Dirk Ifenthaler
Pablo Pirnay-Dummer
Norbert M. Seel
Editors

Computer-Based Diagnostics and Systematic Analysis of Knowledge

Springer

Computer-Based Diagnostics and Systematic Analysis of Knowledge

Dirk Ifenthaler · Pablo Pirnay-Dummer ·
Norbert M. Seel
Editors

Computer-Based Diagnostics and Systematic Analysis of Knowledge

Springer

Editors
Dirk Ifenthaler
Universität Freiburg
Abt. Lernforschung
Rempartstr. 11
79085 Freiburg
Germany
ifenthaler@ezw.uni-freiburg.de

Pablo Pirnay-Dummer
Universität Freiburg
Abt. Lernforschung
Rempartstr. 11
79085 Freiburg
Germany
pablo@pirnay-dummer.de

Norbert M. Seel
Universität Freiburg
Inst. Erziehungswissenschaft
79085 Freiburg
Germany
norbert.seel@ezw.uni-freiburg.de

ISBN 978-1-4419-5661-3 e-ISBN 978-1-4419-5662-0
DOI 10.1007/978-1-4419-5662-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009943821

Preface

As instructional psychology is becoming more specialized and complex and technology is offering more and more possibilities for gathering data, instructional researchers are faced with the challenge of processing vast amounts of data. Yet the more complex our understanding of the field of learning and instruction becomes and the more our theories advance, the more pronounced is the need to apply the structures of the theories to sufficiently advanced methodology in order to keep pace with theory development and theory testing. In addition to obtaining a good fit between theory and diagnostics, this task entails making the methodology and tools feasible (easy to use and easy to interpret). Otherwise, the methodologies will only be used by their developers. The development of useful systems has always been a goal for scientists and engineers serving professional communities in the fields of instructional design and instructional systems development.

The progress of computer technology has enabled researchers to adopt methods from artificial intelligence, graph theory, feature analysis, feature tracking, and applied statistics and to use computers to implement computer-based instructional systems. Researchers have now also succeeded in developing more effective tools for the assessment of knowledge in order to enhance the learning performance of students.

The editorial committee has selected a wide range of internationally known distinguished researchers who present innovative work in the areas of educational diagnostics and the learning sciences. The audience for this volume includes professors, students, and professional practitioners in the general areas of educational psychology and instructional technology. Accordingly, the four parts of this book resemble a complete transfer from theoretical foundations to practical application. The tools and their scope of use and practicability for assessment and descriptive and comparative analysis are introduced, tested, and critically discussed.

The book starts with contributions on the *elicitation of knowledge* and continues with *methods for the aggregation and classification of knowledge* and the *comparison and empirical testing of strategies*. It concludes with a diverse overview of best practice and transferable examples for the *application of results*.

Elicitation of Knowledge

The *first part* of the book is about the theoretical foundations and recent developments and tools for the investigation of knowledge. Without a sound theoretical basis, satisfactory research would not be possible due to the complex aspects of the knowledge construct. On the practical side, recent innovations provide many new opportunities for addressing knowledge empirically and, moreover, for complementing existing methods or even providing alternatives in many cases. The key focus in this part is on strategies for finding out what a person knows as opposed to finding out what he or she does not know (as is often the case in classical knowledge assessment and testing).

Aggregation and Classification of Knowledge

The *second part* concentrates on the aggregation and classification of the different kinds of data on knowledge. Additional integrated tools for assessment and analysis are also introduced. Some of the existing tools already have functionality for aggregation, while the interfaces of others can be used with them. Accordingly, the second part also describes data interfaces between different knowledge assessment tools.

Comparison and Empirical Testing of Strategies

Once the data on knowledge has been aggregated, different methods and tools for comparison are available for use, ranging from applied graph theory to computer linguistic models. Possibilities for comparing and empirically testing similarities and differences between individual and group knowledge go far beyond simple frequency measures. The *third part* of the book will help readers apply these methods to their research. Therefore, there will be an emphasis on the practical application of the methodologies and on the interpretation of the results.

Application of Results

The *fourth part* will help readers structure the results from their own research and apply them to their field. Best practice examples and basic interpretation patterns help orchestrate the findings in practice. Thus, emerging development perspectives for the fields in question are also introduced.

Acknowledgements

Without the assistance of several specialists, the editors would have been unable to prepare this volume for publication. We wish to thank our board of reviewers for its tremendous help with both reviewing the chapters and linguistic editing. Our thanks

also go to David Heyde for proofreading several chapters and Andreas Lachner for preparing the chapters to meet the guidelines for editorial style.

