preconceptions and alternate ideas for students of all achievement levels. T. anticipates components of a model that may be conceptually challenging for students.

	6. Setting up and gathering data: Aiming discourse toward a "why" explanation	**6.1.1.** T. set up for inquiry and data collection focuses on directly observable not underlying causal components. **6.2.1.** While circulating T. presses for ideas about data collection or experimental design and for "what" explanations without asking for "why".	**6.1.2.** T. set up for inquiry and data collection is purposeful but not enough focus placed on rich explanatory models. After an inquiry, models and explanations are used as a culminating or finished product that are revealed to Ss or as an outcome of inquiry. After an inquiry, T. builds in background knowledge of key science ideas or facts.. **6.2.2.** While circulating T. presses for descriptions of how data was collected or how a phenomenon occurred.	**6.1.3.** T. set up for inquiry and data collection is purposeful and highlights tentative explanatory models as the basis for investigation and data collection. T. uses model as a touching point before, during and after an inquiry. T. strategically builds in background knowledge of key science ideas and models before, during and following an inquiry. Just in time instruction is used for fine-tuning Ss understanding of complex models. **6.2.3.** While circulating, T. poses tasks and questions that press Ss to explain patterns in data and to explain phenomena. T. can identify gaps/contradictions in student thinking related to an explanatory model (ex. noticing that Ss are only understanding temperature and volume, not pressure, when investigating the ideal gas laws).
	7. Ongoing press for coordinating evidence with explanation	**7.1.1.** Focus on experimental design, measurement, data collection. Data are not discussed as evidence because focus is not on interpreting data in terms of models and explanations **7.2.1.** T. asks Ss to describe what happened; the press might be for a describing, summarizing, or restating a pattern or trend in data without making a connection to any unobservable/ theoretical components. **7.3.1.** T. does not address uncertainty, counterarguments, error.	**7.1.2.** T. focus is on evaluating data and on explanation but not simultaneously. Unobservable/ theoretical components are only addressed tangentially. **7.2.2.** T. helps Ss construct a logical argument including a claim, evidence, and some reasoning. **7.3.2.** T. tangentially addresses uncertainty or error. OR T. places too much focus on error such that the press for explanation is lost in a press for describing flaws in the experimental design.	**7.1.3.** T's focus is on coordinating evidence with explanations. T. can assess students' discourse in the moment and press for why explanations (move students from a what→why or how→why explanation) **7.2.3.** T. helps Ss explain why something happened by writing claims that justify the link between observable data and unobservable/ theoretical components. **7.2.3.** T. helps Ss construct an argument that takes into account counterarguments, alternative hypotheses, variance and error.
Guiding Classroom Discourse: Evaluating and Revising Explanatory Models	**8.** Revisiting or revising explanatory models based on evidence	**8.1.** No coordination of evidence and models. T. may revisit the hypothesis but this is only done for the purpose of confirming or disconfirming the hypothesis.	**8.2.** T. assists Ss in revisiting explanatory models but the focus is not on enriching or revising the model, simply reviewing the connection between evidence and the part of the model that was investigated.	**8.3.** T. assists Ss in enriching and revising models by pulling in background information that is complementary to evidence collected and targeted scientific theories. T assists Ss through multiple iterations of revisiting/revising models and explanations. T. places sense-making conversations about patterns in the data and explanations after the first round of data collection such that inquiries can be revised to be more rigorous in terms of method and theory. T. also helps Ss see how their evidence + explanations evolved and how they make sense with more sophisticated models or with multiple models. T. lead Ss in conversations about the tentative nature of data and Ss discuss one another's data in terms of models. T. also lead Ss in conversations in which students compare and contrast models and explanations
	9. Application of model/extension of model	**9.1.** T. ask Ss to write a new question based on the inquiry but does not press for the question to focus on an explanation. T. does not differentiate between questions about experimental design (such as changing one variable) and questions aimed at the underlying mechanisms.	**9.2.** T. asks Ss to write a question or make predictions without asking students to consider evidence from the inquiry as well as the explanations.	**9.3.** T. leads Ss in conversations about designing new questions and predictions that are informed by both evidence and model-based explanations following an inquiry. T. also leads conversations about explanations and their relative attributes/generalizability and limitations.
	10. Assessing students' understanding about models	**10.1.1.** T. uses scientific models with Ss but does not lead explicit discussions about the nature and function of models with students. **10.2.1.** T. generally assesses Ss use scientific language; evaluating Ss science vocabulary use	**10.1.2.** T. leads discussions about the nature and function of models but it is not in the context of an inquiry. **10.2.2.** T. assesses if Ss can use scientific language; evaluating the depth of Ss explanations or depth of descriptions of how and why models are used.	**10.1.3.** T. assists Ss in understanding science models as objects of critique and revision, and how models are used in science. T. helps Ss move from particular to more general models with a focus on understanding the general nature and function of models. **10.2.3.** T. assesses individual Ss scientific language development, tracking development of the construction of scientific explanations, use of representational language and ability to generalize and theorize.
Reflective Practice	**11.** Evaluating one's MBI pedagogy through examination of student work	**11.1.1.** T. examines Ss work for "right and wrong" answers, not nuanced understandings of the content. Does not differentiate between grading and examining student work for understanding. **11.2.1.** While examining Ss work, T. fails to see ways in which scientific models could have been a part of their investigation. **11.3.1.** T. sees examination of Ss work as unproblematic and not informative to their practice.	**11.1.2.** T. examines Ss work for nuanced understandings of the content but not in the context of models or theories. **11.2.2.** While examining Ss work, teachers have a limited view of which scientific models could have been a part of their investigation. **11.3.2.** T. sees examination of Ss work as informative to their practice. Take-aways are limited to adjustments in tasks, such as re-wording a prompt, rather than ways in which lessons can be re-worked to help Ss understand explanatory models.	**11.1.3.** T. examines individual Ss work for nuanced understandings of models or theories. **11.2.3.** T. asks questions of one's practice when examining Ss work (ex. What might be going on here that I was not aware of? Are there some subtle changes in the Ss work that I am not seeing? What more could I do to challenge students at all levels of ability? What was the nature of the instruction or questions on the assignments that might have influenced my Ss staying at the same levels? What could I do to better support under-achieving students?) **11.3.3.** T. follows through on what he/she would like to change about teaching.

Figure 1. Model-Based Inquiry learning progression of teaching performances.

constructing explanations (NRC, 2005; Schwarz & White, 2005; Windschitl & Thompson, 2006). To describe the middle anchors, we observed novice teachers in secondary science classrooms during their student teaching and their first year of teaching. We identified patterns of less and more sophisticated versions of each criterion on the MBI LP, noting when partial practices were adopted, when the language but not the practice was appropriated, and when teaching practices were stifled or pulled in new directions by the influence of school context. For example, for criteria 4.1 and 4.2, we added details about how teachers partially appropriated practices related to building scientific models—several teachers used students' initial ideas to construct a scientific model but did not successfully layer theoretical constructs onto these models. As a second example, for criteria 2.2, 3.2, and 9.1, we accounted for an unexpected contextual finding that influenced how conceptual ideas were examined in the classroom—many novice teachers felt pressured to teach to mandated assessments or from textbooks that emphasized the design of scientific experiments rather than the underlying explanation of a phenomenon. This contextual emphasis on experimental design influenced teachers' performances as they worked to engage students in MBI. More than just "fleshing out" the middle cells of the LP, we also added specific examples from our observations to the upper/ theoretical anchor based on instances when the teachers in the study provided new and creative visions of these practices.

Challenge 1: Creating a Format for a LP That Has Instructional Utility

As a part of this study, we wanted novice science teachers to use the LP to locate specific elements of their inquiry teaching practice along a continuum and then to envision "next steps" for advancing their practice. We also wanted to help them develop and share a common language with others using the LP. We quickly realized, however, that our complete LP might overwhelm novice teachers who are just learning to make sense of their classroom practice. In order to scaffold novice teacher learning, we developed a reduced version of our LP to bring teachers' attention to a core set of complex, ambitious, and equitable practices that have the most impact on student learning (see Figure 2). The mathematics education community refers to these teaching practices as high-leverage practices—those most likely to stimulate significant advancements in student thinking and teacher learning. Working from the literature on high-leverage practices (Ball, Sleep, Boerst, & Bass, 2009; Franke & Chan, 2007; Hatch and Grossman, 2009), we used the following criteria to decide which practices to include in the reduced set of ambitious practices (see also Windschitl, Thompson, & Braaten, 2011b):

- Help improve the learning and achievement of all students,
- Support student work that is central to the discipline of the subject matter,
- Are used frequently when teaching,
- Apply to different approaches in teaching the subject matter and to different topics in the subject matter,
- Are conceptually accessible to learners of teaching,

- Have facets that can be articulated and taught,
- Are able to be practiced by beginners in their university and field-based settings,
- Can be revisited in increasingly sophisticated and integrated acts of teaching, and
- Have features that readily allow novices to learn from their own teaching.

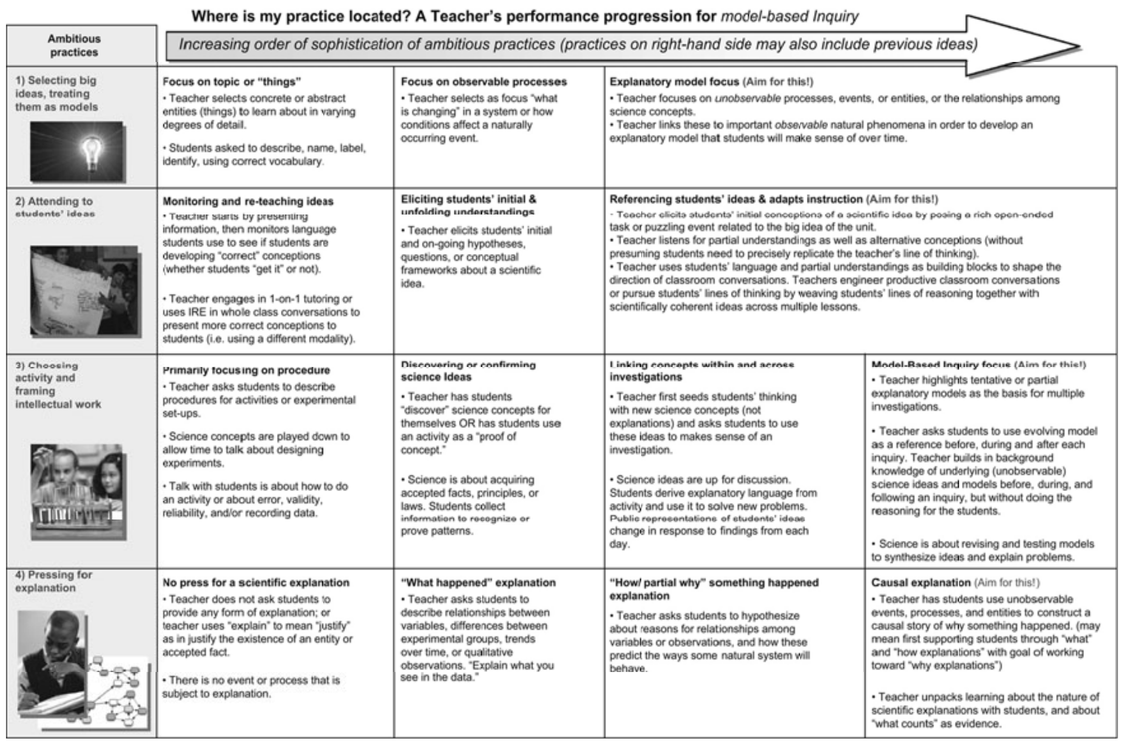
As we observed novice teachers attempting to use reform-based practices in their student teaching and their first year of teaching, we identified "clunky" or incomplete implementations as well as contextual pressures that influenced and often distracted them from full implementation of the most ambitious versions of these practices. In addition to the above criteria, we purposefully selected the practices that novice teachers had the most difficulty in implementing. We describe these practices and the novice teachers' associated struggles in the following paragraphs.

1) Selecting big scientific ideas and treating them as models. The first practice novice teachers struggled with was identifying the big ideas to teach. By "big ideas" we mean substantive relationships among concepts in the form of scientific models that help learners understand, explain, and predict a variety of important phenomena in the natural world. Standard curricula tend to focus on topics and sometimes processes but rarely on theoretical ideas. For example, one teacher was given a curriculum unit on the ocean's tides and taught the definition of high and low tides, how often they occur, and where they occur. Then the teacher had students investigate a correlation among tides and moon phases but not the overarching theories involving gravitational forces. Another teacher with the same topic emphasized the big idea behind this phenomenon and had students test models that described the gravitational relationships between the earth, moon, and sun (a theory focus). Only four teachers (of 11) were able to modify standard curricula to focus on a "big idea"—i.e., theory-focused—and were able to help students use scientific models to generate explanations and test ideas (see also Windschitl & Thompson, 2006). In 81 classroom observations, we found only 25 instances in which these new teachers adapted the central topics of the curriculum, and only eight instances in which they moved beyond superficial topics or broad themes to identify more substantive ideas. Most teachers adhered to their activity-centered curricula. Moreover, if the teachers did not modify topic- or process-focused curricula to be more theory-focused, they were unable to attempt other forms of ambitious practices.

2) Eliciting and attending to students' initial and unfolding ideas in adapting instruction. The second struggle for our beginning teachers was sustaining science discourse in the classroom. The teachers often knew how to initiate student conversations (with a puzzling question or demonstration that elicited students' experiences), but they rarely knew what to do with student responses. We found three variations of this practice ranging from least to most effective: (1) no eliciting of ideas, rather an ongoing monitoring, checking, and re-teaching for "correct" answers; (2) eliciting students' initial understandings but with no follow-up; and (3) eliciting initial ideas and then adapting instruction based on these ideas. Of the 81 classroom observations, we found 24 instances where beginning teachers used the most sophisticated version of this practice; however, only three of the 11 teachers did so consistently. Apprenticing teachers– in particular novice teachers– into this type of teaching is complicated by the "hard wired" routines of low-level questioning in classrooms and discourses of teacher control (Baldi et al., 2007; Banilower, Smith, Weiss, & Pasley, 2006; Horizon Research International, 2003), by the lack of clear models of more sophisticated practices, and by inexperienced educators' limited understanding of students' capacities to engage in challenging work (Elmore, 2005).

3) Framing science activities and intellectual work with a model-based inquiry focus. The third struggle for our beginning teachers was a direct result of the curricula provided to them by their local schools and by their lack of a clear vision of how to change these curricula. We found that standard science inquiries from our teachers' curricula tended to gloss over the testable, revisable, and conjectural nature of scientific knowledge (Windschitl & Thompson, 2006). Instead, curricula directed students only to confirm or "discover" a known scientific idea. Most teachers enacted these curricula without making major revisions, although some teachers took a step in the opposite direction, emphasizing experimental design and material procedure at the expense of scientific ideas. This "method" focus is inherent in most curricula and is also emphasized in Washington State's standardized tests (Windschitl et al., 2008). Two of the 11 teachers attempted to revise standard curricula to conduct an inquiry on the development of scientific ideas. They treated scientific ideas as testable and revisable (although without using models) and chose to emphasize content-rich background knowledge prior to an investigation. Students were then asked to consider how their understanding of the phenomenon evolved in light of the findings.

4) Pressing for causal scientific explanations. The final issue novice teachers struggled with also related to the curricula. Most curricula do not press students to explain underlying causes of events and processes. At best, most activities and laboratory investigations ask students to look for patterns and trends in data. With this as an ultimate focus, it is not surprising that most classroom conversations tended to be about "what happened" in an investigation. There were very few conversations about "why a phenomenon occurred." We recognized this challenge early in our research and therefore developed an Explanation Tool for teachers to use in examining students' written explanations. We made this tool a focus in our first year of induction when teachers participating in our study returned to the university to analyze and discuss students' evidence-based explanations (see Windschitl, Thompson, & Braaten, 2011a). The Explanation Tool helped teachers distinguish among levels of explanation and levels of integration of evidence with scientific explanations. Compared to other ambitious practices, the teachers showed the most gains over time in supporting students' use of evidence-based explanations. Six of the 11 teachers were able to consistently shift classroom talk toward how and why phenomena occur (31 instances of 81 observations). Based on classroom observations and interview data, we attribute these gains to the use of the Explanation Tool and its routines; however, even with the tool, five teachers made no attempts to press for explanations, favoring instead a focus on activity and procedures.



Where is my practice located? A Teacher's performance progression for *model-based Inquiry*

Ambitious practices	*Increasing order of sophistication of ambitious practices (practices on right-hand side may also include previous ideas)*			
1) Selecting big ideas, treating them as models	**Focus on topic or "things"** • Teacher selects concrete or abstract entities (things) to learn about in varying degrees of detail. • Students asked to describe, name, label, identify, using correct vocabulary.	**Focus on observable processes** • Teacher selects as focus "what is changing" in a system or how conditions affect a naturally occurring event.	**Explanatory model focus** (Aim for this!) • Teacher focuses on *unobservable* processes, events, or entities, or the relationships among science concepts. • Teacher links these to important *observable* natural phenomena in order to develop an explanatory model that students will make sense of over time.	
2) Attending to students' ideas	**Monitoring and re-teaching ideas** • Teacher starts by presenting information, then monitors language students use to see if students are developing "correct" conceptions (whether students "get it" or not). • Teacher engages in 1-on-1 tutoring or uses IRE in whole class conversations to present more correct conceptions to students (i.e. using a different modality).	**Eliciting students' initial & unfolding understandings** • Teacher elicits students' initial and on-going hypotheses, questions, or conceptual frameworks about a scientific idea.	**Referencing students' ideas & adapts instruction** (Aim for this!) - Teacher elicits students' initial conceptions of a scientific idea by posing a rich open-ended task or puzzling event related to the big idea of the unit. • Teacher listens for partial understandings as well as alternative conceptions (without presuming students need to precisely replicate the teacher's line of thinking). • Teacher uses students' language and partial understandings as building blocks to shape the direction of classroom conversations. Teachers engineer productive classroom conversations or pursue students' lines of thinking by weaving students' lines of reasoning together with scientifically coherent ideas across multiple lessons.	
3) Choosing activity and framing intellectual work	**Primarily focusing on procedure** • Teacher asks students to describe procedures for activities or experimental set-ups. • Science concepts are played down to allow time to talk about designing experiments. • Talk with students is about how to do an activity or about error, validity, reliability, and/or recording data.	**Discovering or confirming science Ideas** • Teacher has students "discover" science concepts for themselves OR has students use an activity as a "proof of concept." • Science is about acquiring accepted facts, principles, or laws. Students collect information to recognize or prove patterns.	**Linking concepts within and across investigations** • Teacher first seeds students' thinking with new science concepts (not explanations) and asks students to use these ideas to makes sense of an investigation. • Science ideas are up for discussion. Students derive explanatory language from activity and use it to solve new problems. Public representations of students' ideas change in response to findings from each day.	**Model-Based Inquiry focus** (Aim for this!) • Teacher highlights tentative or partial explanatory models as the basis for multiple investigations. • Teacher asks students to use evolving model as a reference before, during and after each inquiry. Teacher builds in background knowledge of underlying (unobservable) science ideas and models before, during, and following an inquiry, but without doing the reasoning for the students. • Science is about revising and testing models to synthesize ideas and explain problems.
4) Pressing for explanation	**No press for a scientific explanation** • Teacher does not ask students to provide any form of explanation; or teacher uses "explain" to mean "justify" as in justify the existence of an entity or accepted fact. • There is no event or process that is subject to explanation.	**"What happened" explanation** • Teacher asks students to describe relationships between variables, differences between experimental groups, trends over time, or qualitative observations. "Explain what you see in the data."	**"How/ partial why" something happened explanation** • Teacher asks students to hypothesize about reasons for relationships among variables or observations, and how these predict the ways some natural system will behave.	**Causal explanation** (Aim for this!) • Teacher has students use unobservable events, processes, and entities to construct a causal story of why something happened. (may mean first supporting students through "what" and "how explanations" with goal of working toward "why explanations") • Teacher unpacks learning about the nature of scientific explanations with students, and about "what counts" as evidence.

Figure 2. Reduced learning progression of four core ambitious practices for novice science teachers. Adapted from "Working toward a Stronger Conceptualization of Scientific Explanation for Science Education," by M. Braaten and M. Windschitl, 2011, *Science Education, 46*, p. 666. Copyright 2011 by John Wiley and Sons. Reproduced with permission of John Wiley and Sons via Copyright Clearance Center.

The development of a LP around these four high-leverage practices has two functions—one for teachers and one for teacher educators. First, if used by novice teachers, the LP could provide descriptions of slight but significant differences in practice that may orient them to a core set of high-leverage practices and outline the steps needed to move toward these practices. We hypothesize that in their second year of teaching, teachers may be able to use the complete LP in collaborative settings and, with feedback on practice, make more detailed assessments of their current practice and what the next level of practice looks like. Second, teacher educators could use classroom observational data mapped onto the LP to decide how best to develop a complementary set of tools that would support teachers as they move from one cell to the next cell on the LP. In the next two sections we describe the successes and other challenges that arose when teachers used the Reduced MBI LP to examine their practice. We also describe how we, as teacher educators, attempted to develop pedagogical tools to help teachers move forward on each criterion of the Reduced MBI LP.

Challenge 2: Developing a Vision Tool for All Teachers

We propose that teachers can: (1) locate elements of their own practice within the LP and (2) use the "next levels" on those continua to imagine what their practice could become. Before we discuss how teachers in our study used LPs to talk about advances they might make in their practice, we address the prerequisite capability that teachers need if they are to accurately identify their current location on a given continuum. We asked teachers to identify where their practice fell on the four criteria of the Reduced MBI LP. We found that, with very few exceptions, they could not only accurately self-identify where their teaching was along several criteria, but they could also describe instances from their teaching which deviated from a level of sophistication they normally "occupied." (This deviation to a lesser level of sophistication was often due to their lack of familiarity with the subject matter in a particular unit, or because they were required to teach a particular unit without much modification.) Teachers could re-tell stories of pedagogical shifts, describing when and how they adopted a new practice or when they started regressing related to earlier progress.

Moreover, many teachers used the LP to state how they would like to improve their practice during their second year of teaching; most teachers asked to keep copies of the Reduced MBI LP as a reminder of how they wanted to advance their practice. However, one of the struggles was that not all teachers talked about improving their practice merely as a result of identifying where they were on the Reduced MBI LP. Some were quite satisfied with where their practice fell on the various criteria and used the Reduced MBI LP to justify their current practice. The following analysis reveals how and why some novice teachers used the Reduced MBI LP as a vision tool. Since we developed the Reduced MBI LP in conjunction with classroom observations and interviews with novice teachers, we did not ask the teachers to use the Reduced MBI LP until the end of their first year of teaching. The findings described below are based on how teachers interacted with the tool as a vision guide, or lack thereof, during the last few months of their first year of teaching.

LP as a vision tool. Two-thirds of the novice teachers used the Reduced MBI LP in three productive ways: (1) to justify their location—not only their practice but also their identities as teachers—based on their teaching and learning core commitments; (2) to identify missing gaps in their practice that might prevent them from moving forward on the LP; and (3) to raise questions about student learning dilemmas they are currently addressing and have yet to resolve.

Rather than stating their curricular visions (Darling-Hammond & Bransford, 2005), the novice teachers used the language and structure of the Reduced MBI LP to express who they are as teachers and who they would like to become. They used the Reduced MBI LP to engage in identity work. They imagined work that was yet to be done and, in some cases, identified the steps needed to work toward that future vision. Sfard and Prusak (2005) have noted that it is in this space—in which individuals are able to work on closing the gap between current and future narratives about oneself—that learning occurs. For example, Rachel described being frustrated that her location on the LP regressed during her first year of teaching when compared to progress in her student teaching.

> Rachel: [referencing LP criterion # 2] I know that there is a couple times when I have waffled over here [pointing to lower anchor] and I have been unhappy about that. I think that is definitely my goal. But yeah, that is where my mind sits. I think planning wise, thinking wise, and capability wise I have the capability of kind of being mostly being here and working towards that [upper anchor], but reality wise for the past six months I gotta say I am here. I don't like being there.