Freiburg, Germany

Dirk Ifenthaler, Pablo Pirnay-Dummer, and Norbert M. Seel

Contents

Contributors

Eva L. Baker CRESST, University of California, Los Angeles, CA, USA, baker@cse.ucla.edu

Andrew F. Chiarella Athabasca University, Athabasca, AB, Canada, andrewc@athabascau.ca

San-hui Sabrina Chuang Graduate Institute of Teaching Chinese as a Second Language, National Taiwan Normal University, Teipei, Taiwan, sanhuich@hotmail.com

Roy B. Clariana The Pennsylvania State University, University Park, PA, USA, rclariana@psu.edu

Trevor A. Cohen University of Texas, Austin, TX, USA, trevor.a.cohen@uth.tmc.edu

Vanessa P. Dennen Florida State University, Tallahassee, FL, USA, vdennen@fsu.edu

Oktay Donmez Florida State University, Tallahassee, FL, USA, od06c@fsu.edu

Athanasios Drigas NCSR Demokritos, Institute of Informatics and Telecommunications, Net Media Lab, Paraskevi, Greece, dr@iit.demokritos.gr

John Epling State University of New York – Upstate Medical University, Syracuse, NY, USA, eplingj@upstate.edu

Kostas Hrissagis-Chrysagis CERETETH, University of Thessaly, Thessaly, Greece; Innovalia, Chios, Greece, kostas.chrysagis@gmail.com

Dirk Ifenthaler Albert-Ludwigs-University of Freiburg, Freiburg, Germany, ifenthaler@ezw.uni-freiburg.de

Allan C. Jeong Florida State University, Tallahassee, FL, USA, jeong@mail.coe.fsu.edu

Tristan E. Johnson Florida State University, Tallahassee, FL, USA, tejohnson@fsu.edu

Mohammed K. Khalil University of Central Florida, Orlando, FL, USA, mkhalil@mail.ucf.edu

Yoon-Jeon Kim Florida State University, Tallahassee, FL, USA, yk06c@fsu.edu

Tiffany A. Koszalka Syracuse University, Syracuse, NY, USA, takoszal@syr.edu

Susanne P. Lajoie McGill University, Montreal, QC, Canada, susanne.lajoie@mcgill.ca

YoungMin Lee Sookmyung Woman's University, Seoul, South Korea, ymlee@sookmyung.ac.kr

Apostolos P. Leros TEI of Chalkida, Chalkida, Greece, lerosapostolos@gmail.com

Iskandaria Masduki Florida State University, Tallahassee, FL, USA, iskandaria@gmail.com

Anne Mendenhall Florida State University, Tallahassee, FL, USA, anne.mendenhall@gmail.com

Harold F. O'Neil CRESST, University of Southern California, Los Angeles, CA, USA, honeil@usc.edu

Pablo Pirnay-Dummer Albert-Ludwigs-University of Freiburg, Freiburg, Germany, pablo@pirnay-dummer.de

Roger W. Schvaneveldt Arizona State University, Mesa, AZ, USA, schvan@asu.edu

Norbert M. Seel Albert-Ludwigs-University of Freiburg, Freiburg, Germany, seel@ezw.uni-freiburg.de

Valerie J. Shute Florida State University, Tallahassee, FL, USA, shute@mail.coe.fsu.edu

Eric G. Sikorski Florida State University, Tallahassee, FL, USA, egs04@fsu.edu

J. Michael Spector Learning and Performance Support Laboratory, University of Georgia, Athens, GA, USA, mspector@uga.edu

Anna Strasser Institute of Philosophy, Humboldt-University, Berlin, Germany, anna.strasser@t-online.de

Peter Tittmann Hochschule Mittweida, Mittweida, Germany, peter@htwm.de

John Vrettaros NCSR Demokritos, Institute of Informatics and Telecommunications, Net Media Lab, Paraskevi, Greece, jvr@iit.demokritos.gr

Chen-Yen Wang Florida State University, Tallahassee, FL, USA, cw06k@fsu.edu

Michael Yacci Rochester Institute of Technology, Rochester, NY, USA, may@it.rit.edu

About the Authors

Eva L. Baker (baker@cse.ucla.edu) is a Distinguished Professor of Education at UCLA. She directs the Center for Research on Evaluation, Standards, and Student Testing (CRESST), an internationally recognized R&D organization in both assessment and technology research. It is supported by government and private sources. Dr. Baker was recently the president of the American Educational Research Association and is currently a member of the National Academy of Education. She has an extensive publication record.