Rachel placed herself and her "mind" rather than her practice on the scale by referring to where she was located and where she did not want to be located. Rachel placed a firm stake in the upper anchors. In this way, it is possible that the Reduced MBI LP offered a way for some teachers to state their pedagogical identities to themselves and to others. The language and structure of the LP supported these statements about teachers' pedagogical selves.

None of these teachers simply mentioned that they would like to move forward on the Reduced MBI LP. They talked instead about working toward new visions. Some teachers were able to express pieces of plans or at least starting places for enacting change. This means that the teachers not only accurately located themselves on the Reduced MBI LP but also assessed the gap between cells on the LP. For example, Barbara considered how she could move away from one of the lower anchors for criterion 3 that focuses primarily on scientific procedures. To fill the gap between this lower anchor and other cells on the Reduced MBI LP that asks students to reason with science ideas, Barbara considered a practice her department head uses.

> Barbara: [referencing LP criterion # 3] Reflecting on this year I have been trying to think about how I can get away from that [the scientific method] so much because I have heard a lot of feedback from the students saying like

> gosh, we do this every single time. The same kind of set up and routine and I have been trying to get away from that. I am looking at how our department head teaches and he asks a couple questions and then that leads into an experiment, students just sort of try to answer those questions instead of coming up with this whole process I am trying to think of how I can go about doing that, but it is still something I am trying to wrap my brain around.

While Barbara has not fully developed the vision, at least she has some specific ideas about how she might move forward. Her comments about her future vision suggest, at least in part, that she had been working on a vision before interacting with the Reduced MBI LP. It is possible that the Reduced MBI LP simply surfaced this move she was considering making, but it is also possible that interacting with the LP provided further justification for working toward this future vision. Moreover, it is clear from Barbara and other participants that developing a shared vision across contexts is critical to the uptake of ambitious practices (see Thompson, Windschitl, & Braaten, 2010). This activity requires mentoring support that extends beyond just working with teachers on an individual basis.

The teachers also used the Reduced MBI LP to discuss students' learning dilemmas associated with their forward movement on the LP. They took a critical stance toward the LP and problematized their own forward movement on the LP by considering how students learn. For example, Emily located herself on the upper anchors of the third criterion of the LP (Pressing for Explanation) but paused to consider how students learn scientific explanations – a consideration that was just beginning to be part of her professional vision.

> Emily: [referencing LP criterion # 4] I think I sometimes really struggle with what is causing something to happen and I can usually sort through that pretty okay, but where I really get stuck is now what should I expect from kids? What does a causal explanation look like for a 15–16 year old? It shouldn't look the same, but I don't know how it should look. Then you can take it the next step once you have figured out how should it look, okay now how do you help them get there? I haven't gotten close to that one.

Emily said she needed to understand how to calibrate learning for high school students. This seems like a dilemma for a productive discussion since she had to reason about the depth of scientific explanation to which she would like to hold her students accountable. She used her interaction with the LP to describe a dilemma that she had yet to resolve in her teaching. She was skeptical of the upper anchor and questioned what the implications were for student learning if she placed herself squarely in that cell. Probably she and other teachers did not have a fully formed vision of what the uppermost practice in the Reduced MBI LP looked and sounded like or what might be possible with scaffolding. Thus it seems that the Reduced MBI LP has the potential to surface new and potentially productive lines of thinking for teachers. Perhaps answers to students' learning dilemmas lie not in a teacher development LP but rather in a student LP.

Moreover, in one case, Sarah used multiple criteria of the Reduced MBI LP to generate a vision of what her practice could look like the following year. In using the

Reduced MBI LP she reasoned through the relationships among the upper cells for all four criteria. She was the only teacher who described the ambitious practices as integrated. She was also the only teacher who read down a column, not across a row, of the Reduced MBI LP. Then she quickly set a new goal to work on multiple criteria simultaneously in the upcoming school year. She began by stating that she imagined herself as the kind of teacher who worked in the upper anchors of the LP.

> Sarah: [referencing LP criteria # 1–4] Well I always wanna get to the top level. So it kind of goes hand in hand with the pressing for explanations [criterion 4] like I would say that I probably do a lot of process focus and then have been thinking a lot about theory focus [criterion 1] this year especially in using evolutionary theory as like an overarching theory that we can hook ideas onto. I guess it makes me think that what I really should be working on is when I'm planning my units for next year to really figure out how to include more of this idea of model testing [criterion 3] or kind of revisiting ideas and …adjusting thinking as we go through a unit [criterion 2]. That would be a goal for next year…if you're doing all then they should all move together. And so I think it's really important for them [students]…if I'm working towards having this theory focus or fluency then it's really important that those aren't just separate things that somehow they're making those connections between the ideas like building a framework and hanging their little ideas on it. Which is part of how that whole theory of how students learn.

Sarah used the LP as a way to develop sophisticated ideas about teaching and learning science. It may be that Sarah accepted the tools we provided without question. However, it may also be likely that her goals matched the descriptions of teaching and learning advocated in the LP. Thus she could begin interacting with the LP in a different way because she was constantly searching for language and structures that helped her better articulate a sophisticated theory of science teaching and learning. This alignment may have given her the opportunity to set a new goal for herself–a goal she has consistently worked on in her second year of teaching.

LP as a non-vision tool. One challenge in using the Reduced MBI LP with teachers was that only two-thirds of the teachers used their interaction with the Reduced MBI LP as an opportunity to re-state, clarify, productively complicate, revise, and/or expand their visions of ambitious teaching and learning in their classrooms. The other one-third of the teachers merely used the Reduced MBI LP to assess their practice. These teachers located their practice in lower anchors on the Reduced MBI LP, citing their own and their students' limited abilities as justification. They expressed satisfaction with their current practices and did not express any need to alter their practices. For example, Adam located his teaching for criteria 1 and 4 as a matter of fact. He cited first his limitations and then his students' limitations, both of which prevented him from moving beyond the lower anchor for criterion 2.

> Adam: [referencing LP criteria # 1, 2, 4] Okay, definitely more on this side, but sometimes going into that. Because I'm not pressing them [students] for why

> and I'm just focusing on vocab… I don't think I will feel comfortable at all to go like with this one [upper anchor of criterion 2]. I think I have to be pretty advanced before that happens. I feel like, especially with the class I've never taught before I really need to have that planned out. I'd be skeptical to just let them…I don't know what these students are like yet. I know they are definitely considered usually underachievers.

In talking about an elective class he will teach for the first time in the next year, Adam cited his limitations as a novice teacher and mentioned that students might be incapable of engaging in more open-ended conversations. It may be that a cultural vision of low expectations and the need for simplistic classroom dialogue that controls classroom behavior (Gilbert & Yerrick, 2001) is working against the development of a more sophisticated vision for what Adam can do with underserved students. Adam did not regard the Reduced MBI LP as a useful tool for helping him envision his future practice; instead, the Reduced MBI LP reinforced his own hesitancy about attempting more ambitious teaching practices.

This third of the teachers were also more likely to provide a slightly inflated assessment of their own practice (particularly on the first criterion, Selecting and Treating Big Ideas as Models). We think this is partly because they may not share a vision of what was meant by each cell in the Reduced MBI LP. Moreover, they only nominally used the language from the LP (see Thompson et al., 2010), as compared to other teachers who had wrestled with language and ideas from the Reduced MBI LP criteria over the course of the previous school year. Without aligned visions for ambitious teaching and the negotiation of these visions with new tools, routines, and language, these teachers were unable to advance their visions or teaching.

As a part of our current study we are considering ways to address the challenge that the Reduced MBI LP does not support a shared vision of high-leverage practices for all novice teachers. For example, we filmed four teachers using ambitious practices in a two-week unit of instruction. As a part of their science teaching methods course in our teacher education program, pre-service teachers viewed these films and answered targeted questions that orient their reflections toward teaching moves that are critical to enacting upper anchor performances. We are also working with mentor teachers to develop shared visions of the practices. In this way, as the novices first attempt ambitious practices during student teaching, they will receive feedback from experienced teachers. But just enhancing visions of the high-leverage practices may not be enough. If enhancing one's vision is dependent on evaluating one's pedagogical identity, then we will need also to use teachers' self-narratives about their teaching practices. We will need to help teachers recognize that their current theories of teaching and learning overlap or are juxtaposed with the high-leverage practices and their effect on student learning. The teachers' current narratives of teaching and learning were most likely developed during their "apprenticeship of observation" (Lortie, 1975) and condition what they then learn in training experiences (Linn, Eylon, & Davis, 2004). If this initial understanding is not engaged with or confronted during teacher preparation, teachers may fail to grasp new concepts about teaching and learning or they may learn them for the purposes of a test only to return later to their preconceptions (Darling-Hammond & Bransford, 2005).

Challenge 3: Developing Multiple Tools to Fill Vision-to-Practice Gaps

The same teachers who used the Reduced MBI LP as a vision tool also used the Explanation Tool and socio-professional routines to imagine the next steps in their teaching. (The Explanation Tool helped teachers diagnose levels of evidence-based explanations that students used in written work; the Critical Friends Group meeting supported teachers in analyzing patterns in students' work). Yet the use of these tools led to different visions for teachers. The Reduced MBI LP helped most teachers set goals for moving their practice forward, but the details of how to accomplish these shifts remained vague. The Explanation Tool and socio-professional routines, however, had four features that supported the appropriation of a high-leverage practice (Pressing for Explanation) and the vision behind the practice. Specifically these features (1) included pedagogical ideas congruent with ambitious teaching, (2) focused on the relationship between teaching practice and student thinking, (3) were used directly to plan for, enact, or assess instruction, and (4) were used in collaborative settings in pre-service and in-service contexts (Thompson et al., 2010). Because the Explanation Tool was closer to practice, the teachers used this tool to support experimentation with the high-leverage practice as well as to support tool-based pedagogical innovation. By pedagogical innovation we mean that the teachers used the tools and routines as expected, but after quickly adopting both the vision and the practice, they began to re-purpose the tool and create a related set of practices that supported the vision of pressing for evidence-based explanations. For example, a couple of teachers used the Explanation Tool, originally designed for teacher analysis of student work, as a way to structure explanations in classrooms. Students in one classroom debated the difference between providing a "how" explanation and a "why" explanation for a phenomenon.

Mapping observational data onto the Reduced MBI LP helped identify critical developmental targets for the development of a suite of tools for novice teachers. We "tracked the fate" of each high-leverage practice over time and were able to determine which practices were the most difficult to attempt, which were easier to attempt regardless of mandated curricula, which practices seemed more sensitive to other pedagogical tools we used, and which practices showed regressive tendencies toward more traditional forms of teaching during the teachers' first year of teaching compared to their year of student teaching. Thus the Reduced MBI LP illuminated dimensions of teacher learning that needed further support through tool development. For example, few teachers were able to use scientific models with theoretical components in their practice. Often curricula did not support teachers in developing a big scientific idea that could be treated as an explanatory model for a class of scientific phenomena. To this end, we developed a Big Idea Tool that teachers can use when planning to convert topics and processes to explanatory models. We also found that teachers either made clunky attempts or did not attempt the following types of conversations in classrooms: (1) eliciting students' ideas to adapt instruction, (2) helping students make sense of classroom labs and activities, and (3) pressing students for evidence-based explanations. In response we developed discourse tools that decompose instructional acts; that is, such tools are effective in "breaking down complex practice into its constituent parts for the purposes of teaching and learning" (Grossman et al., 2009, p. 8). These tools now help our current cohort of novice teachers plan for and enact the ambitious teaching practices that present the greatest challenges for the previous novice teacher groups. When combined with ongoing analyses of student learning, supported by tools like the Explanation Tool and the Reduced MBI LP and the socio-professional routines of collaborative inquiry within a Critical Friends Group, we hope to create a powerful context that supports novice teacher learning of ambitious and equitable teaching practices (see tools4teachingscience.org). With support from a LP, among other tools, we hope that all novice teachers can advance their practice early in their careers rather than experience a time of stasis or regression toward conservative teacher-centered instruction as is often documented for new teachers (see Abd-El-Khalick & Lederman, 2000; Bransford & Stein, 1993; Goodlad, 1990; Grossman et al., 2000; Nolen et al., 2009; Simmons et al., 1999).

CASE 2: A LP TO SUPPORT EVERYDAY ASSESSMENT PRACTICES IN TEACHING NATURAL SELECTION

The goal of the second study – called the Daphne Project - was to create a LP to support everyday assessment practices in teaching natural selection. The project engaged a department of biology teachers to explore and develop a LP and to create a suite of formative assessment tools designed to tap the ideas represented in the LP. It was hypothesized at the outset of the study that engaging in the process of developing this LP would help teachers be better prepared to recognize and act upon student thinking, which in turn would affect student learning by closing what has been called the 'feedback loop' in the formative assessment process (Black & Wiliam, 1998; NRC, 2001).

Previous research indicates that, even in the presence of a framework for how student ideas develop over time in a conceptual domain, teachers struggle to enact formative assessment and to provide timely feedback to students (Furtak et al., 2008). Based on this finding, the Daphne Project sought to support teachers in adopting the following ambitious teaching practice: listening and attending to student thinking on a daily basis for the purpose of modifying instruction while it was in progress.

Seven biology teachers and their students at Springfield High School participated in the project. The high school is an ethnically and socioeconomically diverse school near a large city in the western United States. These teachers have a wide range of backgrounds and experience, from a second-career student teacher to a 15-year veteran biology teacher. Students in their classes were enrolled in one of three levels of biology: 9th grade pre-International Baccalaureate Biology, 10th grade General Biology, and 10th grade English Language Learner (ELL) Biology.

The study took place in three academic years: the first year (Y1) was a pilot study with other teachers in the district, and the second and third years (Y2 and Y3) involved only the teachers at Springfield High School.

The Daphne Project took a design-based approach to the development of the LP (Brown, 1992; Sandoval & Bell, 2004; Sloane, 2008) in which draft documents were refined and adapted during the study in response to the context in which they were used. Thus the teachers did not follow an established protocol in enacting formative assessments in their classrooms; rather, they used the tools in ways they thought best suited to the students. Then the group discussed the results in order to refine the tools.

Challenge 1: Creating a Format for a LP That Has Instructional Utility

The Daphne Project began with a draft LP, which was developed by the research team based on existing literature (Anderson, Fisher, & Norman, 2002; Bishop & Anderson, 1990; Dagher & Boujaoude, 2005; Ferrari & Chi, 1998; Geraedts & Boersma, 2006; Shtulman, 2006) and a pilot study. This LP, called a "map of student ideas" in the study, was presented to the Springfield High School Biology Department as an organizer for instruction during Y2 of the study. Its four basic levels start with unclear or undifferentiated ideas (Level 1), followed by the classic "misconceptions" about natural selection (Level 2), a blending of these "misconceptions" with terms used to describe natural selection (Level 3), and the classic "correct" understanding of natural selection (Level 4). Although many "flavors" of student misconceptions about the origin of new traits exist, it is unclear in the research if any LP exists for the multiple, common misconceptions students have about natural selection (e.g., "environmentally-induced" versus "anthropomorphic" changes). The draft or hypothesized LP contained the list of student ideas and example student responses collected during the Y1 pilot study (see Figure 3).

Level	*Category*	*Description*
4	Natural selection	Changes in populations and species occur gradually over time, due to the fact that all organisms produce more offspring than are able to survive. These offspring vary among themselves, and some of the variation can be passed on to the next generation. thus, those offspring that are able to survive will happen to be those best suited their environments. Over many generations, accumulated changes will lead to changes in isolated populations and, in some cases, the generation of new species.
3	Blended (natural selection and need-based change)	Mixture of ideas based on features of natural selection as well as need-based change, possibly including descriptions of the principles of 'survival of the fittest' and/or 'non-inheritance of acquired traits'
2	Need-based change	*Anthropomorphic:* Changes arise from deliberate acts by the parent organism or the species as a whole *Environment:* Changes emerge because of environmental conditions; i.e., the environment causes changes in organisms to occur *Teleological/Essentialism:* Organisms change over time to reach a given endpoint/essence of the species
1	Unclear/ other	Ideas that are confusing, ambiguous, or do not fall into other categories

Figure 3. Original learning progression for natural selection.

However, it quickly became apparent in Y2 that the content of the LP we gave the teachers was insufficient to scaffold their instruction in the unit. Primarily the LP contained ideas along a continuum towards understanding where new traits originate and the consequences of those traits in terms of selection. Teachers immediately suggested that the misconceptions be linked to the "Five Facts and Inferences" about natural selection that form Darwin's argument for natural selection (Mayr, 1997) as an orienting framework for structuring instruction. These facts summarize the struggle for existence among individuals within a population, differential survival, and reproduction, and show how these changes build over many generations. We discussed ways that these two different representations might be merged and ultimately developed a combination core LP that reflected a progression of correct ideas horizontally, as well as the hypothesized developmental paths from misconceptions to those correct ideas, from less sophisticated at the bottom to more sophisticated at the top (see Figure 4).

The revised LP combines features of LPs designed as frameworks for curricula with lower anchors as the foundational correct understandings (Catley, Lehrer, & Reiser, 2005) and features of LPs anchored by students' common prior ideas, as suggested by Shepard (2009) and Wilson (2009). In this way, the LP includes both an overarching framework to inform scope and sequence as well as a map of the common prior ideas that can often obfuscate the path forward for teachers. The horizontal sequence of correct ideas, taken directly from Mayr (1997), lists the ideas as they should be learned, from left to right. It also includes two vertical sequences (*Origin & Inheritance of Traits* and *Selective Force* constructs) that were identified through analyses of student responses to the assessments in the Daphne Project in comparison with prior research on student ideas about natural selection. These vertical sequences start at the bottom with common student misconceptions and build to the correct ideas represented in the horizontal sequence. The LP is paired with accompanying tools intended to scaffold instruction, as well as a professional development model to support teachers in changing their practices. The LP for natural selection includes a set of formative assessment prompts designed to be embedded in the curriculum that provide opportunities for teachers to elicit student thinking and map that thinking to the LP, common student responses to each assessment at each level, and suggested feedback strategies for students with different ideas (Furtak, 2009).

The result was a set of formative assessments that targeted different ideas on the *Origin & Inheritance of Traits* and *Selective Force* constructs, including a constructed-response prompt ("How do species change over time?"), a multiple-choice plus-justification prompt, an evaluation of Peter and Rosemary Grant's data on Darwin's finches in the Galapagos Islands, and interpretations of real-life scenarios. These formative assessments are explained in greater detail in Furtak (2009); the real-life scenarios formative assessment is shown in Figure 5. Piloting and analysis of the implementation of these prompts led to a collection of student responses for each level of the LP, as well as suggested feedback strategies (see Table 1).

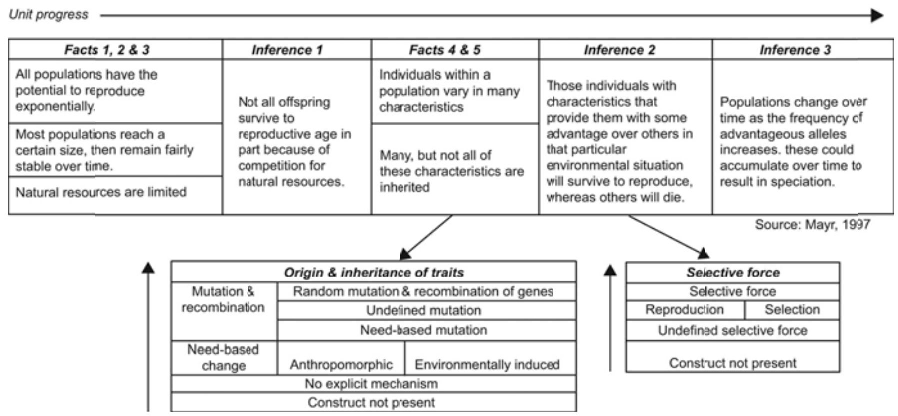


Figure 4. Revised learning progression for natural selection.

Table 1. Sample Student Responses and Feedback Strategies Linked to Need-Based Conceptions on the "Origin & Inheritance of Traits" Construct.

Description	*Example Student Responses*	*Suggested Feedback*
Anthropomorphic: Changes arise from deliberate acts by the parent organism or the species as a whole	"The moths become darker because of bark." "The moths would change their color to a darker color to blend in with the dark bark."	Examine evidence collected in class activities and focus upon where variation originated Explore Lamarck's idea of descent with modification, and compare to modern ideas about genetic inheritance
Environmentally Induced: Changes emerge because of environmental conditions; i.e., the environment causes changes in organisms to occur	"When food and other resources become scarce, the finches develop over a period of time different beaks to eat the hard seeds that seem to be abundant." "Animals mutate to fit in with their natural surroundings. So becoming darker helps to keep them in camouflage."	Explore differences between mutations in an individual and population changes over time Discuss instances in which individuals change in response to the environment (e.g. arctic hares), and how this differs from changes in the proportion of genes in a population

As we continued to refine the formative assessments in Y2 and Y3, the LP proved to be a continuously changing target. As soon as formative assessments were enacted, teachers returned to suggest ways in which they might be changed and improved. These new formative assessments yielded new examples of student ideas, and, in one case, surfaced a new category of student idea that we had not previously identified in the literature. This idea, which the teachers named "eugenic," has to do

5	Arctic hares live in the harsh environment of the North American tundra. In the winter, the hares have a white coat of fur, helping them to blend in with ice and snow. In the summer, their coat turns a gray-blue that matches rocks and plants in their environment.	☐ Natural Selection ☐ Individual Change *Why do you think so?*
6	Antibiotics are used to kill bacteria. Since the 1950's, however, scientists have noticed that some bacteria are not be killed by some kinds of antibiotics. When you take an antibiotic, all the non-resistant bacteria are quickly killed and the remaining resistant bacteria are then left to reproduce and thrive.	☐ Natural Selection ☐ Individual Change *Why do you think so?*
7	While some Canada Geese still fly south for the winter, other 'resident' populations of Geese live in the same place all year. Migration – the seasonal journey in response to food availability, weather, or habitat - is a learned behavior, so once the adults in a population stop migrating, their offspring will not also not migrate south.	☐ Natural Selection ☐ Individual Change *Why do you think so?*

Figure 5. Sample formative assessment prompt.

with a misunderstanding of sexual selection. Students with this idea argue that, in changing environments, animals select mates on the basis of characteristics that will be beneficial in the new environment. The research team found itself challenged to keep up with the pace of changes being made, and at each meeting, new copies of the LP or formative assessments were provided. The revised LP was based on teachers' expressed needs when teaching the natural selection unit. Thus the LP became a tool used to support teachers in adopting formative assessment practices.

Challenge 2: Developing a Vision Tool for All Teachers

The Daphne Project addressed the challenge of developing a vision tool for the teachers by viewing their diverse expertise as a resource for the study. The project employed a model of sustained professional development intended to engage teachers in a process of inquiry into how to create opportunities to elicit and respond to students' ideas in their own classrooms. In Y2, the professional development took the form of a professional learning community (PLC) consisting of the principal investigator (PI), the seven biology teachers, and two research assistants. These stakeholders brought myriad perspectives to the PLC, allowing for the active exchange of ideas based on their diverse expertise. The PLC met monthly during the school year in Y2 and Y3. PLC activities were modeled on problem-solving and assessment development cycles from mathematics education (Borko, 2004; Webb, Romberg, Burrill, & Ford, 2005). The purpose of these activities was to facilitate teachers' reflections on their current practices and their adoption of new practices supported by the LP.

Figure 6 illustrates the four-stage iterative professional development model employed in the Daphne Project. The first step is for teachers to *Reflect* upon their current practice, sharing instructional strategies and materials with each other and discussing their advantages and disadvantages (Loughran, Mulhall, & Berry, 2004). In the second step, teachers *Explore* student thinking in their classes and compare those ideas with the LP. In this step, a particular focus is placed on deepening teachers' content knowledge as well as on understanding students' common misunderstandings, alternative conceptions, and the sources of those

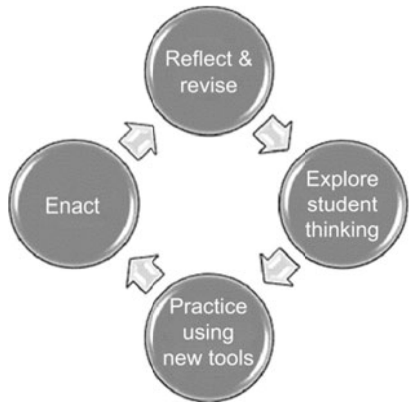


Figure 6. Professional development model.

ideas. In the third step, teachers *Practice* using the associated tools and formative assessments, categorizing samples of student work, anticipating feedback, and discussing videotapes of each other in enacting the assessments. At this stage, it is essential to encourage teachers to discuss not only the identification of student ideas but also the kinds of feedback needed by students with each category of thinking. In the fourth step the teachers *Enact* their natural selection units with the tools contained in the LP, after which they continue the cycle by reflecting upon their enactment and *Revise* their plans for next time.

All these professional development activities were intended to focus teachers on discussing and anticipating the feedback related to particular ideas in the LP. These experiences indicate that thinking about feedback in advance is incredibly difficult, even in a small group of seven teachers and three researchers. In the end, while some ideas were tossed around in meetings with teachers, the researchers, in consultation with a biologist, suggested most of the feedback ideas. Furthermore, when they were asked to anticipate the feedback they would give to a class, considering the seven responses (see below) they had considered in the post-interview at the end of Y2, most teachers described giving feedback to a whole class, rather than tailoring feedback to individual students.

Despite the challenges we faced in getting teachers to discuss feedback in interviews and in the PLC, we still observed transformations in the PLC in terms of changes in the language teachers used to describe student thinking and in the modifications of how the instructional practices were shared (see Furtak, Morrison, Henson, & Roberts, 2010). For example, teachers began the study by sharing stories about what had happened in their classrooms when students raised questions based on creationist theories about the origin of life. Later, teachers shared specific examples of students with misconceptions and talked about ways those ideas might be addressed through instruction. In addition, they moved away from talking about student ideas that were “wrong” toward using the language in the LP to identify and discuss different student ideas.

While the involvement of multiple stakeholders – biology teachers, researchers, and a biologist – proved effective in anticipating some feedback strategies, the extent to which the biology teachers actually tried these feedback strategies in their classrooms varied as much as their approaches to enacting the formative assessments. That is, despite the efforts to develop a LP that would scaffold teachers’ adoption of the ambitious teaching practice of formative assessment, the extent to which teachers changed the way they viewed student thinking and enacted new teaching practices varied greatly in the PLC. For example, teachers enacted the formative assessments in various ways, some of which were better suited to getting students to share their ideas than others. Furthermore, several teachers retained a right-or-wrong attitude about student ideas, merely creating more detail in the “wrong ideas” category.

Challenge 3: Developing Multiple Tools to Fill Vision-to-Practice Gaps

The formative assessments developed collaboratively in the PLC were intended as the tools that would scaffold teachers in eliciting, identifying, and responding to the student ideas represented in the LP. The process of collaboratively developing these formative assessments, looking at student responses, matching elicited ideas with those represented on the LP, and discussing the sources of those ideas helped teachers move the vision of student understanding represented in the LP into their own practice.

The case of one teacher, Lisa, illustrates the process by which the teachers used formative assessment activities developed in the PLC in their classrooms and then used the LP to categorize the student ideas that were shared in response. Lisa explained that she had asked her students to correct examples of student responses taken from the draft LP. Then she categorized the student work into three groups. She called these groups “they got it,” “they sort of got it,” and “what the hell?” In examining the content of each group’s responses, she found the “they got it” responses were the correct answers and the “they sort of got it” responses included student statements similar to “The moths changed” but without explication of the mechanism the student was applying for how the moths changed. She found the “what the hell?” responses included either misconceptions or just something that she couldn’t understand. This categorization procedure indicates that Lisa was beginning to distinguish between the types of student alternative conceptions – in effect, using the formative assessment tools to inform her teaching practice.

Post-interview results suggest that Lisa’s categorization of student thinking was not unique. At the end of the study, the teachers were each asked to complete a sorting task in which they were given seven different student responses from one of the formative assessments. They were asked to sort these responses into categories that made sense to them. Results of this interview task indicate that the teachers agreed on three of the seven responses (7, 3, and 6) as "correct." In general, they grouped the ideas based on alternative conceptions together. However, the teachers fell into three groups with respect to the way they subdivided these categories. These subdivisions were aligned along a continuum roughly corresponding with the LP according to teachers who created similar categories clustered together (see Figure 7). Two teachers (Robyn and Rachel) sorted the responses into four groups representing a continuum of student thinking; three teachers (Theresa, Chris, and Alison) created three categories that included correct answers, incorrect answers, and unclear answers; and two teachers (Lisa and Megan) grouped the student responses as "right" and "wrong" answers.

For example, Robyn explained that the responses represented "Kind of a continuum, as far as I'm concerned." In contrast, Megan explained that half the responses were "Closer to being the right idea and the right concept," whereas the other responses were "the most way off." These results indicate that teachers did not agree on a single interpretation of the different pieces of student work. In future iterations of the project, we plan to speak more explicitly about the LP as a representation of ideas that goes beyond right and wrong and the importance of identifying and acting upon different student responses. Furthermore, additional tools could further support teachers in acting upon different student responses, such as video cases of teachers eliciting student ideas, identifying responses, and giving feedback. In addition, structured practice sessions could give teachers opportunities to try out different feedback strategies tailored to student ideas on the LP.

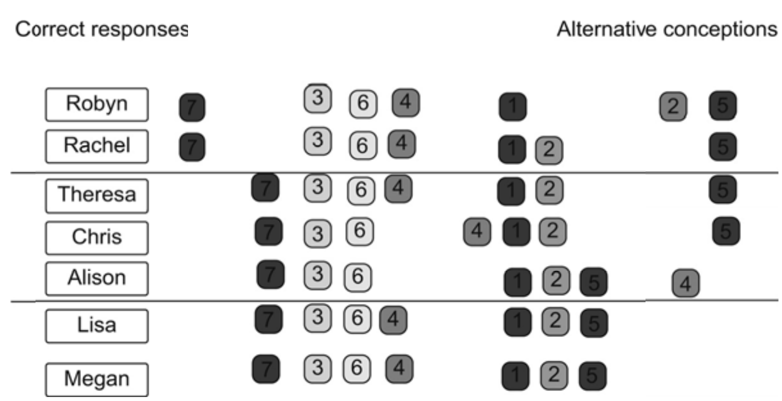


Figure 7. Teachers' responses to post-interview sorting task.

CONCLUSIONS AND IMPLICATIONS

In this chapter, we have discussed two types of LPs that can be used to support teacher development: LPs that map ambitious teaching practices from novice to expert (Case 1), and LPs that represent student ideas to support teachers' ambitious teaching practices (Case 2). These cases use LPs as tools that represent the knowledge needed to teach a particular concept: therefore, they are tools that help teachers develop the knowledge and practices they need to become more effective teachers. The struggles experienced by the teachers in the two cases indicate that the educational research community needs to focus on the "user" end when developing LPs. These struggles also raise a number of questions for discussion by the LP research community. For example, what are the advantages and disadvantages for teachers of the two LP types–those mapping teacher development and those mapping student ideas intended to support teacher development? What are the trade-offs involved with simultaneously developing and using LPs at a school site? Is it possible to include sufficient information in a LP to support teachers in enacting ambitious teaching practices? If so, how would a teacher who is not involved in the development of the LP use it as an instructional tool?

DESIGN PRINCIPLES: LPS TO SUPPORT AMBITIOUS TEACHING PRACTICES

Based on our experiences in the studies described above, we suggest three design principles to inform the creation of LPs that support teacher development. Such LPs should:

1. provide a pedagogical vision,
2. support closing vision-to-practice gaps, and
3. be embedded in ongoing professional development in a community of inquiry.

We describe each design principle next.

Design Principle 1: LPs should provide a pedagogical vision. LPs intended for teacher development should represent a pedagogical vision–that is, a progression of content and practices that is developed over time and that is anchored at the bottom by preliminary ideas or practices and at the top by the most advanced/sophisticated ideas and practices. The content and practices should represent the hypothesized trajectory of development, thus specifying the starting point for teachers as well as the various intermediate stages that are important for achieving competence. Furthermore, the simple act of creating a LP can give the false impression that it dictates a linear progression of development even though teacher and student practices have been shown to develop in multiple ways and not necessarily in the same sequence as a LP suggests. Thus LPs should contain multiple paths to competence as part of a more complete pedagogical vision.

Design Principle 2: LPs should support closing vision-to-practice gaps. Before teachers can take instructional action on the basis of student ideas, they need the appropriate tools that support them in adopting the pedagogy represented in the LP that connect with what they are already doing - that is, tools that will help them close the gap between their vision and current practice. Such tools should have a variety of formats and should be linked to the LP in order to elicit the ideas that appear in the LP. In this way, levels can be identified. Tools linked to the LP should be formative in nature; that is, such tools should provide for feedback that helps close the gap between current practice and the ultimate vision. The tools should also be generative by inviting students to share a range of ideas.

Design Principle 3: LPs should be embedded in ongoing professional development in a community of inquiry. The two cases in this chapter suggest that collaborative inquiry among teachers centered on a LP – either built from teaching practices (Case 1) or built by teachers around student ideas (Case 2) – can support teacher development by requesting that teachers take an inquiry position (Cochran-Smith & Lytle, 2009) toward their practice and be responsive to particular groups of teachers. Creating a culture of teachers who inquire into their practices around a set of ambitious teaching practices articulated in and around LPs has the potential to help the profession consistently improve instructional efficacy based on evidence of student thinking (Huberman, 1995). Furthermore, making LPs the centerpiece of communities of teacher inquirers may support teachers who, with their colleagues, make meaningful changes to their practice that are centered on the LP vision. Such an approach means that schools and school districts must provide the time and resources teachers need if they are to participate in professional development that supports their adoption of ambitious teaching practices.

Finally, we do not, by any means, suggest that LPs for teacher development alone are sufficient to achieve lasting changes in teacher practice. Rather, we argue that LPs embedded in suites of tools and used in the context of sustained professional development have the potential to support teachers who adopt ambitious teaching practices. Although much interest is currently focused on the development of LPs, the education community should proceed with caution. LPs, like all instructional tools, are not a universal solution to the challenges presented by science education reform. Much research is still needed to help determine the utility of LPs across multiple contexts for a variety of users.

ACKNOWLEDGEMENTS

This chapter is based upon work supported by a Research Fellowship from the Knowles Science Teaching Foundation and work supported by the Carnegie Corporation, the Annenberg and Rockefeller Foundations, and the National Science Foundation (Grant No. DRL-0822016). Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of these organizations.

NOTES

1. See Grandy & Duschl, 2007; Kitcher, 1993; Lehrer & Schauble, 2006; Longino, 1990; Nersessian, 2005; Smith, Maclin, Houghton, & Hennessey, 2000; Stewart, Passmore, Cartier, Rudolph, & Donovan, 2005; Windschitl & Thompson, 2006.
2. See Duschl, 2008; Metcalf, Krajcik, & Soloway, 2000; NRC, 2005; Schwarz & White, 2005.
3. See Berliner, 2001; Bransford & Stein, 1993; Darling-Hammond & Bransford, 2005; Grossman et al., 2000; Linn, Eylon, & Davis, 2004; Windschitl & Thompson, 2006.

REFERENCES

Abd-El-Khalick, F., & Lederman, N. (2000). The influence of history of science courses on students' views of the nature of science. *Journal of Research in Science Teaching, 37*, 1057-1095.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching, 39*, 952-978.

Baldi, S., Jin, Y., Skemer, M., Green, P., Herget, D., & Xie, H. (2007). *Highlights from PISA 2006: Performance of U.S. 15-year-old students in science and mathematics literacy in an international context*. Retrieved from National Center for Education Statistics website: http://nces.ed.gov/pubs2008/2008016.pdf

Ball, D. L., Sleep, L., Boerst, T. A., & Bass, H. (2009). Combining the development of practice and the practice of development in teacher education. *The Elementary School Journal, 109*, 458-474.

Banilower, E., Smith, P. S., Weiss, I. R., & Pasley, J. D. (2006). The status of K-12 science teaching in the United States: Results from a national observation survey. In D. Sunal & E. Wright (Eds.), *The impact of the state and national standards on K-12 science teaching* (pp. 83-122). Greenwich, CT: Information Age Publishing.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463-483.

Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching, 27*, 415-427.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5, 7-74.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Bransford, J. D., & Stein, B. S. (1993). *The IDEAL problem solver* (2nd ed.). New York, NY: Freeman.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*, 141-178.

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution*: Paper commissioned by the National Academies Committee on Test Design for K-12 Science Achievement.

Cochran-Smith, M., & Lytle, S.L. (2009). *Inquiry as stance: Practitioner research for the next generation*. New York, NY: Teachers College Press.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Crawford, B., & Cullin, M. (2004). Supporting prospective teachers' conceptions of modeling in science. *International Journal of Science Education, 26*, 1379-1401.

Dagher, Z. R., & Boujaoude, S. (2005). Students' perceptions of the nature of evolutionary theory. *Science Education, 89*, 378-391.

Darling-Hammond, L., & Bransford, J. (2005). Preparing teachers for a changing world: What teachers should learn and be able to do. San Francisco, CA: Jossey-Bass.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic and social learning goals. *Review of Research in Education, 32,* 268-291.

Elmore, R. (2005, February). *Our best ideas about instructional practice do not operate at scale*. Presentation at NSF K-12 Math, Science, and Technology Curriculum Developers Conference, Washington, DC.

Ferrari, M., & Chi, M. T. H. (1998). The nature of naive explanations of natural selection. *International Journal of Science Education, 20*, 1231-1256.

Franke, M. L., & Chan, A. (2007, April). *Learning about and from focusing on routines of practice.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Furtak, E. M. (2009, April). Learning progressions to support teacher learning. In A. C. Alonzo (Chair), *Learning progressions in science: Tools for teaching, learning, and assessment.* Symposium conducted at the annual Meeting of the American Educational Research Association, San Diego, CA.

Furtak, E. M., Morrison, D. L., Henson, K., & Roberts, S. A. (2010, July). *Centering a professional learning community on a learning progression for natural selection: Transforming community, language, and instructional practice.* Paper presented at the International Conference of the Learning Sciences, Chicago, IL.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*, 360-389.

Geraedts, C. L., & Boersma, K. T. (2006). Reinventing natural selection. *International Journal of Science Education, 28*, 843-870.

Gilbert, A., & Yerrick, R. (2001). Same school, separate worlds: A sociocultural study of identity, resistance, and negotiation in a rural, lower track science classroom. *Journal of Research in Science Teaching, 38,* 574–598

Goodlad, J. (1990). *Teachers for our nation's schools*. San Francisco, CA: Jossey-Bass.

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96*, 606-633.

Grandy, R. & Duschl, R. (2007). Reconsidering the character and role of inquiry in school science: Analysis of a conference. *Science & Education, 16*, 141-166.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 1-25.

Grossman, P., Valencia, S., Evans, K., Thompson, C., Martin, S., & Place, N. (2000). Transitions into teaching: Learning to teach writing in teacher education and beyond. *Journal of Literacy Research, 32*(4), 631–662.

Hatch, T., & Grossman, P. (2009). Learning to look beyond the boundaries of representation: Using technology to examine teaching (Overview for a digital exhibition: Learning from the practice of teaching). *Journal of Teacher Education. 60*, 70-85.

Horizon Research International. (2003). Special tabulations of the 2000-2001 LSC teacher questionnaire and classroom observation data. Chapel Hill, NC: Author.

Huberman, M. (1995). The model of independent artisan in teachers' professional relations. In J.W. Little & M.W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, and contexts* (pp. 11-50). New York, NY: Teachers College Press.

Kitcher, P. (1993). The advancement of science: Science without legend, objectivity without illusions. New York, NY: Oxford University Press.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and scientific literacy: Supporting development in learning in context. In W. Damon, R. M. Lerner (Series Eds.) & K. A. Renninger, & I. E. Sigel (Volume Eds.), *Handbook of Child Psychology: Vol. 4. Child psychology in practice.* (6th ed., pp. 153-196). Hoboken, NJ: John Wiley and Sons.

Linn, M. C., Eylon, B.-S., & Davis, E. A. (2004). The knowledge integration perspective on learning. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education* (pp. 29-46). Mahwah, NJ: Erlbaum.

Little, J. W., & Horn, I. S. (2007). 'Normalizing' problems of practice: Converting routine conversation into a resource for learning in professional communities. In L. Stoll & K. S. Louis (Eds.), *Professional learning communities: Divergence, depth, and dilemmas* (pp. 79-92). New York, NY: Open University Press.

Longino, H. (1990). Science as social knowledge: Values and objectivity in scientific inquiry. Princeton, NJ: Princeton University Press.

Lortie, D. (1975). *Schoolteacher: A sociological study*. Chicago, IL: University of Chicago Press.

Loughran, J., Mulhall, P., & Berry, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching, 41*, 370-391.

Mayr, E. (1997). *This is biology*. Cambridge, MA: Harvard University Press.

McDonald, S. P. (2008). Seeing the science: Professional pedagogical vision for instructional leaders. *Catalyst for Change*, *35*(2), 12-18.

McLaughlin, M., & Talbert, J. (2001). *Professional communities and the work of high school teaching*. Chicago: University of Chicago Press.

National Research Council. (1996). *National science education standards*. Washington, D.C.: National Academy Press.

National Research Council. (2001). *Classroom assessment and the National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2005). *How students learn: Science in the classroom*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.

Nersessian, N. (2005). Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In M. Gorman, R. D. Tweney, D. Gooding, & A. Kincannon (Eds.), *Scientific and technological thinking* (pp. 17-56). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nolen, S. B., Ward, C. J., Horn, I. S., Childers, S., Campbell, S., & Mahna, K. (2009). Motivation in pre-service teachers: The development of utility filters. In M. Wosnitza, S. A. Karabenick, A. Efklides, & P. Nenniger (Eds.). *Contemporary motivation research: From global to local perspectives* (pp. 265-278). Ashland, OH: Hogrefe & Huber.

Sandoval, W. A., & Bell, P. (2004). Design-based research methods for studying learning in context: Introduction. *Educational Psychologist*, 39, 199-201.

Schwarz, C., & White, B. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*, 165-205.

Sfard, A., & Prusak, A. (2005). Telling identities: In search of an analytical tool for investigating learning as a culturally shaped activity. *Educational Researcher, 34*(4), 14-22.

Shepard, L. (2009). *From advocacy to adoption: Getting formative assessment into more classrooms.* Discussant remarks at the American Educational Research Association Annual Meeting, San Diego, CA.

Shtulman, A. (2006). Qualitative differences between naive and scientific theories of evolution. *Cognitive Psychology, 52*, 170-194.

Simmons, P., Emory, A., Carter, T., Coker, T., Finnegan, B., Crockett, D., … Labuda, K. (1999). Beginning teachers: Beliefs and classroom actions. *Journal of Research in Science Teaching, 36*, 930–954.

Sloane, F. C. (2008). Randomized trials in mathematics education: Recalibrating the proposed high watermark. *Educational Researcher, 37*, 624-630.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade students' epistemologies of science: The impact of school science experiences on epistemological development. *Cognition and Instruction, 18*, 349–422.

Smith, C. L., Wiser, M., Anderson, C.W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1-98.

Stewart, J., Passmore, C., Cartier, J., Rudolph, J., & Donovan, S. (2005). Modeling for understanding in science education. In T. Romberg, T. Carpenter, & F. Dremock (Eds.), *Understanding mathematics and science matters* (pp. 159–184). Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, J., Windschitl, M. & Braaten, M. (2010, March). *Developing a theory of ambitious teacher practice*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Webb, D. C., Romberg, T. A., Burrill, J., & Ford, M. J. (2005). Teacher collaboration: Focusing on problems of practice. In T. A. Romberg, T. P. Carpenter, & F. Bremock (Eds.), *Understanding mathematics and science matters* (pp. 231-251). Mahwah. NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009, April). Learning progressions to guide systems of formative and summative assessment. In A. C. Alonzo (Chair), *Learning progressions in science: Tools for teaching, learning, and assessment.* Symposium conducted at the annual meeting of the American Educational Research Association, San Diego, CA.

Windschitl, M., & Thompson, J. (2006). Transcending simple forms of school science investigations: Can pre-service instruction foster teachers' understandings of model-based inquiry? *American Educational Research Journal, 43,*783–835.

Windschitl, M., Thompson, J. & Braaten, M. (2008). How novice science teachers appropriate epistemic discourses around model-based inquiry for use in classrooms. *Cognition and Instruction, 26, 310-378*.

Windschitl, M., Thompson, J., & Braaten, M. (2011a). Ambitious pedagogy by novice teachers: Who benefits from tool-supported collaborative inquiry into practice and why? *Teachers College Record. 113*(7), 1311-1360.

Windschitl, M., Thompson, J., & Braaten, M. (2011b). *The beginner's repertoire: Proposing a core set of instructional practices for teacher preparation*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Orlando, FL.

JACOB FOSTER AND MARIANNE WISER

THE POTENTIAL OF LEARNING PROGRESSION RESEARCH TO INFORM THE DESIGN OF STATE SCIENCE STANDARDS

Many U.S. states will revise their science standards in the next few years. In making these revisions, the states have a unique opportunity to adopt the most up-to-date understandings about student learning drawn from learning progression research (e.g., Corcoran, Mosher, & Rogat, 2009). Learning progressions offer a connected view of the development of students' thinking and skills in core domains over extended periods of time. Learning progressions may provide added coherence, empirical grounding, and cognitive depth to standards development.

Student performance, as determined by science assessments, remains stubbornly low as well as disparate between student sub-groups; therefore, any insights into fostering meaningful student learning over time are critical (Corcoran et al., 2009). While many factors may account for such student performance, research suggests that some concepts would be better learned if they were introduced in a different grade band (e.g., Wiser & Smith, 2008) and that other concepts require prior understandings not specified by current standards (e.g., Smith, Maclin, Grosslight, & Davis, 1997). We need to use any available research (e.g., learning progression research) in standards development to address the disparity between students' conceptions and scientific conceptions and to ensure that the two are bridged successfully. This requires giving standards a stronger cognitive basis than they have at present. In the development of standards, a learning progression perspective may facilitate working from both a cognitive basis and a scientific basis. In particular, this perspective may guide the development of standards that explicitly target the reconceptualizations needed to transform students' ideas into sound scientific understandings. Some research in this regard is better than none, and even incomplete or prototype learning progressions may be useful.

Understanding the challenges of applying findings from learning progression research to standards development may help researchers structure, focus, and communicate their work in ways that will be more efficient, more powerful, and more useful to standards developers in the future. This chapter articulates some challenges the learning progression community, in collaboration with the states and other standards developers, should consider in order to ensure that learning progression research is used to inform the development or revision of state science standards. This chapter considers two sets of challenges: one related to the early state of learning progression research and the other related to the purpose and

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 435–460.

limitations of the states' science standards. For our discussion of these challenges, we use the Massachusetts standards revision process as our example. After a brief overview of the Massachusetts revision process, the chapter explains each set of challenges and describes how Massachusetts is responding to them. The purpose of presenting an early example of strategies for addressing these challenges is to stimulate a dialogue in the broader education community about additional possible responses.