Andrew F. Chiarella (andrewc@athabascau.ca) is an Assistant Professor of Educational Psychology in the Centre for Psychology at Athabasca University, Canada. Dr. Chiarella's research interests include social software, distributed cognition, self-organising systems, and educational technology. In particular, his current research examines how social software can be used to create opportunities for the self-organisation of digitally evolving artifacts that scaffold learners.

San-hui Sabrina Chuang (sanhuich@hotmail.com) received her Ph.D. in Educational Psychology and Technology from the Rossier School of Education at the University of Southern California. Her instruction/research specialty is educational evaluation and assessment, focusing on the role of feedback. She has worked as a research consultant in planning and executing educational research at the UCLA Center for Research in Educational Standards and Student Testing (CRESST), and the University of Southern California. Currently, she works as a lecturer in East Asian Languages and Cultures at USC.

Roy B. Clariana (rclariana@psu.edu) is a Professor in the Department of Learning and Performance Systems in the College of Education at the Pennsylvania State University, USA. Clariana's research interests include feedback in instruction, context effects on memory, and measuring knowledge structure (i.e., mental models). He developed computer-based shareware for the analysis of concept maps (ALA-Mapper) and of natural language text such as students' essays (ALA-Reader).

Trevor A. Cohen (trevor.a.cohen@uth.tmc.edu) is an Assistant Professor in the School of Health and Information Sciences at the University of Texas, Houston. Dr. Cohen's research interests include distributional semantics, knowledge discovery, information retrieval, and biomedical cognition.

Vanessa P. Dennen (vdennen@fsu.edu) is an Associate Professor of Instructional Systems in the Department of Educational Psychology and Learning Systems at Florida State University. At FSU she teaches courses on discourse analysis, program evaluation, learning theory, and instructional design. Her research focuses on online discourse, cognitive apprenticeship, and online communities of practice.

Oktay Donmez (od06c@fsu.edu) received his MA in Computer Education & Instructional Technology from Hacettepe University, Turkey in 2005. Currently, Oktay is a doctoral student in the Department of Educational Psychology and Learning Systems at Florida State University. Currently, Oktay is focusing on the assessment and modeling of systems thinking skill within an immersive game for middle school students.

Athanasios Drigas (dr@iit.demokritos.gr) is a Senior Researcher at IIT-NCSR Demokritos, Greece. He has been the Coordinator of Telecoms and the founder of Net Media Lab since 1996. From 1985 to 1999 he was operational manager of Greek Academic network. As the Coordinator of several International & National Projects, in the fields of ICTs – Telecoms, e-services (e-learning, e-psychology, e-government, e-inclusion, e-culture, e-business, etc.), he has published more than 200 international & national articles in ICTs, 7 books, 25 educational CD-ROMs, & several patents. He has been a member in several International & National committees for design and coordination Network & ICT services & activities, and also in several committees of international conferences & journals. He has also received several distinctions for his scientific work (articles, projects, patents).

John Epling (eplingj@upstate.edu) is an Associate Professor of Family Medicine and Public Health & Preventive Medicine at the State University of New York – Upstate Medical University in Syracuse, New York, USA. He teaches epidemiology, biostatistics, evidence-based medical decision-making, quality improvement and research methods to medical students. His research interests include practice-based clinical research, educational research, quality improvement methods, and clinical prevention.

Kostas Hrissagis-Chrysagis (kostas.chrysagis@gmail.com) has been Visiting Prof. at the University of Thessaly, Research Adviser at CERETETH and recently President of Innovalia. His research interests include Modeling & Simulation of Systems and Optimization under Uncertainty, Automation & Mechatronics. He has been involved for more than 18 years in Science, R&D of co-funded programs: His work has flourished in taking part in more than 70 Projects. Dr. Hrissagis has been in various EU Committees preparing or reviewing work-programs and projects. He has co-authored MARVEL–eLearning in Mechatronics, "Remote programming and configuration of a robotic system: A workplace oriented case study".