AN OVERVIEW OF THE MASSACHUSETTS STANDARDS REVISION PROCESS

Massachusetts's current Science and Technology/Engineering ("science") standards (Massachusetts Department of Education [MADOE], 2006) are organized around four strands for grades PreK-8 (earth and space science, life science, physical science, and technology/ engineering) and five "introductory" high school courses (earth and space science, biology, chemistry, introductory physics, and technology/engineering).[1] Massachusetts's science standards were first articulated in 1996. A full revision was completed in 2001, and a second full revision, begun in 2009, has an expected completion date in 2013.[2]

A panel of volunteers is involved in the revisions of science standards for public schools in Massachusetts. The panel members are exemplary representatives from the state's science education community (consisting of K-12 teachers of science, school and district administrators, higher education faculty, and organizational/ community representatives). The panel identifies and references national documents (e.g., American Association for the Advancement of Science [AAAS], 1993; College Board, 2009; International Association for the Evaluation of Educational Achievement [IEA], 2007; International Technology Education Association [ITEA], 2000; National Assessment Governing Board [NAGB], 2008; National Research Council [NRC], 1996, 2007). The panel drafts and redrafts standards based on expert recommendations and gathers public comments on those drafts. The responsibility of the Massachusetts Board of Education is to review the draft standards; with the Board's approval, the draft standards are then formalized as the new state science standards.

The current revision process of the science standards in Massachussets is attempting to reflect a learning progression perspective. The goal of this process is to use primary source education research, including learning progression research, explicitly and systematically in order to ensure effective support for students' learning of key concepts and practices within and across grade bands.

CHALLENGES RELATED TO THE EARLY STATE OF LEARNING PROGRESSION RESEARCH

Challenges in Using Learning Progression Research to Specify the Content of Standards

While learning progression research holds great potential for informing standards development, using findings from this research, which is still in early stages, presents significant challenges. Existing research is not yet sufficient to support the full range of domains and grade spans typically addressed in the states' science standards. Because operational definitions of learning progressions differ among researchers, it is also difficult to present standards developers with a view of learning progression research that is coherent enough to be useful. There is a lack of consensus among learning progression researchers working on the same or related domains about what students should know by high school graduation and which *core ideas* comprise the large goals of PreK-12 science instruction. Making the best use of the existing research requires addressing each challenge. We examine these challenges next.

Challenge 1: The limited scope of available learning progressions to inform PreK-12 standards. Very few learning progressions have been considered, much less fully articulated, on the PreK-12 scale required by the states' science standards. Currently, learning progression research has explored only some science domains and only some grade bands within those domains. This does not mean that every learning progression should span the full PreK-12 grade bands; multiple learning progressions that address different grade bands can be linked as long as their constituent components are compatible (see Challenge 2 below).

To fill the gaps in current learning progression research, one can refer to existing literature that documents students' "alternative conceptions" (e.g., Driver, Guesne, & Tiberghien, 1985; Pfundt & Duit, 1994), "misconceptions" (e.g., Novak, 1993), lack of mastery of basic concepts and principles[3] (e.g., Hestenes, Wells, & Swackhamer, 1992), and "facets of knowledge" (e.g., Minstrell, 1991). This literature provides insights into student thinking and learning that can be recast within a learning progression perspective. For example, Wiser and Smith (2008) began their work on a learning progression for matter by analyzing the literature on students' ideas about atoms and molecules from a conceptual point of view. They established patterns among students' beliefs and traced the root of those patterns to other beliefs. This process revealed that an important source of students' difficulties with understanding atoms and molecules is the deep incompatibility of their conceptualization of matter at the macroscopic level with a scientific account (e.g., with the scientific concepts of weight and density and the principle that all matter has mass). Using this base, Wiser and Smith then looked at the cognitive development literature and hypothesized how, with proper instruction, by the end of elementary school very young children's knowledge about objects and substances could undergo the deep and broad reconceptualizations needed to develop a scientifically sound understanding of matter (at the macroscopic level). A challenge for standards developers is to apply this process in a way that leads students to a basic but solid understanding of a scientific theory without the extensive cycles of research that Wiser and Smith conducted.

Challenge 2: Integrating different conceptions of what constitutes a learning progression. Research groups working on different learning progressions have yet to reach consensus on what progresses in a learning progression. Existing learning progressions differ in whether they represent how students' ideas actually develop or in whether they represent ideal developmental paths. In the first view, the focus is on students: learning progressions are descriptions of how students' knowledge evolves, given their current curricular experiences. For example, Jin and Anderson (2007) describe stages that only some students reach in thinking about energy in socio-ecological systems. Lee and Liu (2010) use standardized test items to compare sixth, seventh and eighth graders' ideas about energy sources, transformation, and conservation. In the second view, the focus is on the knowledge itself: a learning progression describes a hypothetical succession of knowledge states where each state requires reconceptualization to transform to the next state. For example, Wiser and Smith (2008) treat their learning progression on matter as a hypothesis about the way knowledge ideally evolves from young children's ideas about materials and objects toward an understanding of atomic molecular theory. The two views of what progresses are not incompatible; learning progressions of the first kind are probably needed to begin developing and refining learning progressions of the second kind (Stevens, Delgado, & Krajcik, 2010).

Additionally, research groups working on learning progressions in the same or related domain(s) structure how knowledge progresses in different ways. They may organize a domain around different core ideas (see Challenge 4) or use different units of analysis. Units of analysis as used here are theoretical constructs used to parse knowledge and skills in ways that support meaningful interpretations with respect to classes of phenomena.[4] From the researcher's point of view, units of analysis must have a meaningful developmental path (i.e., one should be able to follow their progression from grade span to grade span as they increase in complexity and get closer to scientific knowledge).

For example, several learning progressions involve energy. Wiser and colleagues structure their learning progression around several interrelated ideas in a process they call "looking at phenomena through the energy lens" (Wiser, Tobin, & Lacy, 2011). In elementary school, this learning progression includes the search for pairs of correlated energy changes and the development of an awareness of energy dissipation as precursors to the high school concepts of energy conservation and entropy, respectively. Their unit of analysis is students' use of the components of the energy lens in different contexts. Jin and Anderson (2007) focus on reasoning patterns for the processes that transform carbon in socio-ecological systems (photosynthesis, digestion, combustion, etc.). Their units of analysis are the content and nature of students' explanations. While younger children focus on the macroscopic level, describing plants and animals as subject to different rules than inanimate objects, and food and fuel sources as enablers of life processes, more advanced students use chemical models that connect organisms and inanimate matter. Lee and Liu (2010) focus on *knowledge integration* (i.e., the ability to connect relevant ideas in explaining phenomena). Their unit of analysis is the pattern of answers to classes of questions. They find that questions about energy sources (which require attention to only one form of energy), questions about energy transformation (which require links between two forms of energy), and questions about energy conservation (which require links between all the energy transformations in a system) present increasing levels of difficulty for middle school students.

Variations in units of analysis and in the types of phenomena that form the domains of these learning progressions on energy make it difficult to use these different learning progressions to inform standards development. Greater coordination in learning progression researchers would both enrich our understanding of science learning and facilitate the work of standards developers.

Challenge 3: Establishing the upper anchor—what students should achieve. The process of formulating state standards for high school students is partly sociopolitical and partly empirical. The process requires consideration of which scientific ideas (e.g., global warming, endangered species) high school graduates should understand if they are to make informed political and personal decisions; this is the sociopolitical aspect of the process. The process also requires determining if high school students can reach those ideas, and, if so, the level of sophistication of those ideas; this is the empirical aspect of the process.

Since the 1980s, sociopolitical considerations have appeared in discussions about scientific literacy (e.g., AAAS, 1990; Hazen & Trefil, 1991; NRC, 1996) and, more recently, in discussions of career and college readiness (e.g., College Board, 2009; Conley, 2005). There is now some consensus about the scientific knowledge students should have when they graduate from high school. For example, see the commonalities in *Benchmarks* (AAAS, 1990), *National Science Education Standards* (NRC, 1996), *Science Matters* (Hazen & Trefil, 1991), and the College Board (2009) standards. These commonalities are generally also found in the states' science standards.

While there is general sociopolitical consensus on the scientific ideas high school graduates should have learned, there is limited empirical evidence that shows they can realistically master those ideas in their PreK-12 education. Many scientific theories and models, as understood and used by scientists, are too advanced for students with a PreK-12 education, particularly given the limited amount of classroom time usually devoted to each science topic. For pragmatic and well as cognitive reasons, it may not be realistic to expect students to master so many complex scientific ideas.

While the upper anchors of learning progressions may ultimately inform high school standards, few current learning progressions include a high school component or account for sociopolitical considerations.[5] For now, standards continue to be defined from a primarily sociopolitical perspective. However, learning progression research has the potential to bring an empirical perspective to this issue to better inform states about the ideas students can master at different grade levels. Early learning progression research suggests, at least in some domains, that students can learn more sophisticated concepts and practices than typically supposed (e.g., Metz, 2011). In contrast, other research suggests how difficult it is to help students move toward ideas compatible with scientific theories (Smith et al., 1997).

Challenge 4: Defining core ideas and representing them in standards. Various educational researchers have different criteria for what constitutes a core idea (NRC, 2009). For some, core ideas reflect major theoretical tenets (e.g., the conservation of energy, the atomic structure of matter). For others, core ideas articulate the major aspects of the explanatory power of theories (e.g., how the interaction of genes and the environment causes physical and behavioral traits; how the structure of matter accounts for its properties and behaviors). Another consideration is whether core ideas should be about specific classes of objects or phenomena in a domain, for example, the relationship of form and function in the respiratory system and the relationship of form and function in the circulatory system [MADOE, 2010]) or whether core ideas are more abstract and should apply across a whole domain (e.g., the relation between structure and function at different levels of biological systems [Michael, Modell, McFarland, & Cliff, 2009]). While it is clear that the two views are not antithetical and that abstract ideas have to be approached first in specific contexts, privileging content areas over abstract principles leads to learning progressions, and ultimately to standards, that are structured differently. This difference in structure can be particularly vexing when learning progressions focused on the same core idea take different approaches. Without some consistency or common agreement on what constitutes a core idea, it is difficult to for the states to develop coherent standards based on learning progressions.

Capturing the coherence that core ideas may provide across grades and grade bands is also a challenge for standards developers. Standards typically delineate specific concepts and practices that can be learned in a year or two. But core ideas are broad—they are composed of multiple concepts and practices that are learned over many years. In addition, the concepts and practices that comprise core ideas develop in interaction with each other—the meaning of one component depends on the meaning of others. To do justice to a core idea, standards expressing the different components of a core idea should be related explicitly to each other so that educators appreciate how students' knowledge can be meaningfully constructed over time. This is particularly important because core ideas are often expressed in the form they take in the upper anchor (i.e., as important scientific ideas expressed scientifically). The precursors to these core ideas (i.e., less complex, less abstract, less general versions) in earlier grades may be unrecognized if standards from a particular grade span are viewed in isolation. Most curriculum developers and teachers do not refer to a complete set of standards – vertically or horizontally – when planning lessons or curricula. The temporal continuity across grades from earliest precursors to intermediate components of a core idea to its scientific version is an essential aspect of a learning progression approach. Core ideas are more than the sum of their parts—the connections between their parts and the way they evolve over time are inherent. It is a significant challenge to represent this temporal continuity in standards.

Challenge of Integrating Knowledge and Practice Elements in Standards

A recent goal of standards developers is to express standards as performance expectations that integrate scientific content and practices. Science standards should involve the integration of scientific content and practices because understanding scientific theories is inseparable from understanding how scientific knowledge is generated (NRC, 2007). Declarative statements focused on content alone do not capture what it means to understand a topic. Learning science is learning to use one's understanding in particular ways. The integration of scientific content and practices, consistent with how scientists apply their knowledge, is not typically reflected in current science standards. Science standards have long presented inquiry skills separately from content knowledge (e.g., AAAS, 1993; NRC, 1996). One rationale for this separation was that inquiry is a process for learning content; therefore any (or all) inquiry skills could be matched with any content in teaching and learning (AAAS, 1993; NRC, 1996). The curriculum designer or teacher could pair skills and content based on which skill was appropriate for which grade. Presenting content knowledge as separate from inquiry, even with the expectation that they will be combined in curricula, establishes that content knowledge and inquiry are independent. This organization is antithetical to the very nature of scientific knowledge. This is an area in which learning progression research has just begun to make significant progress.

Two important themes in recent learning progression research are that learning science includes developing content knowledge and scientific practices, and that scientific practices may have different levels of epistemological sophistication (Smith, Wiser, Anderson, & Krajcik, 2006). For example, reconceptualizing weight as an extensive quantity in early elementary school depends on understanding principles of measurement as well as privileging measuring weight with a scale over hefting (i.e., learning not to trust the unaided senses). In later schooling, reconceptualizing *force* so that students understand that trees can pull on ropes and that walls can push on bodies requires a relatively sophisticated understanding of the relationships between theoretical entities (e.g., force) and our perceptions. These examples illustrate how scientific ideas are integrally linked to the epistemological beliefs in scientific practices. Lehrer and Schauble's (2006) seminal research, which argues for integrating modeling and communication into science activities from kindergarten forward, calls attention to these broad ideas.

This recent learning progression research may be used to combine science content with scientific practices that provide the epistemological context necessary to construct specific scientific concepts and principles. This combination would achieve the three goals of making scientific concepts and principles meaningful, of providing students with scientific practices that are adapted to both their cognitive resources and the content they are learning, and of making scientific practices meaningful and purposeful. However, using learning progressions to guide the development of standards that accomplish these goals presents a significant challenge.

Challenge 5: Until learning progression research advances to the point that it provides clear and systematic insights that match content and practice, there is little guidance for specific knowledge-practice integration in standards. Some current learning progressions systematically match content and practice (e.g., Gotwals, Songer, & Bullard, this volume) or address epistemological questions (e.g., Schwarz, Reiser, Acher, Kenyon, & Fortus, this volume). Research such as this is an important start. However, much more research is needed, particularly research in which the findings are specific enough for standards development. In the meantime, the literature on students' content-related ideas and beliefs, on the one hand, and their epistemological understanding and mastery of scientific practices, on the other, can be used to generate standards that integrate scientific content and practices.

The College Board (2009) has recently taken a bold step in its articulation of Science College Readiness Standards™ as performance expectations that match particular scientific practices with essential knowledge statements. Although these new standards do not rely on a learning progression approach per se, the researchers and teachers who wrote them relied on premises similar to those advocated here. The standards use a cognitive framework that is organized around big ideas, that treats practices and content are inseperable, and recommends combining particular practices in a particular domain at a particular grade range. Some practices, such as asking a cogent, scientific question about a topic, may combine productively with any content at any grade level. As an example:

> LSM-PE.3.1.6a Formulate a scientific question that addresses the relationship between the number of organisms in a population and the physical (abiotic) factors of their environment. (College Board, 2009, p. 66)

In contrast, running a simulation may be relevant in a restricted number of domains only and may be inappropriate for all grade levels. As an example:

> LSH-PE.1.2.1 Construct a model or run a simulation that represents natural selection in terms of how changes in environmental conditions can result in selective pressure on a population of organisms. (College Board, 2009, p. 55)

The two practices of "construct a model" and "run a simulation" can be applied across topics, but the types of questions or simulations may be quite different depending upon the specific content. Thus practices have specific meaning when connected to the particular content and are especially useful to student learning of particular content. The College Board's work on the Science College Readiness Standards™ shows it is unnecessary to wait for learning progressions to specify the development of integrated, practice-content standards. However, once more learning progressions that integrate content and practices are developed and tested empirically, standards developers will have evidence that supports their work.

Developing Concept and Skill Progressions to Inform Massachusetts' Standards Revision

In its attempt to use learning progression research to inform standards revision, Massachusetts used a tool called a *concept and skill progression*. This tool was developed to address the challenges (as described in this chapter) as well as a somewhat more practical concern. In particular, a significant problem in efforts by the states to ensure that standards revisions are based on empirical evidence is that no department of education in any state has the resources or expertise to locate, access, and analyze the relevant primary source literature, particularly within the time frame established for the states' standards development (generally, about two years). Therefore, Massachusetts asked educational researchers to develop summaries of this research for particular science topics in order to identify and organize the development of key concepts and skills in different domains. The researchers were given a generic template for this purpose (see Figure 1).[6] They were asked to use learning progression research where possible. In instances where there was no learning progression research for the components of the template and/or grade ranges, they were asked to offer an initial hypothesis based on other research, including pre/misconceptions and conceptual change research and insights from their own work.[7] Another education researcher reviewed each summary, and revisions were made as needed. Concept and skill progressions were written for 17 science topics.[8] These summaries use learning progression research, but they are not learning progressions in the research sense. Massachusetts refers to the summaries as concept and skill progressions (MADOE, 2010). It is possible to view these summaries as the initial step in the ongoing cycles of research and curriculum development required to articulate a fully validated learning progression.[9]

Each concept and skill progression includes both *narrative storyline* and *concept and skill details* sections to convey how students' conceptual growth occurs over time (see Figure 1); both sections tell the same story but at different levels of specification. The concept and skill progressions reflect the work of Wiser and Smith (2009) in which a learning progression is described as

> mak[ing] use of data on students' ideas and performances as a function of instruction to say "on the basis of extensive and intensive cognitive analyses of patterns of beliefs within students exposed to curriculum X, and between students exposed to different curriculum, and on the basis of our theoretical approach to conceptual development, here is a way knowledge could evolve" (p. 8).

The structure of the concept and skill progression template captures several key features of learning progressions. The template proposes a *lower ancho*r that takes into account young children's initial ideas and an *upper anchor* that articulates the concepts and skills students should achieve through their K-12 education. The template organizes concepts and skills around *core ideas* that provide coherence within and across grade bands and outlines a series *of stepping stones*[10] that could link the lower anchor to the upper anchor. The stepping stones have key conceptual shifts that a student needs to make in a particular grade band as a result of instruction and thus represent the targeted concepts and skills at the end of the grade band. The knowledge students achieve when they reach a stepping stone is: 1) conceptually closer to the upper anchor than knowledge in the previous stepping stone; 2) coherent and meaningful; and 3) productive in that it gives students the conceptual tools to make progress toward the next stepping stone. The concept and skill progressions do not, however, qualify as bona fide learning progressions because they are typically hypothetical (except for the portions based on learning progression research).

In addition, relying on one or two authors to develop each concept and skill progression is somewhat problematic as the process limits the number of perspectives represented. For this reason, and because of their hypothetical nature, the concept and skill progressions should be viewed as stopgap measures. However, they provide a fairly efficient way for researchers specializing in a particular science topic to compile information on how students learn and to develop an informed hypothesis about how instruction may be structured and sequenced so that students progress toward scientific understanding. This makes concept and skill progressions useful in the relatively short timeline of the states' standards development process.

The following sub-sections use the topic of genetics to illustrate the development and use of a concept and skill progression. This topic was part of Massachusetts's science standards revision process. Genetics is a topic with a solid pre/misconception research base, an emerging consensus on core ideas that are important to scientific literacy, and some learning progression research. Aaron Rogat of Teachers College, Columbia University, wrote the genetics concept and skill progression.[11]

Addressing Challenges 1 Through 4

In developing concept and skill progressions, authors must begin by defining the core ideas (Challenge 4). These have to be established first because they structure the progression: they are traced across stepping stones toward the upper anchor. Moreover, the specific concepts and skills in the progression and their interrelationship across time depend on the evolution of the core ideas. Massachusetts relies on the authors to identify the core ideas in the domain assigned to them (i.e., to decide how to parse the scientific theory for that domain into a few main ideas with strong explanatory power that can effectively structure the progression).[12] The genetics concept and skill progression is structured around five core ideas: (1) organization, location, and function of DNA and genes; (2) relationships between genes, nuclear division, and passage of genetic information; (3) effects of DNA mutations and variation; (4) gene variation and implications for phenotypic variation; and (5) transmission of genetic information and patterns of inheritance. Rogat described the underlying basis for these five core ideas as follows (A. Rogat, personal communication, June 4, 2010):

> I relied on three sources of insight in articulating the five core ideas of genetics: 1) an earlier paper that Ravit Duncan and I worked on where we identified big ideas in genetics [Duncan, Rogat, & Yarden, 2009]; 2) my own expert content knowledge in molecular biology and genetics since often I reflected on the ideas that I used as a scientist to understand the phenomena, investigations, and data generated by others; and 3) the existing MA standards and national science standard. (I attempted to relate what was in the MA standards with the theoretical work that Ravit and I did).
>
> The work that Ravit and I initially did was informed by research on naive conceptions in genetics as well some research into interventions in this

> discipline. Ravit and I realized that there are several studies that show many students fail to connect genes to proteins and phenotypes in productive ways and yet this connection is foundational to how modern day molecular biologists and geneticists think about biological phenomena. So we knew this was a core idea to focus on and we made it explicit in our learning progression. In the summary I prepared for MA it is most explicit in the core idea that focuses on the organization, location, and function of DNA and genes.

A similar approach was applied to the other four core ideas. Throughout this process, Rogat maintained a focus on several themes in identifying the core ideas of the genetics concept and skill progression. Rogat continued:

> I think there are three other themes that characterize our thinking: 1) We [Duncan and Rogat] wanted to identify ideas that held great explanatory power within the discipline and that would help to explain or predict a number of the natural phenomena in biology that ordinary citizens might encounter in their personal lives or in the news (e.g., disease, biotechnologies, new advances in genetics). 2) Many core ideas are actually a set of smaller ideas or components that together comprise a more complete picture of the mechanisms underlying the genetic phenomena. We notice that several research studies show that many students develop incomplete pictures of these mechanisms, partly because the pictures are fragmented and put into different section in curriculum (e.g., deep discussion of genes and proteins) and in some cases the important ideas are lost or deemphasized, so we developed core ideas that would address some of these problems and made those ideas more explicit and more complete. 3) We are more interested in deep conceptual understanding rather than surface understanding, so we focus on more meaningful, or big picture, ideas rather than more fragmentary and fine grained ideas.

In their earlier work, Duncan and Rogat addressed Challenges 2 and 4 in order to identify and agree upon the core ideas of genetics. Rogat then applied those core ideas to the development of the genetics concept and skill progression, taking into account a few additional factors noted in a standards context.

In the discussion of Challenge 4 above it was noted that representing these core ideas in standards is difficult because the standards are written at the level of particular concepts and skills. This level of detail may be antithetical to the nature of core ideas, which consist of a network of concepts and beliefs that take their meaning from each other. Keeping in mind that the progression has to be translated into individual statements, the template given to the authors required that they group concepts and skills together, with each row representing one core idea. This grouping implies that corresponding standards can also be grouped into a set that represents the components of a core idea, with an overarching explanatory statement that articulates their relationship. While this statement would not be part of the standards themselves, it would make clear how the concepts and skills comprise a core idea. The use of a core idea is a typical structure in many current state standards that group concepts by topic, but the explanatory text would focus on how a set of standards relate and contribute to the development of a core idea. This is certainly not a perfect solution, as discussed in Challenge 4, but it is a realistic beginning.

Once the core ideas were identified, the authors then determined how they develop over time, representing that progression from column to column of the concept and skill progression. To define the upper anchor of the genetics concept and skill progression (Challenge 3), Rogat had to specify the sophistication level vis-à-vis core ideas that graduating high school students should have, the concepts and skills in the core ideas at that level, and how students could use the core ideas. In addressing Challenge 3, the template required that Rogat (and other progression authors) consider the upper anchor from the perspective of research and of core ideas and also to consider the current Massachusetts high school science standards (with the understanding that those standards were originally developed, at least in part, with reference to national documents as well as local sociopolitical considerations). The authors' recognition of these constraints ensured respect for the concepts already included in the Massachusetts standards. At the same time, it was clear that research might inform significant changes intended to increase the likelihood of students' learning of targeted concepts.

As expressed in the narrative storyline, the final version of the upper anchor of the genetics concept and skill progression reads:

> *High school* students understand the relationship of genes to phenotypes. Students can explain how events during cell division are important in explaining why we see certain gene combinations in predictable patterns. They understand only one copy of a trait needs to be present to show a dominant phenotype or both copies present to show a recessive phenotype. They can predict possible combinations of alleles (and potential phenotypes) for the progeny of two parents with a given set of alleles. Students can relate DNA duplication and nuclear division to the passage of DNA and consequently inherited traits. They understand that there are two types of division that cells can undergo (mitosis and meioses) and that these occur in different cell types and result in different end products in terms of the number of chromosomes or genes produced. Students recognize that all organisms have DNA and genes. Students can explain the role and function of genes in living organisms. They can relate and distinguish between chromosomes, genes, DNA, and nucleotides in animal and plant cells. They understand the order of nucleotides in a gene determine the structure of a protein, and consequently the function of a protein in cells. Students can explain how a mutation in a gene can affect the structure, function, or behavior of a cell or an organism by influencing the structure or function of proteins.

A number of these concepts exist in the current Massachusetts high school standards for biology. However, Rogat expands on those standards in other sections of the genetics concept and skill progression. In particular, there are new expectations in his genetics concept and skill progression: high school students should be able to differentiate genetic structures in animal and plant cells, and should be able identify the structure, function, and the role of proteins in cells. These new key concepts in the core ideas of genetics make the concepts in the current standards more meaningful.

Initially, the upper anchor was not fully developed to the extent just presented. Having sense of the upper anchor helped articulate the lower anchor and stepping stones. Refinement of the upper and lower anchors and stepping stones was an iterative process. Ultimately, the lower anchor was defined to emphasize that students often come to initial instruction in genetics (which may start in early elementary school) with a number of concepts that can be productively built upon, as well as a number of misconceptions. The text below is excerpted from the lower anchor (in the narrative storyline) of the genetics concept and skill progression:

> *Before instruction* students typically have a theory of inherited kinship; with this they can distinguish some inherited characteristics versus socially determined characteristics. They are, however, likely to believe that daughters get their characteristics from mothers, and likewise sons from fathers. Students can identify some critical aspects of organisms required for living, including the ability to reproduce. Students are also likely to understand that all living things share some of these things in common, as well as physical attributes. They are unlikely, however, to attribute those to any common mechanism such as DNA or genes, even though they are likely to have heard about genes and DNA.

To formulate the lower anchor, Rogat reviewed the research on young children's ideas about genetics. The purpose of this review was to identify prevalent beliefs that could be productively built upon. He read the empirical evidence and researchers' hypotheses about the effects of children's beliefs on learning; where evidence was lacking, he used his own judgment. This process, along with the subsequent development of the stepping stones, required Rogat to piece together information from the research and to use educated guesses where research was unavailable.

Concept and skill progressions attempt to explain how learners can make the transition from initial ideas to more scientifically accurate understandings (assuming appropriate instruction is provided) of a core idea. This transition requires descriptions of intermediate productive ideas (or stepping stones)—those ideas that are not as sophisticated or complex but are conceptually closer to the upper anchor than the ideas in the lower anchor. For example, the genetics concept and skill progression specifies stepping stones for the end of elementary school, early middle school, and late middle school that are conceptually related over time.[13]

> *Elementary school* students understand that siblings do not always look identical to each other or their parents, but have a combination of characteristics from their parents. They can apply this to people, animals, and even reptiles or insects.[14] They may, however, believe that plants are not living and therefore do not have genes. Students can identify critical aspects

of organisms needed to live, and know that genes somewhere inside living organisms provide information about an organism's development. Students can explain that "information" in genes about how organisms look provide are passed on from one generation to another. They do not, however, understand how this works.

Early middle school students understand that genes are linked to a theory of kinship. They know that traits are physical characteristics of organisms that are influenced by genes (e.g. "a gene for eye color" is not actually an eye color, but has information about eye color). Students understand that mutations are changes in genetic information that can confer "different" traits or functions. Students may associate reproduction with copulation and may believe it only occurs in animals. Students are able to explain that genes are inside cells, but may believe they are in only a few cell types. Students understand that genes provide information about the development of traits or cellular entities, but likely believe genes are active "particles."

Late middle school students understand the mechanism of how traits are passed between generations and relate kinship to genes. Students understand that genes are found in cells of all organisms and associate genes with chromosomes. They can explain that chromosomes carry genetic information from generation to generation during cell division. Students understand that for each trait we have two version of the gene (alleles). They understand that duplication of chromosomes occurs before cell division to maintain an equal amount of genetic information. They know there are two types of cell division: mitosis and meiosis, but may not properly connect these to specific cell types. Students realize egg and sperm fuse in sexual reproduction to produce a new cell that goes on to develop as the offspring. Students understand the role of genes in transmission of information and in influencing proteins and cell function. They understand that a mutation in genes can result in a change in proteins or cell function.

The three stepping stones in the genetics concept and skill progression have limited empirical evidence, particularly for genomes, mutations, and patterns of inheritance. In his document on the learning progression, Rogat identifies the content based on his informed hypotheses and the content based on empirical evidence. This identification is necessary due to the lack of a full PreK-12 progression for genetics, as discussed in Challenge 1. Rogat gives citations for concepts and skills that have a clear evidentiary base; he notes where he uses his informed judgment.

As with the development of the upper anchor, Rogat was asked to create stepping stones taking into account the current Massachusetts standards and research evidence. He was also asked to use his informed judgment, especially in situations requiring justification of significant deviations from the evidence. Rogat focused on research about concepts that students may find difficult. He also examined research about the traditional sequencing of concepts and/or typical grade level(s), in particular evidence about sequences in which an introduced concept was not understood and a different approach was supported. For example, the concept of proteins is not in Massachusetts's current standards. However, this concept is in the genetics concept and skill progression because there is evidence of its importance to core ideas of genetics and to students' ability to explain genetic phenomena. Additionally, based on research evidence, each stepping stone integrates concepts that differ from concepts in the same grade span in Massachusetts's current standards. For example, the current middle school standards in Massachusetts include the idea that genes influence traits but defer the idea on the role of mutations to the high school standards. In Rogat's genetics concept and skill progression both ideas are included in the middle school standards. Again, while not a perfect solution, it allows authors of concept and skill progressions to use research literature and their expert judgment to help Massachusetts make informed decisions about standards revision.

This broad overview of the genetics concept and skill progression is useful as an illustration of how Massachusetts worked with authors to address the first four challenges identified above. Rogat was able to draw on several previous learning progressions. Additionally, in building his genetics concept and skill progression, he referred to additional research to fill gaps in the K-12 range, made and justified decisions about the concepts and skills for particular grade bands, reconciled the sociopolitical aspects of the current Massachusetts high school standards with the empirical evidence on what is learnable, and referenced research evidence or convincing personal hypotheses. Ultimately this procedure will help support the changes to the Massachusetts science standards.

Addressing Challenge 5

The current Massachusetts standards are worded as performance expectations. Therefore, they use verbs that are typically associated with a cognitive taxonomy, such as Blooms' taxonomy (Anderson & Krathwohl, 2001) and that are matched with content. Together the standards articulate what students should be able to do with their knowledge. In its standards revision process, following the College Board's (2009) model and earlier discussions on the importance of content, scientific practice, and scientific epistemological elements in learning progressions, Massachusetts expresses performance expectations in the context of scientific practices. Thus fewer generic, cognitive verbs are used; more verbs from scientific practices are used. Matching practices with content, however, cannot be arbitrary.

In some cases, a concept and skill progression already contains the content-practice link. This suggests that the progression is likely to support learning. For example, a draft standard in Massachusetts states: "Analyze genetic data (from a pedigree or using a Punnett Square) to determine the effect of dominant and recessive alleles on the phenotype of an organism." This standard is closely associated with a similar statement from the genetics concept and skill progression. There is a deliberate match between the analysis of data and a key genetics concept. Practices such as analyzing data appear fairly consistently in concept and skill progressions; it is not difficult to match them with relevant content.

Other practices, such as “use evidence to support a claim” are not always included in concept and skill progressions. In these cases, the concept and skill progressions offer little substantive justification or guidance for matching any given practice with particular content. Indeed, a range of content might be used to instantiate the use of evidence to support a claim, resulting in a standard such as “Use evidence from genetics, such as variation in characteristics across generations, to support a claim about the different gene combinations resulting from sexual and asexual reproduction”. Such standards specify both the content of the claim and the type of evidence that should support the claim. Because these content-practice matches were not included in all concept and skill progressions, careful consideration must be given to these matches to ensure that they help students learn both the practice and the associated content.

Massachusetts is working to ensure that scientific practices are included in a number of standards that are distributed across science domains and grade spans. The expectation is that students will have multiple opportunities to apply each practice in different contexts. It is not easy to match content and practices so that the learning expectations support students’ cognitive engagement and content learning while at the same time ensure that all practices are adequately represented.

Section Summary

The challenges posed by the early state of learning progression research are significant but not insurmountable. The development and use of concept and skill progressions represent initial solutions to these challenges. However, structuring and communicating a large amount of information about science learning, using both scientific and cognitive perspectives to support student learning, is a large task. This work poses a very significant challenge for standards developers. Ultimately, the research community is best positioned to address these challenges since researchers can develop learning progressions in more domains and more grade bands. Researchers can also come closer to consensus about the nature of learning progressions and can develop a common format for the findings of this research.

CHALLENGES BASED ON THE PURPOSE, DESIGN, AND LIMITATIONS OF STATE STANDARDS

Challenges of Representing the Complexity of Reconceptualization in State Standards

We have briefly mentioned the challenge of representing core ideas in state standards. Core ideas typically consist of concepts and skills whose content and relationships change from grade band to grade band. Thus, rather than representations of a collection of isolated concepts and skills, core ideas may be thought of as knowledge networks. However, given the typical structure and presentation of the states’ standards, knowledge networks may be difficult to describe. Therefore, a related and even more difficult challenge is to represent the reconceptualizations required in students' knowledge networks. Both the interconnected nature of the concepts and skills in learning progressions and their sometimes pre-scientific nature are difficult to include in the states' standards. In addition, traditional standards present content and practices separately. Therefore, in explaining new types of standards (which integrate content and practice) to their key constituents, standards developers face yet another challenge. We discuss these three challenges next.

Challenge 6: Standards are typically presented as numbered lists that are not conducive to representing interconnections between concepts and skills. The traditional way of representing standards as numbered lists is intrinsically incompatible with the structure of learning progressions. A key to understanding reconceptualization is the recognition that concepts and skills derive part of their meaning from their relationship to other concepts and skills. Hence reconceptualization requires working on more than one concept or skill at a time. Representing this dynamic interplay of student ideas, scientific concepts/skills, and new understandings in the states' standards is tricky. Standards may group interrelated concepts for each core idea within a grade band (horizontally) and across grade bands (vertically). However, the interconnections and sequences among standards are impossible without the addition of distracting notations. Other representations of standards that explore these relationships are needed.

Challenge 7: Stepping stones are not always scientifically accurate, but standards are typically designed to be scientifically accurate. Learning progressions consist of a series of stepping stones that progressively approach the upper anchor, support coherent and meaningful conceptualizations of the domain, and ensure that students can progress toward the upper anchor in the next grade band. "How students get there" is about reconceptualization. Stepping stones are important guides for standards developers because they can help them structure standards around ideas that are conceptually productive.

Some components of stepping stones are scientifically correct and easily incorporated into the states' standards. For example, "Any piece of material, however small, occupies space" is a step toward the upper anchor of a matter learning progression and is scientifically correct. Similarly, "If two objects at different temperatures are placed in contact, heat flows from the higher temperature object to the lower temperature object" is a scientific principle that has been proposed as part of an elementary grades stepping stone in an energy learning progression (Lacy, Wiser, & Doubler, 2010).

To be conceptually productive, however, some stepping stones are scientifically inaccurate. Both scientists and science educations will likely challenge such stepping stones. "Any piece of matter, however small, has weight" is a case in point. Wiser and Smith (2009) argue that the idea that the idea of matter having weight is a crucial conceptual step toward a scientific understanding of matter. The scientifically correct "Any piece of matter has mass" is not since young students do not understand the concept of mass. They need "Any piece of material, however small, has weight" to construct the concepts of mass and matter. For them, "matter" is solids and liquids that can be touched and seen. If they are to believe that gases are matter, based on empirical evidence, they need to understand that gases can be material and have weight like solids and liquids even though they cannot be touched or seen. Imagining gases as tiny pieces of matter, *each having weight* but too distant from each other to be detectable with the unaided senses, makes the materiality of gases more believable. Once a concept of matter is achieved ontologically, it can be quantified meaningfully. The concept of mass can emerge by combining the concepts of gravitational force, matter, and the amount of material. At that point, the statement "Any piece of matter, however small, has weight" can be revised to be "Any piece of matter has weight in a gravitational field because it has mass". This new statement is a meaningful and relatively straightforward. In short, this example illustrates that some components of stepping stones are extremely generative and bootstrap themselves into their own scientific versions, even though they are not scientifically accurate.

Standards developers are typically concerned with scientific accuracy, in part due to a concern that students may develop misconceptions or misunderstandings. Additionally, because the states do not want to present wrong ideas in their standards, they usually invite scientists and others who are concerned with scientific accuracy to review and comment on the proposed standards. Formulating potentially contentious stepping stones that respect scientific accuracy may be an important goal in standards development.

Challenge 8: Standards that integrate content and practice can be perceived as curriculum or instruction rather than as outcomes. As explained earlier, consistent with both the new thinking about standards and recent learning progression research, standards should integrate content, practices, and epistemological knowledge. If content-practice combinations are perceived to be about curriculum or instruction, however, a concern may arise that the standards recommend certain curricular or instructional materials and methods rather than stress learning outcomes. We believe that such perceptions may be due to misinterpretations about the nature of practices and their intended use in standards. For example, recent conversations with Massachusetts educators introduced them to discussion versions of selected standards, including the Grade 6–8 standard– "Construct a model to explain how the tilt of the earth and its revolution around the sun result in an uneven heating of the Earth." The educators overwhelmingly perceived the standard as suggesting students should make a diorama of the earth's tilt relative to the sun. This was not the intention of "construct a model" as a scientific practice. Such a tendency to misinterpret practices is likely related to the ambiguity of 'scientific inquiry' more generally. Scientific inquiry and many related terms are used to describe both skills students should develop and instructional or curricular methods (e.g., NRC, 2000). The lack of differentiation between the two meanings of 'inquiry' makes it difficult for many educators to differentiate learning outcomes (articulated as practices) from instructional or curricular methods. Care is needed in writing the states' standards as achievable outcomes rather than as classroom activities. The application of the standards to curricula is a key task that must always be kept in mind in standards development. The standards should be directed to the desired outcomes of educational experiences rather than to the specification of those experiences.

New Strategies for Massachusetts's Standards Revision

Addressing challenge 6. Like most states, Massachusetts has developed standards as sequentially numbered lists of concepts. Traditionally, these standards group the lists into related topics but do not reflect any other relationships between concepts. It is clear, however, that if the current revision process is to honor a learning progression perspective, another way to represent the standards and the relationship between them is needed. Thus, Massachusetts converted its current standards into strand maps[15] based on the *Atlases for Science Literacy* (AAAS, 2001, 2007). Massachusetts uses this format to frame the revision process. Strand maps help make explicit the progressions of concepts and skills across grade spans that collectively support and lead to the desired high school outcomes. As visual representations, the strand maps are very effective in allowing standards developers to see whether concepts and skills in a topic progress in a purposeful manner across grade spans. Arrows in the strand maps show the linkages between standards. The strand maps can also be arranged in different configurations that highlight how ideas and concepts relate, even across disciplines. For example, Massachusetts is developing a thematic climate strand map that shows the relationship of physical science, earth and space science, and life science concepts that contribute to a broad understanding of climate. The continued use of strand maps is likely to help Massachusetts's standards developers, as well as curricular and instructional staff, approach teaching and learning so that the goal is student reconceptualization rather than mastery of discrete ideas.

Addressing challenge 7. We have argued that the stepping stones in the concept and skill progressions include some statements that, while not scientifically accurate, are conceptually productive (e.g., "matter has weight"). One main challenge with using those statements in standards is to word them so that they do not provoke outcries from those who demand scientific accuracy. In Massachusetts we have tried (a) to phrase standards carefully so that while the science concept is inaccurate, there is little debate about its relationship to scientific understandings and (b) to place the science concept in the context of a model because models, by definition, apply only in certain contexts and are meant to be revised. Additionally, one could provide evidence for the productivity or even pedagogical necessity of an inaccurate idea in a certain grade band. It is not clear how this third approach should be represented and documented in standards. Massachusetts has not implemented that approach.

The wording of a standard is key to whether it is viewed as scientifically accurate, even if intended to express an inaccurate idea in a stepping stone. This is the key to the first solution we propose. For example, the animate/inanimate distinction, which is very salient from infancy on, has no place in physics;

however, it is important to address this distinction in developing the concept of force, since many young students believe that only animate objects exert force (Gunstone & Watts, 1985). Two important steps in constructing a scientific view of force are to understand that dynamic inanimate objects and then static inanimate objects can exert force. Standards that reflect this construction can be phrased as "Students recognize that all objects in motion can exert force when they come in contact with other objects" and "Students recognize that a static object (e.g., a table) can exert a force on another object." This wording expresses a component of a stepping stone in scientific terminology and thereby helps describe students' development of scientific ideas in a way acceptable to standards developers and the scientific community.

Our second solution recognizes that, in modeling, model developers make choices about what to represent and what to ignore. Therefore, models are never completely accurate. "All models are wrong, but some are useful" (Box, 1979). For example, the idea that "Matter has weight" may be unacceptable to scientists who make a clear distinction between mass and weight. However, "All matter has weight" is an idea that is valid in some contexts (in this case, in a gravitational field, for example, on earth) and not in others. It is acceptable in a standard to phrase the statement as "Matter on earth has weight," rather than "All matter has weight".

Addressing challenge 8. Massachusetts's current science standards emphasize content knowledge and relegate scientific skills to the introductory sections of its science framework that is separate from the standards. Thus these standards focus on the terms used for scientific practices to avoid associations with curricula and instruction as much as possible. It is also important, in standards development processes, to define clearly each scientific practice in terms of outcomes in order to reduce the possibility of misperceptions and, thus, the chance that misinterpretations will be used to remove standards during the adoption process.

That said, the use of learning progressions in standards development poses an interesting quandary. A key purpose of standards is to drive the development of curricula and instruction. Learning progressions may specify key learning experiences and/or be very specific as to the order in which concepts and relationships between concepts should be explored. If standards development is to rely on learning progressions, as we have argued in this chapter, it is possible, and perhaps inevitable, that the distinction between standards and curricula will blur. This has been a struggle in the Massachusetts standards revision process. The struggle relates to how much the concept and skill progressions should reference, but not drive, particular instructional strategies that address the cognitive needs of students. Standards are fundamentally about student learning outcomes; typically standards do not offer pedagogic recommendations. In future work on learning progressions, it is necessary to examine the interplay between maintaining the intent that standards drive curricula and recognizing that the curricula drive standards (through reliance on learning progressions).

Section Summary

In addition to challenges posed by the early state of learning progression research, there are challenges presented by the nature of standards that the states have developed. These are challenges that relate to the translation of the richness of learning progression research into fairly fixed and constrained state standards. Finding solutions to the challenges presented by the dynamic interplay of students' conceptions and scientists' conceptions requires experimentation and time. The strategies that Massachusetts is developing to respond to these challenges, by taking a learning progression perspective, while clearly not complete or problem-free, suggest a more explicit and systematic means of supporting student learning.

TOWARD STANDARDS THAT HONOR A COGNITIVE BASIS

State standards that join a scientific perspective and insights from learning progression research promise to be effective tools in science education. Learning progression research can help standards developers identify the concepts students should know (upper anchor), account for student pre-conceptions in a domain (lower anchor), and draw roadmaps that provide a more effective guide to student learning (via stepping stones).

In this chapter, we argue that the use of learning progressions to revise science standards can begin now. We offer partial solutions to the challenges posed by incorporating learning progressions in this revision and describe those challenges in the early development of learning progression research and in the history of standards development. The Massachusetts standards revision, a fairly limited process, exemplifies one way of approaching these challenges. The tensions and controversy over the use of learning progressions in the states' science standards will continue for some time.

It is also clear that the states cannot do this work alone. Standards development by the states requires that the research community produce research findings that inform their standards development. Several steps are needed: consistency in definitions and forms of learning progressions, assurance that a range of domains and grade spans are studied, and specification of the implications for curricula and learning outcomes. Communication with those who adopt, implement, and assess new science standards is also critical. For example, the National Science Foundation recently awarded a RAPID grant to the Consortium for Policy Research in Education (Rogat, 2010). This project is designed to develop hypothetical learning progressions for a few topics and to consider the formats and products that are useful for different end users. Similarly, this chapter is intended to promote discussion between and across communities such that the research and products of learning progression research are as useful and timely as possible.

If learning progression researchers and users (including the states, curriculum developers, assessment developers) explore practical solutions collaboratively, even the thornier issues may be resolved. Researchers may enrich the education community's understanding of various learning progression issues, including their usefulness and their translation into curricula and standards. Standards developers, in turn, may enrich their understanding of the merit of organizing standards around core ideas, the complexity and difficulty of making conceptual changes, the use of stepping stones, and the need to structure learning experiences in particular ways. Such a dialogue may lead to a reconceptualization of both learning progressions and science standards.

Ultimately, this dialogue may produce a model of how to translate learning progressions into standards-developer-friendly documents that transcend even important differences between the structures of different progressions. Learning progression researchers may systematically recognize important sociopolitical concerns in defining the upper anchors and the intermediate knowledge targets (e.g., the issue of global warming may influence learning progressions on energy, making energy dissipation a core idea and focusing on precursors and stepping stones for entropy). Dialogue may also contribute to resolution of the issue of using inaccurate stepping stones. Groups of researchers, standards developers, teachers, and others may agree on how to formulate standards that capture cognitively-sound stepping stones without producing lasting misconceptions. Such agreement would support standards developers' emphasis on cognitive development.

Some issues are more difficult. Basing standards on learning progressions may, for example, be perceived as making them too directive. This issue is linked to a debate in the learning progression community—how unique are learning progressions? Opinions differ widely on this issue. The opinion held depends in part on the view taken of knowledge development, including the effects of individual differences on learning trajectories or the explicit role of curricula in a learning progression. If the belief is that elements in a knowledge network strongly constrain each other, it is easier to think, "few roads lead to Rome". The fewer the roads leading to the upper anchor, the finer the grain needed to specify standards. Standards based upon learning progressions may also be more curricular-like if progress toward the upper anchor depends on specific key experiences. This issue, which is empirical, requires comparison of curricula based on different learning progressions, with different key experiences, for the same domain. A learning progression approach to state standards prioritizes student learning and meaning-making by "focus[ing] on central concepts and practices in science [and]… provid[ing] the careful scaffolding required for students to develop integrated and sophisticated understanding of science content" (Corcoran et al., 2009, p. 12). The use of learning progressions in the development/revision of state science standards may be critical in making students' ideas about science central to how the curriculum and instruction engage them in learning.

NOTES

1 The inclusion of technology/engineering as a discipline of science, equivalent to earth, life, or physical science, is a unique aspect of the Massachusetts science standards. The current Massachusetts science standards are available at www.doe.mass.edu/frameworks/scitech/1006.pdf.

2 Massachusetts began this revision process before the National Research Council and Achieve, Inc. initiated its work on the "Next Generation Science Standards" (http://www.achieve.org/next-generation-science-standards). The 2013 completion date was changed from an earlier date in order to syncronize the state's timeline with the recently proposed timeline for this national effort. Jacob Foster is the lead facilitator of Massachusetts's standards revision process; Marianne Wiser was an advisor to Massachusetts's standards revision process.

3 The "lack of mastery" perspective may lead to the articulation of outcomes or expectations for the progression (a "negative" research base can be turned into a "positive" element of a progression). What students do not know can be as informative as their misconceptions in a science domain. Such lack of knowledge may signal conceptual blocks that can be accounted for in analyses of other ideas that students do/do not have in that domain. In other words, lack of knowledge about a concept or principle is part of the data used in conducting conceptual pattern analyses that reveal students' knowledge network in a domain.

4 For example, a unit of analysis could be the type of explanation a student offers for a class of phenomena (such as carbon-transforming processes). When asked to explain any given phenomena in the class (e.g., photosynthesis), the student should understand the request and offer the same account for all phenomena in that class.

5 A notable exception is the research conducted by Anderson and his colleagues (e.g., Mohan, Chen, & Anderson, 2009).

6 The researchers volunteers for this work.

7 Each concept and skill progression attempts to explicitly identify which elements or components of the concept and skill progression are based on learning progression research and which elements or components are more speculative.

8 Concept and skill progressions were developed for the following topics: Earth processes; Plate tectonics; Earth in the solar system; Anatomy & physiology; Cell biology & biochemistry; Genetics; Evolution & biodiversity; Ecology; Force & motion; Conservation & transformation of energy; Matter & its transformations; Atomic structure & periodicity; Chemical bonding & reactions; Solution chemistry; Engineering design; Manufacturing technologies; and Materials. These are available at www.doe.mass.edu/omste/ste/default.html. The 17 concept and skill progressions are related to topics in the current Massachusetts standards, although a concept and skill progression was not developed for all topics covered in the standards.

9 Several authors referred to an initial, research-based learning progression as a "hypothetical learning progression" (e.g., Stevens et al., 2010, p. 687) or a "prototype learning progression" (NRC, 2010, p. 7-1).

10 The term "stepping stones" refers to 'qualitatively different levels' of student understanding or thinking akin to "levels" that other learning progression researchers refer to.

11 The full genetics concept and skill progression is too large to include in this chapter. It is available at www.doe.mass.edu/omste/ste/default.html.

12 The current Massachusetts standards are organized by topic and do not articulate core ideas.

13 Most concept and skill progressions maintain the typical grade span structure of grades K-2, 3-5, and 6-8.

14 While this statement is not scientifically accurate (people, reptiles, and insects are animals), it reflects how elementary school students are likely to classify living things (e.g., Driver, Squires, Rushworth, & Wood-Robinson, 1994).

15 The strand maps of the state current science standards can be accessed at: www.doe.mass.edu/omste/maps/default.html.

REFERENCES

American Association for the Advancement of Science. (1990). *Science for all Americans*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2001, 2007). *Atlas of science literacy* (Vols. 1-2). Washington, DC: Author.

Anderson, L., & Krathwohl, D. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Allyn and Bacon.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201-236). New York, NY: Academic Press.

College Board. (2009). *Science College Board Standards for College Success*™. New York, NY: Author.

Conley, D. T. (2005). College knowledge: What it really takes for students to succeed and what we can do to get them ready. San Francisco, CA: Jossey-Bass.

Corcoran, T., Mosher, F. A., & Rogat, A. D. (2009, May). *Learning progressions in science: An evidence-based approach to reform*. (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Driver, R., Guesne, E., & Tiberghien, A. (1985). *Children's ideas in science*. Milton Keynes, UK: Open University Press.

Driver, D., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). Living things. In *Making sense of secondary science: Research into children's ideas* (pp. 17–26). London, UK: Routledge.

Duncan, R. D., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th-10th grades. *Journal of Research in Science Teaching, 46*, 655–674.

Gunstone, R., & Watts, M. (1985). Force and motion. In R. Driver, E. Guesne, & A. Tiberghien (Eds.), *Children's ideas in science* (pp. 85-104). Milton Keynes, UK: Open University Press.

Hazen, R. M., & Trefil, J. (1991). *Science matters: Achieving scientific literacy*. New York, NY: Doubleday.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher 30*, 141–58.

International Association for the Evaluation of Educational Achievement. (2007). *Trends in International Mathematics and Science Study (TIMSS) 2007 assessment frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

International Technology Education Association. (2000). *Standards for technological literacy: Content for the study of technology.* Reston, VA: Author.

Jin, H., & Anderson, C. W. (2007, July). *Developing a learning progression for energy in socio-ecological systems*. Paper presented at the Knowledge Sharing Institute of the Center for Curriculum Studies in Science, Washington, DC. Retrieved from Environmental Literacy website: http://edr1.educ.msu.edu/environmentallit/publicsite/files/Energy/En%20ResearchPaper/Energy Paper.doc

Lacy, S., Wiser, M., & Doubler, S. (2010). *Rethinking how to teach energy: Laying the foundations in elementary school.* National Science Foundation award #1020013.

Lee, H.-S. & Liu, O. L. (2010). Assessing learning progression of energy concepts across middle school grades: The knowledge integration perspective. *Science Education, 94*, 665-688.

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 371-388). New York, NY: Cambridge University Press.

Massachusetts Department of Education. (2006). *Massachusetts science and technology/engineering curriculum framework*. Retrieved from www.doe.mass.edu/frameworks/scitech/1006.pdf

Massachusetts Department of Education. (2010). *Science and technology/engineering concept and skill progresssions*. Retrieved from www.doe.mass.edu/omste/ste/default.html

Metz, K. E. (2011). Disentangling robust developmental constraints from the instructionally mutable: Young children's epistemic reasoning about a study of their own design. *Journal of the Learning Sciences, 20*, 1–50.

Michael, J., Modell, H., McFarland, J. & Cliff, W. (2009). The "core principles" of physiology: What should students understand? *Advances in Physiology Education, 22*, 10–16.

Minstrell, J. (1991). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110–128). Kiel, Germany: IPN/Institute for Science Education.

Mohan, L., Chen, J., & Anderson, C. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

National Assessment Governing Board. (2008, September). *Science framework for the 2009 National Assessment of Educational Progress.* Retrieved from http://www.nagb.org/publications/frameworks/science-09.pdf

National Research Council. (1996). *National science education standards*. Washington, D.C. National Academy Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.

National Research Council. (2009). *Organizing K-12 science education around core ideas.* Retrieved from the National Academies website: www7.nationalacademies.org/bose/Core_Ideas_Homepage.htm

National Research Council. (2010, July). *A framework for science education: Preliminary public draft*. Retrieved from the National Academies website: http://www7.nationalacademies.org/bose/Standards_Framework_Homepage.html

Novak, J. D. (1993). *Proceedings of the third international seminar: Misconceptions and educational strategies in science and mathematics. Vol. III*. Ithaca, NY: Department of Education, Cornell University.

Pfundt, H., & Duit, R. (1994). *Bibliography: Students' alternative frameworks and science education* (4th ed.). Kiel, Germany: IPN/Institute for Science Education.

Rogat, A. (2010). *Developing learning progressions in support of the implementation of new science standards: A RAPID workshop series proposal*. National Science Foundation award #1051144.

Smith, C., Maclin, D., Grosslight, L., & Davis, H. (1997). Teaching for understanding: A comparison of two approaches to teaching students about matter and density. *Cognition and Instruction*, *15*, 317–393.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

Wiser, M., & Smith, C. L. (2008). Learning and teaching about matter in grades K-8: When should the atomic-molecular theory be introduced? In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 205-239). New York: Routledge.

Wiser, M., & Smith, C. L. (2009, August). *How does cognitive development inform the choice of core ideas in the physical sciences?* Commissioned paper presented at the National Research Council Conference on Core Ideas in Science, Washington, DC. Retrieved from the National Academies website: http://www7.nationalacademies.org/bose/Wiser_Smith_CommissionedPaper.pdf

Wiser, M., Tobin, R., & Lacy, S. (2011). *Looking at physical phenomena through the energy lens: A proposal for structuring an energy learning progression*. Manuscript in preparation.

AMELIA WENK GOTWALS

LEARNING PROGRESSIONS FOR MULTIPLE PURPOSES

Challenges in Using Learning Progressions

The using strand addresses the many different ways learning progressions are expected to impact science education. Learning progressions have been proposed as frameworks that can inform the development of standards, large-scale assessments, classroom assessments, curriculum development, and teacher preparation/ professional development. By providing a common framework, learning progressions have the potential to bring coherence to these multiple facets of science education (National Research Council [NRC], 2007). Thus the use of learning progressions may significantly influence other strands of learning progression work. When making decisions in these other strands, researchers and educators necessarily consider how learning progressions will be used. Similarly, learning progressions that are translated into products and tools for particular uses draw on theories and empirical research about defining, assessing, and modeling learning progressions. More specifically, questions that tie the other learning progression strands to the using strand are: How do we define learning progressions for particular uses? Which types of learning progression-based assessments are most appropriate for such uses? Which measurement models provide the best information for stakeholders when they make decisions? The authors of the chapters in the using strand focus on how learning progressions are used for specific purposes. They discuss the challenges they faced and the decisions they made in addressing these challenges.

Wiser, Smith, and Doubler (chapter 16) describe the challenges they encountered in developing a curriculum based on a learning progression for matter and the design decisions they made in addressing these challenges. They describe the iterative process of elaborating a section of a larger learning progression and designing curricula based on this elaborated piece. This process required the following actions: (1) choosing a core concept that is both generative for students and critical to scientists' understandings in a domain; (2) identifying stepping stones (intermediate, coherent sets of ideas) to use as learning targets for specific grade levels; (3) identifying lever concepts (core ideas that require reconceptualization and promote movement from one stepping stone to the next) that drive the design of activities; and (4) identifying key representations, mental models, and symbolic tools that support reasoning and conceptual change for integration into curricula.

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 461–472.

In their chapter, Furtak, Thompson, Braaten, and Windschitl (chapter 17) describe two types of learning progressions—learning progressions for students and learning progressions for teachers—and illustrate how both types can promote ambitious teaching practices. The Model-Based Inquiry (MBI) project involves the development of a teacher learning progression that presents increasingly sophisticated teaching practices. The authors explain how teacher candidates and novice teachers use the learning progression to describe and improve their practice. In the Daphne project, which the chapter also describes, a group of teachers refined and used a learning progression for evolution as a way to inform their formative-assessment practices. Their chapter describes the design decisions that both projects made in response to the following three challenges: (1) creating a format for learning progressions that are instructionally useful, (2) developing the learning progression as a vision tool that teachers from many different backgrounds can use, and (3) developing a system of tools that supports teachers in the use of the learning progression.

In their chapter, Foster and Wiser (chapter 18) describe how the State of Massachusetts uses current research on learning progressions in the design of its new, state science standards. The authors discuss two main challenges: (1) working with a learning progression research base that is incomplete and that operationalizes the learning progression construct in different ways and (2) working within the confines of typical standards documents. Specifically, they explain that the format of current standards documents is not always conducive to illustrating connections between topics or including objectives that integrate content and practices. The authors describe the design decisions they made in their creation of the standards and offer suggestions for how future research on learning progressions could benefit the revisions of standards.

In the following sections I discuss three major challenges that are raised by the chapters in the using strand and that were discussed at the Learning Progressions in Science (LeaPS) conference. I highlight the similarities and differences in the approaches to these challenges based on the proposed use of the learning progression. The first challenge arises when developing learning progressions for specific uses—in particular, deciding how learning progressions should "look" in a particular context. The second challenge arises when translating and making use of the misconceptions or inaccurate ideas that are included in the lower anchor and middle levels of learning progressions. The third challenge arises when identifying the research needed to make learning progressions useful for specific purposes.

DEVELOPING LEARNING PROGRESSIONS FOR SPECIFIC USES

Given the myriad of proposed uses for learning progressions, it is important to examine what learning progressions must "look like" to be useful frameworks in a particular context. It is possible that a learning progression that works well for one application may not be easily translated to a different context or used for a different purpose. In fact, because of the many potential uses of learning progressions (e.g., standards, large-scale assessments, classroom assessments, curriculum development, teacher preparation/ professional development), it is doubtful that a single learning progression can meet the needs of all users.

Learning progressions may have different purposes and hence different characteristics. For example, in order to inform standards, some learning progressions must cover a broad range of content. Therefore they must describe how core ideas develop over long time frames (e.g., multiple years or even K-12). Since learning progressions that can inform standards need to emphasize larger themes and patterns in how students learn big ideas, they tend to have a larger grain size. Thus such learning progressions do not include as much detail about student understanding at particular points in time as learning progressions with finer grain size. In addition, the shifts between levels are likely large because they capture "major shifts in worldviews" (Mohan & Plummer, this volume, p. 142). The large breadth and coarse grain size of this type of learning progression, while necessary for a long-term vision, may be less useful for classroom teachers who have to respond to questions and ideas raised in classroom discussion or in embedded assessments (Furtak et al., this volume). Learning progressions for smaller scale use, such as classroom assessments or smaller scale curricular units, are likely less broad (i.e., with less content) and have a smaller grain size. These learning progressions contain greater detail at each level describing the nature of student thinking. In addition, the shifts between levels may capture more "discrete changes in conceptual networks" (Mohan & Plummer, this volume, p. 142). Yet coherence is needed between the large-scale applications (e.g., standards, large-scale assessments, and comprehensive curriculum development) and small-scale applications (e.g., classroom assessments and instruction).

Some of the discussion at the LeaPS conference in the using strand focused on how learning progressions should "look" for different uses. The participants proposed two main ideas: (1) "zooming in and out" of learning progressions and (2) embedding learning progressions in products and tools.

Zooming In and Out

As described above, the main differences between learning progressions that are best suited to inform different uses are their breadth and their grain size. However, learning progressions that can inform different uses might be thought of as nested rather than separate. Thus one way to reconcile the differences in breadth and grain size of large-scale and small-scale learning progressions is to consider a zooming in and out process. For example, Wiser et al. (chapter 16) start with the large-scale (K-8) learning progression for matter and atomic-molecular theory (Smith, Wiser, Anderson, & Krajcik, 2006), zoom in on the Grades 3–5 portion of this learning progression, and iteratively elaborate the Grades 3–5 learning progression and use it for curriculum development. The larger, "zoomed-out" learning progression includes key questions, big ideas, and components of big ideas for each grade band (Smith et al., 2006). This learning progression is both a vision tool and a large-scale map of where students should be when they begin high school. It offers a macroscopic view of broad levels of student understanding (and common misunderstandings) at three grade bands. Thus, the learning progression is well-designed for informing standards and large-scale assessments because it includes a large core idea that develops over time.

However, this "zoomed-out" learning progression is not as helpful in the design of specific curricula and classroom activities (Alonzo & Gearhart, 2006) because it is not as "magnified." Magnification, in this sense, means that the "zoomed out" learning progression lacks details about how students' conceptions develop over shorter periods of time. The "zoomed-in" learning progression elaborates the ideas in the larger learning progression and includes the microscopic details necessary for designing curricula and classroom instruction. The "zoomed-in" learning progression also includes other tools such as experiences and key representations that teachers use when working directly with students in order to foster reconceptualizations (Wiser et al., this volume).

Not all learning progressions are as neatly nested. Some learning progressions address similar content but have been designed for different uses and, perhaps, by different research groups. These learning progressions may have different breadths and grain sizes. Since these learning progressions have been developed in parallel rather than in tandem, the relationships between these learning progressions may not be as neat as the relationships described for the learning progression for matter. However, it is possible to make important connections post-hoc. For instance, one might think of a learning progression with a narrow focus on plant nutrition (e.g., Alonzo, Benus, Bennett & Pinney, 2009) as nested within a learning progression with a broader focus on carbon cycling in socio-ecological systems (e.g., Gunckel, Mohan, Covitt, & Anderson, this volume). The overall goal might be for students to understand how carbon cycles in socio-ecological systems (including the process of photosynthesis) (e.g., Gunckel et al., this volume). However, in order to reach that goal, students must first understand the role of plants in these systems—the capture of energy from the sun and its transformation to a form that other organisms in an ecosystem can use (Alonzo et al., 2009). Thus an understanding of the role of plants in carbon cycling may be an important stepping stone that, with careful scaffolding by teachers and curriculum materials, can support students in identifying patterns that build toward big ideas with larger explanatory power. These stepping stones may appear in standards as goals for specific grade levels or as instructional units. The upper anchors of broader learning progressions may appear in standards at the end of larger grand bands. Small-scale learning progressions may be used to inform classroom activities while large-scale learning progressions may be more useful in standards design or in large-scale assessments.

Products and Tool Systems

Learning progressions alone are not ready for specific purposes. In addition to providing users (e.g., teachers, curriculum developers, assessment developers) with learning progressions, we have to provide them with either learning progression-based products or tools for translating learning progressions into useful products. Therefore a second way to make learning progressions useful is to create product or tool systems that instantiate learning progression ideas. There are different ways of thinking about such products or tools.

One method is to directly instantiate the ideas from the learning progression into products or tools that accompany the learning progression. For example, in the Daphne project (Furtak et al., this volume) a learning progression was combined with formative-assessment activities and feedback strategies for teachers. These tools (i.e., assessments and feedback strategies) use ideas from the learning progression to elicit students' ideas with respect to the learning progression and include methods to help teachers interpret students' responses and support students in moving toward more sophisticated ideas. Similarly, in the MBI project (Furtak et al., this volume) novice teachers used a "reduced" or simplified learning progression (for teacher learning) as a vision tool to describe and improve their practices. This tool was the learning progression itself (although simplified for use by novice teachers). However, the reduced learning progression was used in a methods class and was accompanied by other tools (e.g., an Explanation Tool) intended to scaffold novice teachers toward more ambitious practices.

The tools developed in the Daphne and MBI projects provided users (i.e., the teachers) with the learning progression. The projects also included other products or tools as part of a system that made the learning progression useful. In addition, in both projects, the learning progressions and associated products and tools were used in social learning environments (methods classes and professional learning communities). Thus, in order to design learning environments that support teachers' use of learning progressions, it may be especially important to recognize the importance of basing the use of products and tools in interpersonal settings (e.g., see Vygotsky, 1978).

Another method for instantiating ideas from a learning progression into products is to use intermediary tools that promote the translation of the learning progression for specific uses. For example, in both a curriculum design process (Wiser et al., this volume) and in standards design (Foster & Wiser, this volume) the researchers identified core ideas, lower and upper anchors, and stepping stones useful in translating the learning progression into products. In their work on standards, educators and policy makers in the State of Massachusetts developed a "concept and skills progression tool" (Foster & Wiser, this volume). Experts use this tool as a template to systematically synthesize relevant research and make inferences about missing information in a format that was useful for developing standards. This tool permits specification of upper and lower anchors as well as identification of stepping stones that link the upper and lower anchors. Experts were also asked to present a narrative storyline and the skill details that showed students' conceptual growth over time (Foster & Wiser, this volume). The information provided by this tool can then be used in the ultimate product, the standards.

Wiser et al. (chapter 16), in their Inquiry Project, also used core ideas, lower and upper anchors, and stepping stones as tools to guide their curriculum design process. These intermediary tools, along with lever concepts and key representations, were used to (iteratively) move from a learning progression to curricula designed to foster reconceptualization. Their curriculum design process, which involved iteratively elaborating a section of the larger learning progression, and designing curricula based on this elaborated section is a also tool that instantiates the learning progression ideas. The Inquiry Project's curriculum design process highlights the importance of documenting not just the decisions and processes for defining learning progressions but also the decisions and processes for translating learning progressions into coherent tools and products for specific users. Other researchers can then use these design process tools to zoom in on a larger learning progression and to develop intermediary tools to create curricula or other products.

WORKING WITH STUDENTS' MISCONCEPTIONS IN LEARNING PROGRESSIONS

Learning progressions combine what we know about core ideas in science and about how students learn in specific domains. While the upper anchors of learning progressions represent scientifically accurate understandings of core concepts, lower anchors and middle levels may include misconceptions or scientifically inaccurate ideas. The lower anchors and middle levels attempt to describe student understandings accurately as students progress towards the upper anchors. However, the lower anchors and middle levels may look very different from a disciplinary decomposition of the domain. By including misconceptions or scientifically inaccurate ideas, learning progressions honor what students bring to learning and respect how their initial ideas interact with instruction. However, when translating learning progressions into tools for specific uses, there are questions as to how to handle scientifically inaccurate ideas. Participants at the LeaPS conference shared ideas, described next, for translating both misconceptions and inaccurate ideas in the lower anchors and middle levels of the learning progression into specific tools.

Honoring Misconceptions Using Educative Features

Students come to the learning process with many rich ideas based on their experiences and their attempts to make sense of those experiences. Because these ideas are often scientifically unsophisticated, student misconceptions can be expected to appear at the lower anchor and middle levels of learning progressions. While documenting the misconceptions that students hold at various levels is essential for accurately representing the nature of their understandings, when misconceptions are included in specific levels, it is sometimes difficult to tease out the most constructive ways for designing instruction and working with teachers. This is because misconceptions may or may not be productive; in fact, some misconceptions may hinder learning. Thus care must be taken so that the lower anchor and middle levels of learning progressions do not present unproductive misconceptions as learning goals. However, for many purposes, learning progression-based tools can honor student misconceptions at different levels by including them as educative features in curricula (Davis & Krajcik, 2005) and by designing assessment items that elicit these ideas (e.g., Briggs, Alonzo, Schwab, & Wilson, 2006; Furtak et al., this volume).

Educative curricular materials often relate common student misconceptions with specific activities as well as provide tools for teachers to probe student understanding of these misconceptions (Davis & Krajcik, 2005). Sometimes, however, these curricula present students' misconceptions as ideas that must be identified and "fixed." Educative curricular materials based on learning progressions may take a different stance towards these misconceptions by suggesting activities that build on them as students move from one level to the next.

Similarly, learning progressions can inform the design of other educative tools such as assessment probes. In the Daphne project, Furtak and her colleagues (chapter 17) created an educative tool based on the learning progression that includes formative-assessment activities and sample student responses. The tool also includes suggested feedback strategies that teachers can use with students who give particular responses. This tool, then, honors students' ideas and suggests how teachers can use these ideas to support students in moving towards upper anchor understandings.

In their MBI project, Windschitl, Thompson and their colleagues (chapter 17) developed a "user-friendly" version of the learning progression that teachers (both novice and candidate) could use as a diagnostic tool to describe their current practice and as a vision tool to imagine their ideal practice. The modified learning progression has a note at the upper anchor: "Aim for this!" This tool scaffolds teachers in self-assessment by helping them to be metacognitive in identifying where their practice currently falls and in seeing the goal practice of model-based inquiry. Using this information, teachers can work with their methods instructors to bridge the gap between their current practice and the upper anchor. In methods courses, instructors teach how to plan, enact, and reflect on upper anchor, model-based inquiry classroom practices. However, instructors acknowledge the realities of the classroom and the challenges that teachers face in implementing model-based inquiry and accept that there are different levels of practice. Instructors also acknowledge movement along the progression and the path that teachers follow as they develop more sophisticated practices.

Using Alternative Ideas in the Learning Process

Learning progressions identify the components of big ideas that can be addressed at different levels of sophistication. Occasionally, these components are out of reach for younger students. Ideas that are not completely accurate from a scientific point of view may be more accessible to such students and, through careful instruction, may be reconceptualized to support scientific understandings. Thus it may be fruitful, and perhaps even essential, to teach students ideas that are not entirely scientifically accurate in order for them to interact with content at a developmentally appropriate level. These scientifically inaccurate ideas are a "means to an end" in which the end is for students to achieve upper anchor understandings. For example, both Wiser et al., (chapter 16) and Foster and Wiser (chapter 18) use the example of teaching the stepping stone that "any piece of matter has weight." While the idea that matter has weight is only true in a gravitational field, the idea that "any piece of matter has mass" is not meaningful for students in elementary grades because mass is not an understandable concept for them. However, they are familiar with the concept of weight, which they can experience in the classroom through "hefting" and using scales (Wiser et al., this volume). Wiser and colleagues argue that the stepping stone "any piece of matter has weight" is essential to developing the scientific concept "any piece of matter has mass" later. However they note that both scientists and science educators may challenge the use of inaccurate stepping stones, especially if such stepping stones appear in standards documents.

Participants at the LeaPS conference agreed that the use of inaccurate ideas as learning goals (either in standards or curriculum materials) must be justified. There has to be substantial empirical evidence that (1) the inaccurate idea is generative and (2) the inaccurate idea can be reconceptualized as a scientifically accurate idea using specific tools or strategies. While there is a substantial (though, certainly not comprehensive) research on student misconceptions, there is little empirical evidence for methods for reconceptualizing inaccurate ideas. Thus research is needed that investigates whether and how this reconceptualization takes place.

MAKING LEARNING PROGRESSIONS READY FOR USE

As some challenges described in the using strand chapters and the discussion in the sections above reveal, many crucial decisions are necessary before learning progressions are ready for use. At the national level, learning progressions were initially an organizing principle in the preliminary draft framework for the new national science education standards (NRC, 2010). However, taking into account feedback from experts in the field, the authors of final framework do not include this role for learning progressions. As the final framework states, "R&D [research and development] on learning progressions in science is at an early stage, many aspects of the core ideas and their progressions over time with instruction (as sketched out in the framework) remain unexplored territory" (NRC, 2011, p. 13–4). In their development of tools to bridge this gap in the research, Foster and Wiser (chapter 18), documented the challenges inherent in this process. Despite these challenges, both Foster and Wiser and the authors of the national framework (NRC, 2011) realize the potential for using learning progressions and call for research to develop learning progressions that could be used in future iterations of standards.

In order for learning progressions to inform standards (and other uses), four criteria must be met. First, learning progressions must center on core ideas, and the core ideas must be clear (Foster & Wiser, this volume). In general, core ideas are ideas that scientists think are important, that informed citizens should know, and that can be used to structure the learning progression from the lower anchor to the upper anchor. Thus core ideas are important in the "learnability" sense as well as the scientific sense (M. Wiser, personal communication, August 31, 2011). However, having stakeholders use these criteria to come to a consensus about what core ideas are may be difficult. While the criterion of centering on core ideas is not new, it is essential if learning progressions are to narrow our focus to the big ideas that help students achieve socio-politically important understandings. It is not useful for either large-scale applications (e.g., standards, large-scale assessments) or smaller scale applications (e.g., tools for formative assessment or classroom activities) if there are learning progressions for every topic in the current science curriculum.

Second, learning progressions must have lower anchors that take into account students' preconceptions in a domain. For standards to move away from a solely top-down decomposition of the domain, learning progressions must honor and incorporate the ideas that students bring to learning and incorporate ideas about how students learn. For learning progressions to be useful at the classroom level, they must provide information to help move students from naïve understandings to higher stepping stones and upper anchor ideas.

Third, learning progressions must have an upper anchor that is based both on socio-political ideas and also empirically shown to be attainable by students graduating high school. Although current standards and many classroom-based tools (e.g., curricula) are somewhat based on socio-political ideas, mapping the upper anchors to specific understandings for scientific literacy can be strengthened by the focus on core ideas. Learning progressions can also provide evidence about the appropriateness of the ideas taught to high school students. Some current standards and classroom-based tools may include ideas that are too complex given new understandings of how students learn, while other standards and classroom-based tools may be made more rigorous with new sequences of stepping stones and instructional techniques.

Fourth, the levels of learning progression must include empirically-based stepping stones that represent significant reconceptualizations toward the upper anchors. These sequences of ideas may be radically different as the result of new conceptions of how students learn and of the effects of learning progression-based instruction.

It is apparent in consideration of these criteria for learning progressions that empirical evidence is necessary for decision-making. Given the impact that standards have on the educational system, evidence to validate learning progressions, especially those that inform standards and other large-scale uses, must be strong. The evidence should be obtained from a representational sample of students. One way to collect evidence about students who are at different grade levels and who come from diverse backgrounds is to use the student responses from state or national assessments (e.g., Alonzo, Neidorf, & Anderson, this volume). Data from such large-scale assessments can be extremely useful since these datasets are very large and include information from a wide range of students. These data are also helpful in testing various new measurement models for use with learning progressions because many models require large datasets. However, such datasets only provide information about what students can achieve with status-quo instruction. In fact, these datasets do not provide information about the instruction students have received on the specific content; some students may have had many quality opportunities to learn the content while others may have had few or none.

Since learning progressions are not developmentally inevitable, data gathered from large-scale assessments are insufficient as explanations of whether a learning progression is useful for standards development or any other application. It is important to develop products and tools to test the ways students move along a learning progression (e.g., curricula, assessments, teacher preparation/ professional development). Testing the products and tools can provide validity evidence about the learning progression itself (i.e., whether students progress in the hypothesized ways) and also provide information about the learning progression-based products and tools (i.e., whether these products and tools scaffold students toward more sophisticated understandings). Thus, it is in these smaller scale uses that much evidence needed to flesh out the learning progressions research base can be gathered.

Design-based research is an essential type of learning progressions research because it allows for iterative design of interventions. Such interventions include curricula, professional development, and revision of the learning progression based on results (Barab & Squire, 2004; Brown, 1992, Cobb & Gravemeijer, 2008). The research that can move learning progressions forward will likely, "start with design experiments situated in classrooms that explore (a) how to specify the knowledge to be acquired by students at particular grade bands and (b) what instructional approaches might best support the proposed progressions" (NRC, 2011, pp. 13–4).

However, conducting design-based research with learning progressions is not straightforward. Because learning progressions are longitudinal, changes to the learning progression or tools at one level will affect content in the learning progression and tools at other levels, both above and below. Wiser et al. (chapter 16) recognize that they may need to reconsider their research on the grades 3–5 section of the learning progression after they have "zoomed-in" on the K-2 section and develop instructional tools for younger students. The reason for this reconsideration is that students who have had curricula that promote reconceptualization of specific aspects of the learning progression for matter before third grade may be at a higher level of the learning progression than students who have not experienced such curricula. "Thus very little is known about what can develop in later grades on the basis of successful implementation of solid learning progressions for a concept in the earlier grades" (NRC, 2011, pp. 13–4). In addition, design-based research on learning progressions may push ideas from traditional sequencing forward and have younger students working with ideas that have traditionally been regarded as out of reach. Conversely, learning progressions may problematize ideas (e.g., mass) that in past were assumed self-evident, thereby requiring more time on such key ideas (Lehrer & Schauble, in press). Similarly, design-based research on learning progressions may result in more effective instruction toward traditional learning goals, or it may redefine what is possible based on new ways of conceptualizing learning (Cobb & Gravemeijer, 2008).

Ultimately, multiple types of evidence must be gathered to validate learning progressions for multiple uses. These data should be assembled in systematic ways in order that learning progressions with different breadths and grain sizes can be coherently connected to multiple aspects of science education. For example, learning progressions that have been informed by what students can achieve with specific instruction can be used to create assessment items that are tested at the larger scale. The data from student responses to these items may inform the validity of the learning progression for larger audiences. The data may also allow researchers to refine their learning progressions and the associated products and tools for scale-up with more diverse audiences. The process of combining data from multiple sources, a process that is neither simple nor linear, is an area requiring additional investigation.

CONCLUSIONS

Learning progressions may help us address the challenges we encounter when standards and curricula are "a mile wide and an inch deep" (Schmidt, McKnight, & Raizen, 1997, p. 122). State or national standards control many aspects of science education. In addition, state and national assessments are developed on the basis of standards (see the assessing strand of this book for a discussion of the challenges in assessments). In backwards design of learning materials and environments (Wiggins & McTighe, 2006), the first step is to articulate the learning goals and outcomes, which in most cases are dictated by standards documents. Thus incorporating learning progressions into standards could create a powerful "trickle down" effect that integrates learning progressions with curriculum development, instruction, assessments, and teacher preparation/ professional development. However, given the strong influence of standards, it is imperative that learning progressions are not pushed into prime time before thorough research has been conducted (e.g., Shavelson & Kurpius, this volume).

Although learning progressions may bring coherence to standards and curricula, we should not forget the promise of learning progressions for smaller scale uses such as classroom assessments. Learning progressions that have a smaller grain size and cover less content (perhaps a "zoomed-in" section of a larger learning progression) can transform how students are taught. Educative tools and products can support teachers by helping them work more closely with student ideas as they scaffold students towards scientific ideas (e.g., the Daphne project, Furtak et al., this volume). Revised curricula can include experiences that also scaffold students in reconceptualizing their ideas and moving them toward scientific understandings (Wiser, et al., this volume).

More research is needed on how broad learning progressions that may inform standards and large-scale assessment can be "zoomed-in" and unpacked for small-scale use, such as curriculum development and classroom assessments. In addition, research is needed on how learning progressions with small and large grain sizes can be connected so that learning progressions meet their promise of bringing coherence to multiple levels of science education.

REFERENCES

Alonzo, A. C., Benus, M., Bennett, W., & Pinney, B. (2009, September). *A learning progression for elementary school students' understanding of plant nutrition*. Paper presented at the biennial meeting of the European Science Education Research Association, Istanbul, Turkey.

Alonzo, A. C., & Gearhart, M. (2006). Considering learning progressions from a classroom assessment perspective. *Measurement: Interdisciplinary Research and Perspectives, 14*, 99–104.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*, 1–14.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*, 141–178.

Cobb, P., & Gravemeijer, K. (2008). Experimenting to support and understand learning processes. In A.E. Kelly, R. Lesh, & J. Y. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching* (pp. 68–95). London, UK: Routledge.

Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher, 34*(3), 3–14.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, *32*, 268–291.

Gravemeijer, K., & Cobb, P. (2008). Design research from a learning design perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 17–51). New York: Routledge.

Lehrer, R., & Schauble, L. (in press). Supporting inquiry about the foundations of evolutionary thinking in the elementary grades. In J. Shrager and S. Carver (Eds.), *From child to scientist: Mechanisms of learning and development*. Washington, DC: APA Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press.

National Research Council. (2010, July). *A framework for science education: Preliminary public draft*. Retrieved from the National Academies website: http://www7.nationalacademies.org/bose/Standards_Framework_Homepage.html

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Smith, C.L., Wiser, M., Anderson, C.W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*, 1-98.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. London, UK: Cambridge University Press.

Wiggins, G., & McTighe, J. (2006). *Understanding by design* (expanded 2nd ed.). Upper Saddle River, NJ: Pearson Education.

CONCLUDING SECTION

ALICIA C. ALONZO AND AMELIA WENK GOTWALS

LEAPING FORWARD

Next Steps for Learning Progressions in Science

The chapters in this book illustrate the complexity of work on learning progressions. Through their discussion of issues and design decisions, the authors of these chapters highlight the challenges of this work. This focus is a departure from typical portrayals in conference presentations and journal articles. Therefore, we appreciate the authors' willingness to share aspects of their work that are rarely made public. While there is certainly great interest in the products of work on learning progressions (e.g., completed learning progressions and associated tools, reports on the impact of learning-progression-based tools), this book takes a step back to carefully consider the challenges in learning progression work. In this sense we embrace the stance that "anything worth doing is difficult" (and, we would add, complex). Work on learning progressions seeks to capture and inform student learning in all its complexity. This lofty goal is not easily accomplished. However, as described in the introductory chapter (Gotwals & Alonzo, this volume), the pay-off is likely to be substantial if learning progressions can achieve their promised influence on science education.

As work on learning progressions advances, we believe that candid discussions are essential. Consideration of the challenges and potential solutions employed by the chapter authors can help the field to mature so that each group is not "reinventing the wheel." Cumulative efforts may then lead to deeper knowledge of how students learn and greater clarity about the learning progression construct. The result may be more valid learning progressions and associated tools. Illustrations of the complexity of work on learning progressions may also create more realistic expectations about the prospects for immediate application of learning progressions in large-scale policies and practices. While quick fixes may be sought, and we may become more efficient in our work on learning progressions, it may be neither desirable nor possible to take shortcuts in the iterative process of developing valid learning progressions and associated tools.

Through descriptions of the challenges they face in their work, the chapter authors show how even seemingly simple aspects of the development, investigation, and use of learning progressions require design decisions—decisions that are constrained by a number of factors and that reflect underlying commitments. For example, defining a learning progression is not simply a matter of synthesizing research literature or hypothesizing an order for scientific topics. Decisions must necessarily be made about the focus and grain size of the learning

Alicia C. Alonzo, Amelia Wenk Gotwals (eds.), Learning Progressions in Science: Current Challenges and Future Directions, 475–490.

progression and about how the levels represent progress towards targeted knowledge and practices. These decisions have consequences for assessing, modeling, and using learning progressions. Consistent with principles of design research (e.g., Gravemeijer & Cobb, 2006), it is essential for those working on learning progressions to articulate their design decisions and underlying commitments so that that both are open for debate.[1] We hope that the chapters in this book provide models for making these aspects of work on learning progressions explicit and subject to discussion.

CONSIDERING CHALLENGES ACROSS STRANDS OF WORK ON LEARNING PROGRESSIONS

In order to structure both the LeaPS conference and this book, we divided work on learning progressions into four strands: defining, assessing, modeling, and using. Each strand represents a different aspect or phase of this work. However, because the strands are tightly intertwined, it is not possible to work on a single strand without considering other strands. Thus the projects described in the exemplar chapters do not represent isolated work in a particular strand; rather, the authors have highlighted certain aspects of their work in this book. As many of the chapters illustrate, the work of defining (and refining) learning progressions necessarily involves design of assessments, development of techniques for modeling students' responses to those assessments, and careful consideration of the uses intended for particular learning progressions. Thus, in this concluding chapter, we pull the strands together in order to consider larger challenges in work on learning progressions.

Challenges that Cut across Strands

Lack of an adequate research base. In 2001, the National Research Council (NRC) report *Knowing What Students Know* called for assessment practices to be more responsive to "the cognitive and measurement sciences" (p. 1). As discussed in the assessing synthesis chapter (Alonzo, this volume) and reflected in other chapters, researchers have heeded this call. These chapters also call for further research that can inform work on learning progressions.

In the defining strand, research on student thinking is limited (both in scope and in nature). As additional research is conducted to explore student thinking across the K-12 curriculum, it may be particularly important to move beyond documentation of misconceptions (Alonzo, 2011) in order to describe student thinking in the "messy middle" (Gotwals & Songer, 2010, p. 277). In addition, there is little longitudinal research that documents student learning (the change in thinking) over time. Eventually, we need to move beyond the description of learning progressions as "partly hypothetical or inferential" (National Assessment Governing Board [NAGB], 2008, p. 90); studies are needed that explore whether student learning actually proceeds as hypothesized in learning progressions.

In the assessing strand, while promising techniques exist for eliciting students' knowledge and practices, novel approaches to assessment are needed. Learning progression assessments must reflect a "learning progression stance" (Alonzo, this volume, p. 247), meaning that they capture the nuances in student thinking rather than simply whether they "get" an idea. In addition, learning progressions often describe complex performances that must be elicited by assessment items. The challenges faced in this strand call for research that investigates the affordances and constraints of new item formats and delivery platforms.

In the modeling strand, new psychometric techniques may be needed to reflect (a) assumptions about the nature of student thinking as reflected in learning progressions and (b) the features of items designed to assess that thinking. In addition to investigating new modeling techniques, research is needed that evaluates methods for comparing fit across models.

Finally, the using strand raises issues about the types of learning progressions that are most useful to various stakeholders. Research is needed that examines the usefulness of learning progression products before they are "rolled out" on a large scale. Because learning progressions represent a non-traditional way of conceptualizing student thinking and learning, communication to stakeholders is particularly important. Otherwise, as Fullan (2000) noted about the large-scale reforms of the 1950s and 1960s, language will be adopted without fundamental changes in teaching practices, and learning progressions will fail to achieve their desired effect on the educational system.

Representing student thinking. The chapter authors ask questions about the nature of student thinking and its representation in learning progressions and associated products. The complexity of student thinking and learning challenges simple description. Learning progressions necessarily reflect some simplifications. The choices we make about what to simplify and what to keep complex may have important implications for the utility of learning progressions in both research and practice. The chapter authors in the defining and using strands consider the role of misconceptions in learning progression levels. On the one hand, if learning progressions are to honor student thinking as it develops, it may be important to include typical and/or productive misconceptions that—while not scientifically correct—arise as students move between naïve ideas and scientific ideas. On the other hand, there may be good reasons not to use these scientifically incorrect ideas as instructional goals in standards and curricula. As discussed in the using synthesis chapter (Gotwals, this volume), to a certain extent, the issue concerns how learning progressions are translated into products for other stakeholders. However, this issue also concerns the nature of the learning progression construct itself and, as such, must be resolved as learning progressions begin to have a larger influence on educational policies and practices.

The issues about the nature of student thinking raised in the assessing and modeling strands are perhaps even more fundamental. Several questions about the representation of student thinking in levels underlie work in these strands. To what extent are levels an appropriate way of representing student thinking? What assumptions do researchers make about these levels? For example, do we assume that students' knowledge and practices are coherent and therefore that students should operate consistently at a given level? Or do we assume that students are likely to exhibit characteristics of a variety of levels, depending upon the situation?[2] While concerns have been raised about levels as a means of representing student thinking, participants at the recent National Science Foundation (NSF) learning progression “footprint” conference[3] asked, “If not levels, then what?” Even if one assumes that levels provide an adequate representation of student thinking, questions arise about the nature of those levels. Are they simply ways of parsing a continuous quantity? Or are they more “step-like”?[4] In addition, as a field, we have not yet grappled with the likelihood that “there may be multiple pathways by which certain understandings can be reached,” where “prior instructional experiences, individual differences, and current instruction” (NAGB, 2008, p. 90) influence the pathway that a particular student takes. How can multiple pathways be represented in a learning progression? And what are the implications for assessing, modeling, and using learning progressions?

As the field addresses these questions about the nature of student thinking and learning and considers how to represent the answers in learning progressions and associated tools, it is particularly important that we state our assumptions. If—as is the case for all learning progression examples to our knowledge—levels are defined, what assumptions are made about student thinking relative to those levels? What do the levels represent? What simplifications are made? Forthrightness in answering these questions is important in communicating with other researchers and stakeholders.

Deciding on grain size. Throughout the book, chapter authors mention the issue of grain size. However, as we worked with them, we realized there were different ways of considering grain size. Clarity of terminology is an important first step towards careful discussion of this issue. It may be that grain size needs to be considered in multiple ways (e.g., Alonzo, this volume), but this can only occur when we are clear about our definitions.

In the defining synthesis chapter (Mohan & Plummer, this volume), grain size refers to the size of the levels—or to the size of the transition between adjacent levels. In the modeling synthesis chapter (Briggs, this volume), grain size refers to the detail used to describe each level. While these different definitions of grain size are clearly related—and related to the overall breadth of the learning progression—they may lead to different decisions and different trade-offs in how learning progressions are assessed, modeled, and used.

Grain size is just beginning to be discussed in the field, and as yet there are no simple guidelines to follow in making grain-size decisions. As with other design decisions, it is important that there is clarity about the choice of grain size. Because grain-size decisions depend on the interplay between empirical evidence and practical considerations, researchers will make different decisions about grain size; the reasons for these decisions should be documented. As the four synthesis chapters discuss, the intended use of a learning progression may determine whether a finer or a coarser grain size is more useful. For example, while teachers may need fine-grained learning progressions to inform instruction, standards developers may find coarse-grained learning progressions more suited to their purposes. However, practical constraints (such as the number of items that can be administered, whether in large-scale or classroom assessments) may limit our ability to differentiate between levels. With a limited number of items, fine-grained diagnoses of student thinking may not be possible. In addition, the nature of student thinking may affect decisions about how best to portray that thinking in learning progression levels. While it may be theoretically possible to make fine-grained distinctions between levels, empirical evidence may indicate that these distinctions cannot be made reliably on the basis of students' performances. In this case, a coarser grain size may be more appropriate. Conversely, empirical evidence may indicate that the broad levels of coarse-grained learning progressions do not capture meaningful differences in students' performances (e.g., Schwarz, Reiser, Acher, Kenyon, & Fortus, this volume). In such instances, depending upon the purpose of the learning progression, a finer grain size may be more useful.

Evaluating/validating learning progressions. As researchers grapple with the challenges described above, they face uncertainty in evaluating their work. Clear criteria do not yet exist for evaluating learning progressions and their associated products (such as curricula and assessments). Anderson (2008) contributes to this discussion by listing possible criteria for use in a validity argument for a learning progression. He starts with three qualities that learning progressions should possess: conceptual coherence, compatibility with current research, and empirical validation. In a table entitled "Criteria for Validity of Learning Progressions" (p. 4), Anderson uses these qualities to develop criteria for three components of learning progressions: individual cells, levels of achievement, and progress variables. For example, applying the empirical validation quality to levels of achievement results in the following criterion: "Levels have predictive power: Students should show similar Levels of Achievement for Learning Performances associated with different Progress Variable [sic]." As another example, applying the empirical validation quality to progress variables results in the criterion: "Progress from one Level to the next can be achieved through teaching strategies that directly address the differences between Learning Performances."

However, there are few examples of such validation work, and the field has not agreed on what "counts" as high-quality (or even adequate) evidence. As Corcoran, Mosher, and Rogat (2009) note, validation efforts vary in the extent to which evidence has been gathered and in the methods used. Perhaps most saliently, some validity evidence has been collected under conditions of "status quo" instruction, while other evidence has been collected in very different (and carefully designed) instructional conditions (Mohan & Plummer, this volume).

Status quo evidence is often used in investigations of individual levels, which may focus on the predictive power of these levels (e.g., Alonzo & Steedle, 2009). Because this type of validation work requires careful consideration of the extent to which different learning performances "hang together" in the same level, the focus is on eliciting individual students' thinking about a range of phenomena. Therefore, methods such as cognitive and think-aloud interviews are commonly used.

In contrast, researchers who provide students with carefully constructed instructional experiences (e.g., Wiser, Smith, & Doubler, this volume) may focus more on the transitions between levels and on whether instruction is effective in achieving these transitions. This work addresses the criterion that teaching strategies can be used to achieve progress between levels. This validation work may involve a variety of methods. Wiser et al. use cognitive interviews to explore how individual students experience their carefully designed instruction. Anderson's group (e.g., Mohan & Anderson, 2009) implements teaching experiments—targeted efforts to advance students from one level of a learning progression to the next—and uses paper-and-pencil assessments to evaluate the progress made by students under these conditions.

These differences in approaches to validation may reflect different stances or simply different stages of work. For example, Anderson's group has explored student thinking under conditions of both status quo and carefully designed instruction. They began with status quo instruction (e.g., Mohan, Chen, & Anderson, 2009), but more recently they have turned to carefully designed instruction (e.g., Gunckel, Mohan, Covitt, & Anderson, this volume; Mohan & Anderson, 2009).

As a field, we need to engage in discussions of the criteria for and the approaches to evaluating learning progressions. These discussions may vary depending upon the different uses of learning progressions. For example, a higher level of certainty may be needed in high-stakes situations in which opportunities for correction are limited (e.g., accountability testing) than in low-stakes situations where corrections can easily be made (e.g., informing day-to-day classroom instruction). Without prematurely restricting the development, refinement, and validation work currently in progress, the field may eventually develop an argument (such as Clements, 2007, makes for curriculum research) for the research needed to provide validity evidence for learning progressions (and for what acceptable evidence of validity might entail).

Challenges in Coordinating across Strands

As has been argued throughout this book, work on learning progressions requires collaboration among those with varied expertise. The work of developing, refining, and validating learning progressions requires iterative cycles with input from science educators, learning scientists, assessment experts, and psychometricians. While all these players should be involved in learning progression work, each strand requires different combinations of expertise. Thus the strands may highlight contributions of specific groups. For example, psychometricians may primarily be involved in the assessing and modeling strands. However, their input is also needed in the defining and using strands.

As we begin to organize science education around "big ideas," traditional disciplinary boundaries may become more porous. As a result, learning progressions may describe knowledge and practices that transcend the divisions among physics, biology, chemistry, earth science, and even mathematics. For example, work on a learning progression for matter and atomic-molecular theory (e.g., Smith, Wiser, Anderson, & Krajcik, 2006; Wiser et al., this volume) acknowledges the importance of students' epistemic knowledge about measurement. This work has important parallels to work on a learning trajectory for measurement in mathematics (Barrett et al., 2012). In addition, there are connections between the work of Lehrer and Schauble (2004) on data-modeling to support students' understanding of evolution and the work of Cobb and colleagues on students' understanding of statistical covariation (e.g., Cobb, McClain, & Gravemeijer, 2003). Thus there may be important connections at the level of individual learning progressions/trajectories that provide opportunities for curricular coherence across subject areas.

Stakeholders outside the research community may begin to apply learning progressions in decisions about standards, assessments, curricula, and teacher professional development. These stakeholders must be part of our conversations about learning progressions. Learning progression work requires unusually high levels of collaboration among those who may not have worked together in the past. Although we may come from communities with very different ways of conceptualizing and talking about student thinking and learning, this diversity can be a benefit. Such diversity may contribute to the development of products that reflect the best thinking in a number of areas. Yet this diversity may also present challenges. We are not always used to working with people who use different professional language and make different assumptions about student thinking and learning. We need "boundary crossers" who can facilitate our work by clarifying the language and assumptions that are essential for productive collaboration.

MOVING FORWARD

Contributing to Policy Conversations

In the preceding discussion and, indeed, in most of the book, we have portrayed the interconnections among the four strands of work on learning progressions. However, there is an inherent tension between the work needed to develop/refine learning progressions and associated tools (the first three strands) and the press for immediate application of learning progressions (the using strand). For example, the NRC (2011) called for a significant role for learning progressions in the development of new science standards. The learning progression construct is already influencing the development of standards in mathematics. The Common Core State Standards (CCSS) Initiative (2010) claims that the development of standards in mathematics "began with research-based learning progressions detailing what is known today about how students' mathematical knowledge, skill, and understanding develop over time" (p. 4). However, as Shavelson and Kurpius (chapter 2) caution, learning progressions may not be ready for prime time. Indeed, the CCSS Initiative acknowledges the limits of existing work on learning progressions: "One promise of common state standards is that over time they will allow research on learning progressions to inform and improve the design of standards to a much greater extent than is possible today" (p. 6). Despite this acknowledgement, the assessment consortia that are developing assessments of the CCSS consider learning progressions a key component of their work (D. Briggs, personal communication, July 20, 2011). Thus, while there are unprecedented opportunities for research to have immediate influence on practice, we are wise to heed Shavelson and Kurpius's caution that learning progressions may be rapidly applied and dismissed if they are not implemented carefully and as intended.

As researchers, it may be easy for us to put our heads down and focus on the esoteric details of our research agendas. However, although we may think our work is not yet ready for immediate application, we cannot afford to remain detached from the policy discussions related to learning progressions. If we do not have a voice in this arena, others will speak for us. Learning progressions are suffering from a "bandwagon effect": this label is being applied to tools that do not meet the criteria for learning progressions held by many in the science education community. Although we do not have complete consensus, documents such as *Taking Science to School* (NRC, 2007) and the Consortium for Policy Research in Education (CPRE) report (Corcoran et al., 2009) have helped achieve some uniformity in how the science education community discusses and operationalizes learning progressions. Thus, while the chapters in this book reflect variation in specific details of the proposed learning progressions (including what is progressing and how long progress may take), they share a similar view of this construct.

The learning progressions in this book focus on "core ideas," those that represent key instructional targets because they are important both scientifically and societally and "can be understood with increasing… complexity" (M. Wiser, personal communication, August 31, 2011). These learning progressions include: (1) a lower anchor that includes students' pre-instructional ideas; (2) an upper anchor that represents either sociopolitical consensus about a desired learning goal or an important "stepping stone" that is the basis for a subsequent learning progression; and (3) empirically-based middle levels between the lower and upper anchors. These middle levels are not simply a decomposition of scientists' knowledge and practices; rather they are attempts to faithfully describe the nature of students' knowledge and practices. As discussed in the defining synthesis chapter (Mohan & Plummer, this volume) and in the using synthesis chapter (Gotwals, this volume), the middle levels may include knowledge and practices that may be considered "incorrect" from a scientific point of view; however, these incorrect knowledge and practices may be important stepping stones that support students' progress towards targeted knowledge and practices.

As learning progressions gain traction and widespread popularity beyond the science education community, we need to (a) take a stand with respect to features of learning progressions that we can agree on and (b) use those features to differentiate our work from that of others who use this label to refer to very different kinds of work. For example, Popham (2011b) defines a learning progression as "a sequenced set of building blocks (that is, subskills and/or bodies of enabling knowledge) it is thought students must master en route to mastering a more remote, target curricular aim" (p. 1). This definition of learning progressions is now used in textbooks for pre-service teachers (e.g., Popham, 2011a) and in large-scale professional development programs for in-service teachers (see, for example, http://datause.cse.ucla.edu/iowa.php). These valuable forms of support for both pre-service and in-service teachers may have big pay-offs for instruction. And, as Shavelson and Kurpius (chapter 2) note, teachers may play an important role in the development of learning progressions across the K-12 curriculum.

However, we must be clear in our use of language. There are crucial distinctions between definitions of learning progressions in the science education community and other definitions that influence current educational policies. Specifically, the learning progressions described in this book are more than "scope and sequence" documents. They attempt to describe the nature of student thinking, not to present a sequence of topics. In doing so, the learning progressions in this book prioritize the messy middle and, thus, attend to the "gray area" between naïve ideas and scientific ideas. Student thinking in the messy middle may include both scientifically correct and scientifically incorrect knowledge and practices. In contrast, many learning progressions that influence current educational policies focus solely on correct building blocks—the knowledge and skills students master at each level. While these learning progressions may be specified at a smaller grain size than that of typical scope and sequence documents, they focus on correct knowledge and practices rather than on typical ways of thinking and performing at different levels of sophistication. The learning progressions in this book identify qualitative distinctions between increasingly sophisticated knowledge and practices. In contrast, other types of "learning progressions" support dichotomous judgments about whether students have mastered building block knowledge and skills.

In many ways, the field is at an interesting (and challenging) point in its development. Even in its short history, the learning progression construct has influenced policy and research conversations in significant ways. However, in this initial period of discovery, researchers are pursuing learning progression work in many different ways. This variety may be due in part because practical work in this area (i.e., developing learning progressions and associated products) is occurring at the same time that the learning progression construct itself is debated and refined (L. Schauble, personal communication, July 26, 2011). While this situation complicates efforts to communicate with policy makers, we cannot wait to engage with policy until we have answers to our research questions and validated learning progressions to share. As Shavelson and Kurpius (chapter 2) warn, by that point, policymakers "will have long ago taken down [their] tents and headed for another apparently greener pasture" (p. 16). The influence that learning progressions will have on educational policies and practices depends upon our ability to communicate the essential features of this approach.

Intersecting with Work beyond Science Education

As work on learning progressions in science matures, it is important to consider lessons from related work in other fields. Work on learning progressions in science has involved those with varied expertise, but we have rarely looked beyond our own "backyard" to consider work in other areas of education. In particular, as highlighted by the recent NSF footprint conference, work on "learning trajectories" in mathematics has been conducted largely in parallel to work on learning progressions in science.

From its inception, the learning trajectory construct in mathematics has had a classroom-level focus. Simon (1995) introduced learning trajectories as a way to support teachers' use of information about student learning in the classroom. In science, learning progressions were introduced with an initial emphasis on the "big picture." The NRC (2006, 2007) has focused on the large-scale potential of learning progressions (for example, through influence on curriculum and assessments) and has specified that learning progressions span 6-8 years of instruction. More recent work on learning progressions has considered classroom applications (e.g., Alonzo, 2011; Furtak, Thompson, Braaten, & Windschitl, this volume). However, work is needed to coordinate the use of learning progressions across different levels of the educational system. In this work and in informing classroom-level instruction and formative assessment practices, we may look to mathematics for guidance about the interplay between large- and small-scale applications of learning progressions.

In addition, researchers in mathematics education have wrestled with many of the issues we identify above. For example, Jere Confrey has worked in the policy arena to influence the development of standards based on preliminary results of research on learning trajectories (Penuel, Confrey, Maloney, & Rupp, 2011). The application of learning progressions to educational policy may occur more rapidly in mathematics because of the greater emphasis on mathematics in K-12 curricula and assessment. Science educators may therefore wish to follow the development of the mathematics CCSS assessments in order to inform efforts to use learning progressions to influence science standards and assessments.

Recommendations for Future Work and Funding Priorities

In conjunction with their contributions to policy conversations, learning progression researchers should continue to focus on the empirical research that advances the field. Much work is still needed in the development, refinement, and validation of learning progressions across the K-12 science curriculum. In this section, we argue for the type of research that can lead to important contributions to policy deliberations.

Long-term efforts. The development, refinement, and validation of learning progressions require long-term work. One of the strengths of learning progressions is that they rely on detailed descriptions of student thinking and learning. However, such descriptions take time and effort to develop. If we take seriously the notion of learning progression work as design research, it is apparent that such work is both time-consuming and exacting (e.g., Cobb et al., 2003). Therefore, research funding is needed not only for development and classroom tryouts of learning-progression-based curriculum materials but also for the refinement of those materials (and the underlying learning progressions). Advancing the field requires clear documentation of both iterative cycles of revision and underlying design decisions (including the theoretical commitments and empirical evidence on which they are based). This design work requires a great investment of time and money, but it is essential for significant changes to current educational policy and practices.

As noted above, longitudinal research is essential to move learning progressions from hypothetical constructs to empirically validated tools. Teaching experiments (e.g., Delgado, 2009; Mohan & Anderson, 2009) may demonstrate that students progress from one level to the next with appropriate instruction. However, teaching experiments are often limited in duration. Ideally, learning progressions would allow students to experience coherent instruction over longer periods of time. As learning-progression-based instruction becomes a reality in classrooms, the knowledge and practices expected of students at a particular grade level may be quite different from what students currently achieve. Therefore, a learning progression that is developed, refined, and validated on the basis of current students making progress from one level to the next may be quite different from one that is developed, refined, and validated on the basis of following students from one grade level to the next over many years. Thus learning progression work may require carefully designed instruction over longer time periods than the duration of a single teaching experiment. In addition, learning progressions that influence educational policies and practices cannot be considered static documents. In particular, learning progressions for the upper grades require revision as students experience learning-progression-based instruction in the lower grades.

Funding for learning progression work should recognize the time required for this work. Design experiments require time for iterative cycles of revision and for the documentation of those revisions. Longitudinal research with the same students over increasingly longer time periods may be required to fully understand how knowledge and practices develop. Typical funding cycles—three to five years—may be insufficient for the development, refinement, and validation of learning progressions. One solution may be to fund the initial stages of work on a learning progression with a commitment to provide additional funding for the collection of validity evidence if initial results are promising.

The work supported by the initial funding cycle may take different forms for different projects. As Plummer (chapter 5) describes, some researchers may begin with a synthesis of available literature and then develop a preliminary learning progression that draws upon this literature. They may supplement this literature review with additional research on the nature of student thinking using interviews and responses to written assessment items. For other researchers (e.g., Wiser et al., this volume), initial work on a learning progression may use the literature to hypothesize new ways of engaging students with a core idea. The learning progression can then be developed and refined through design work to support students' knowledge and practices.

Decisions about additional funding may be made at the conclusion of this initial work. Such decisions depend on considerations of what counts as "promising" research and how to determine whether more resources should be invested in a particular learning progression. Without such additional cycles of funding, we will continue to develop promising, but not empirically validated, tools.

Coordinated collaborative efforts. Work on learning progressions holds promise of achieving much-needed coherence across the educational system, such that student learning is supported over extended periods of time. However, because the development of learning progressions that span a wide age range is a daunting undertaking, it is unrealistic to expect a single research group to conduct the painstaking work required to develop an entire large-scale learning progression. As an example, consider the work on the matter and atomic-molecular theory learning progression (commissioned by the NRC and reported in Smith et al., 2006). This broad learning progression, which spans grades K-8, describes how elementary and middle school students may develop knowledge and practices essential to this key concept. Smith, Wiser and colleagues (e.g., Wiser et al., this volume) have conducted detailed work on a small portion of this learning progression. Their work is a type of nested "zooming in and out" that Gotwals (this volume, p. 463) discusses in the using synthesis chapter.

Although examples of nested work are rare, different research groups are developing learning progressions that describe students' learning of the same knowledge and practices. Post hoc connections between these learning progressions may provide approximations of more ideally coordinated work. However, as Foster and Wiser (chapter 18) note, the choices made by different research groups may impede efforts to "stitch together" learning progressions. Therefore, we recommend that the field consider funding structures that permit coordination of detailed individual projects under the umbrella of an overarching learning progression. This structure may be useful in bringing together findings from individual projects in order to develop broad learning progressions that can inform K-12 science instruction.

While individual learning progressions may provide coherence in students' learning about related knowledge and practices, attention must also be paid to the overall coherence of students' science instruction. Looking across learning progressions, we may ask: Do the knowledge and practices "hang together" in ways that support student learning? In addition to content connections across learning progressions—for example, issues of scale may be important in learning progressions for both astronomy and matter—scientific practices may provide important links between learning progressions. These practices may be developed in conjunction with a range of different content. For example, in a high school biology course, the practice of scientific modeling may be developed in intentional ways as students learn content related to genetics, biodiversity, and evolution. This requires attention not only to the individual learning progressions for genetics, biodiversity, and evolution but also to how students' knowledge and practices develop across learning progressions. Synthesis grants and/or funding for working groups may support these efforts to coordinate across learning progressions in order to promote coherence in K-12 science education.

Networks of efforts to address fundamental questions. Concurrent with long-term coordinated efforts to develop, refine, and validate learning progressions across the K-12 science curriculum, we should continue to conduct research that directly addresses challenges in learning progression work. For example, as noted above, there are important questions about the nature of student thinking and learning, about assessment and modeling techniques, and about various stakeholders' needs. While many of these questions may be addressed through large-scale work focused on learning progression products, there may be an important role for small targeted efforts that address fundamental issues. These smaller studies may be networked and linked to needs in the field, including challenges that arise in larger-scale studies. This structure may promote funding coherence, while still leaving room for smaller innovative work that may lead to generalizable findings that advance the field.

Varied expertise. As noted above, and emphasized throughout this book, work on learning progressions requires collaboration and intersections between cutting-edge research in a variety of different areas. Such collaboration is necessary, both from the beginning of the work and throughout the development, investigation, and use of learning progressions. In particular, assessment and modeling should not be an after-thought. Decisions about these aspects of learning progression work can force us to clarify our underlying assumptions and, thus, may have a significant influence on all aspects of our work.

CONCLUDING COMMENTS

Learning progressions have great potential to improve educational policy and practices. However, that potential will only be realized through collaborative work that tackles the challenges discussed in this book (as well challenges yet-to-be discovered). Funding structures can support the collaboration that is essential for coherence across learning progression work as well as more innovative research to advance the field. This is both an exciting and a daunting time. As we move forward, transparency about our use of language, our underlying assumptions, and our design decisions are particularly important. As we undertake this work, we encourage the field to be as candid as possible about challenges (and ideas about their solutions). Such work requires collaboration and, thus, clear communication. We look forward to the impact that learning progressions may have on students' educational experiences and learning.

NOTES

1. The concept of "design rationales" from the field of engineering (e.g., Moran & Carroll, 1996) may be of particular relevance to the design of learning progressions and associated products. Penuel, Confrey, Maloney, and Rupp (2011) provide an example of a design rationale for the development of a learning trajectory in mathematics.
2. Discussions of learning trajectories in mathematics have taken this position by arguing that levels

> develop gradually out of the preceding level(s) rather than being sudden reconfigurations, and that means that students often can be considered to be partially at one level while showing some of the characteristics of the next, and "placing" them… becomes a matter of making probabilistic judgments that they are more likely to perform in ways characteristic of a particular level than those that come before or after it… does not suggest that ways of thinking or operating characteristics [sic] of earlier levels are abandoned—rather students may revert to them if conditions are stressful or particularly

complex, or perhaps as they "regroup" before they move to an even higher level. (Daro, Mosher, & Corcoran, 2011, p. 24)

[3] Charles Anderson, Principal Investigator; DUE-1132562.

[4] Here, too, we may wish to look to mathematics, where researchers have begun to discuss how levels of learning progressions/trajectories are conceptualized (e.g., Battista, 2011).

REFERENCES

Alonzo, A. C. (2011). Learning progressions that support formative assessment practices. *Measurement: Interdisciplinary Research and Perspectives, 9*, 124–129.

Alonzo, A. C., & Steedle, J. S. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

Anderson, C. W. (2008, February). *Conceptual and empirical validation of learning progressions: Response to "Learning progressions: Supporting instruction and formative assessment."* Retrieved from Consortium for Policy Research in Education website: http://www.cpre.org/ccii/images/stories/ccii_pdfs/learning%20progressions%20anderson.pdf

Barrett, J., Clements, D., Sarama, J., Cullen, C., Witkowski, C., McCool, J., & Klanderman, D. (2012). Evaluating and improving a learning trajectory for linear measurement in elementary grades 2 and 3: A longitudinal study. *Mathematical Thinking and Learning, 14*, 28–54.

Battista, M. T. (2011). Conceptualizations and issues related to learning progressions, learning trajectories, and levels of sophistication. *The Mathematics Enthusiast, 8*, 507–569.

Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula." *Journal for Research in Mathematics Education, 38*, 35–70.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*, 1–78.

Common Core State Standards Initiative. (2010). *Common core state standards for mathematics.* Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.

Daro, P., Mosher, F. A., & Corcoran, T. (2011, January). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (CPRE Research Report #RR-68). Philadelphia, PA: Consortium for Policy Research in Education.

Delgado, C. (2009). *Development of a research-based learning progression for middle school through undergraduate students' conceptual understanding of size and scale* (Doctoral dissertation). Available through the Deep Blue database (http://deepblue.lib.umich.edu)

Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change, 1*, 5–28.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259–281.

Gravemeijer, K., & Cobb, P. (2006). Design research from a learning design perspective. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 17–51). New York, NY: Routledge.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*, 635–679.

Mohan, L., & Anderson, C. W. (2009, June). *Teaching experiments and the carbon cycle learning progression*. Paper presented at the Learning Progressions in Science (LeaPS) conference, Iowa City, IA.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

Moran, T. P., & Carroll, J. M. (1996). *Design rationale: Concepts, techniques, and use.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

National Assessment Governing Board. (2008, September). *Science framework for the 2009 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/publications/frameworks/science-09.pdf

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

National Research Council. (2006). *Systems for state science assessment*. Washington, DC: The National Academies Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: The National Academies Press.

National Research Council. (2011). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Penuel, W. R., Confrey, J., Maloney, A., & Rupp, A. A. (2011, April). *Design decisions in developing assessments of learning trajectories: A case study*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Popham, W. J. (2011a). *Classroom assessment: What teachers need to know* (6th ed.). Boston, MA: Pearson Education, Inc.

Popham, W. J. (2011b, April). *How to build learning progressions: Keep them simple, Simon!* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructive perspective. *Journal for Research in Mathematics Education, 26*, 114–145.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1–98.

APPENDIX A: LeaPS CONFERENCE PARTCIPANTS

The following individuals were formal participants in the LeaPS conference (held June 24–26, 2009 in Iowa City, IA).The leader for each strand is indicated with an asterisk (*).

Defining Strand

Martin Brock	Eastern Kentucky University
Tanya Dewey	University of Michigan
Nick Giordano	Purdue University
Kristin Gunckel	University of Arizona
Audrey Heppleston	McGill University
Gregory Louie	Gravelly Hill Middle School
Katherine McNeill	Boston College
Lindsey Mohan*	Michigan State University
Eun Jung Park	Northwestern University
Thomas Pentecost	University of Colorado
Elisabeth Roberts	University of Arizona
Aaron Rogat	CPRE, Teachers College, Columbia University
Ivan Salinas	University of Arizona
Christina Schwarz	Michigan State University
David Sederberg	Purdue University
Eulsun Seung	Indiana State University
Richard Shavelson	Stanford University
Tiffany-Rose Sikorski	University of Maryland
Carol Smith	University of Massachusetts Boston
Shawn Stevens	University of Michigan
Su Swarat	Northwestern University
Sarah Talle	Lakewood High School
Mary Carolyn Thies	University of Maryland
Roger Tobin	Tufts University
Emmett Wright	Kansas State University

Assessing Strand

Alicia Alonzo*	University of Iowa
Nathan Balasubramanian	Centennial BOCES
Lea Bullard	University of Michigan
David Carraher	TERC
Karen Draney	University of California Berkeley
Janice Earle	National Science Foundation
Sandra Enger	University of Alabama, Huntsville

Andrew Galpern	University of California Berkeley
Daryl Greenfield	University of Miami
Cari Herman Abell	AAAS Project 2061
Hayat Hokayem	Michigan State University
Hui Jin	Michigan State University
Roy Levy	Arizona State University
Greg Light	Northwestern University
Kathy Long	Lawrence Hall of Science
Joi Merritt	University of Michigan
Teresa Neidorf	American Institutes for Research
Paul Nichols	Pearson
Kathy Paget	EDC
Ed Roeber	Michigan State University
Maria Araceli Ruiz-Primo	University of Colorado Denver
Ursula Sexton	WestEd
Nancy Songer	University of Michigan
Tina Stanford	SRI International
Leah Walker	University of California Berkeley
Patti West	Cisco Systems

Modeling Strand

Derek Briggs*	University of Colorado Boulder
Kristen DiCerbo	Cisco Learning Institute
Robert Mislevy	University of Maryland
Fritz Mosher	CPRE, Teachers College, Columbia University
André Rupp	University of Maryland
Shinping Tsai	University of Washington
Mark Wilson	University of California Berkeley

Using Strand

Andrés Acher	Northwestern University
Beth Basista	Wayne State University
Leema Berland	University of Texas, Austin
Janet Coffey	University of Maryland
Seth Corrigan	Lawrence Hall of Science
Susan Doubler	TERC
Erin Furtak	University of Colorado Boulder
Amelia Gotwals*	Michigan State University
Christopher Harris	SRI International
Karin Hess	National Center for the Improvement of Educational Assessment
Michael Jabot	SUNY Fredonia

Bruce Johnson	University of Arizona
Diana Johnson	Lewis County Schools
Suzanna (Suzy) Loper	Lawrence Hall of Science
Jeff Orasky	University of South Florida
Julia Plummer	Arcadia University
Brian Reiser	Northwestern University
Namsoo Shin	University of Michigan
Pier Sun Ho	ConnectEd
Jessica Thompson	University of Washington
James Tomlin	Wright State University
Mark Windschitl	University of Washington
Marianne Wiser	Clark University
Yue (Melody) Yin	University of Illinois, Chicago

APPENDIX B: CHAPTER REVIEWERS

The following individuals provided one or more chapter reviews. We are grateful for their insights and feedback to the authors.

Charles W. (Andy) Anderson	Michigan State University
Jacqueline Barber	Lawrence Hall of Science
Derek Briggs	University of Colorado Boulder
Nathaniel Brown	Indiana University
Britte Haugan Cheng	SRI International
Janet Coffey	University of Maryland
Jimmy de la Torre	Rutgers University
Cesar Delgado	University of Texas at Austin
Sue Doubler	TERC
Karen Draney	University of California Berkeley
Ravit Golan Duncan	Rutgers University
Richard A. Duschl	Pennsylvania State University
Erin Marie Furtak	University of Colorado Boulder
Kristin L. Gunckel	University of Arizona
Karin Hess	National Center for the Improvement of Educational Assessment
Kristen Huff	The College Board
Lisa Kenyon	Wright State University
Katherine McNeill	Boston College
Robert Mislevy	University of Maryland
Lindsey Mohan	Michigan State University
Fritz Mosher	CPRE, Teachers College, Columbia University
Jonathan Osborne	Stanford University
Anthony Petrosino	University of Texas at Austin
Julia Plummer	Arcadia University
Ann Rivet	Teachers College, Columbia University
Ed Roeber	Michigan State University
Aaron Rogat	CPRE, Teachers College, Columbia University
Jo Ellen Roseman	American Association for the Advancement of Science
Maria Araceli Ruiz-Primo	University of Colorado Denver
Leona Schauble	Peabody College of Vanderbilt University
Carol Smith	University of Massachusetts Boston
Shawn Stevens	University of Michigan

APPENDIX B

Jessica Thompson	University of Washington
Mark Wilson	University of California Berkeley
Mark Windschitl	University of Washington
Yue (Melody) Yin	University of Illinois, Chicago