Dirk Ifenthaler (ifenthaler@ezw.uni-freiburg.de) is an Assistant Professor at the Department of Educational Science at the Albert-Ludwigs-University of Freiburg,

Germany. Dr. Ifenthaler's research interests focus on the learning-dependent progression of mental models, problem solving, decision making, situational awareness, and emotions. He developed an automated and computer-based methodology for the analysis of graphical and natural language representations (SMD Technology). Additionally, he developed components of course management software and an educational simulation game (SEsim – School Efficiency Simulation).

Allan C. Jeong (jeong@mail.coe.fsu.edu) is an Associate Professor of Instructional Systems in the Department of Educational Psychology and Learning Systems at Florida State University. He teaches courses on distance education, computer-supported collaborative learning, and courseware development. His research is focused on the development of tools and methods to measure and visualize the processes of learning and discourse in computer-supported learning environments.

Tristan E. Johnson (tejohnson@fsu.edu) is the Director of the International Center for Learning, Education, & Performance Systems and an Assistant Professor of Instructional Systems in the Learning Systems Institute & Department of Educational Psychology and Learning Systems at Florida State University. Dr. Johnson has several years of experience studying team cognition, team-based learning, measuring shared mental models, team assessment and diagnostics, and team interventions. He is also involved in the development of agent-based modeling and the creation of data-driven algorithms for modeling training and instructional effects within performance models.

Mohammed K. Khalil (mkhalil@mail.ucf.edu) is an Assistant Professor at the College of Medicine, University of Central Florida. Dr. Khalil research interest is in the area of learning and instructional technology. He is interested to advance medical education with innovative learning strategies. He conducts applied research on the area of technology integration in medical education and team learning. The overall goal of his research is to develop effective pedagogy that promotes student-centered learning and hence life-long learning.

Yoon-Jeon Kim (yk06c@fsu.edu) is a doctoral student in the Department of Educational Psychology and Learning Systems at Florida State University. She received her B.A. in Educational Technology from Ewha Woman University and M.S. in Instructional Systems from FSU. Her research interests focus on student modeling, educational data mining, and educational technology. She is currently modeling and developing an assessment for creative problem solving to be embedded in a game environment.

Tiffany A. Koszalka (takoszal@syr.edu) is an Associate Professor of Instructional Design, Development and Evaluation at Syracuse University in Syracuse New York, USA. She teaches graduate courses in the instructional sciences, learning theory, research methods, and educational technology integration. Her research interests are focused on understanding the relationships among learning, instruction, and technology that best facilitate different types of learning.

Susanne P. Lajoie (susanne.lajoie@mcgill.ca) is a James McGill Research Professor in the Department of Educational and Counselling Psychology at McGill University. She is a Fellow of the American Psychological Association and the American Educational Research Association. Dr. Lajoie designs technology-rich learning environments for educational and professional practices. She uses a cognitive approach to identify learning trajectories that help novice learners become more skilled in the areas of science, statistics, and medicine.

YoungMin Lee (ymlee@sookmyung.ac.kr) is an Assistant Professor at the Professional School of Human Resource Development for Women, Korea. Dr. Lee's research interests are on the problem solving, expertise, team learning, and performance technology. He has done research projects for government, international organizations, and NGOs.

Apostolos P. Leros (lerosapostolos@gmail.com) is an Associate Professor in the Department of Automation at the Technological Institute of Xalkis, 34400 Psahna, Xalkis, Evia, Greece. His interests concentrate on modelling, simulation and optimization of complex, multivariable, linear and nonlinear decentralized and distributed dynamical large-scale systems and on theory and applications of advanced and innovative technologies of neural networks, fuzzy logic and artificial intelligence.

Iskandaria Masduki (iskandaria@gmail.com) is a doctoral candidate in the Department of Educational Psychology and Learning Systems, Florida State University. As an Instructional Designer for the Center of Information Management and Educational Services, she is currently involved in research and training for public sector employees in Florida. Her research interests include evidence-centered design assessment, educational games, individual/team mental models, and human performance in the workplace.

Anne Mendenhall (anne.mendenhall@gmail.com) is an advanced doctoral student at Florida State University in Tallahassee, Florida. Ms. Mendenhall is a research assistant and project manager at the Learning Systems Institute. Her research areas include team-based learning, shared mental models, global collaborative learning, and open and distance learning in developing countries.

Harold F. O'Neil (honeil@usc.edu) is a Full Professor of Educational Psychology and Technology at the University of Southern California. He is a Fellow of the American Psychological Association, the American Educational Research Association, and a Certified Performance Technologist. A prolific writer, he has recently co-edited four books – *What Works in Distance Learning: Guidelines* (2005), *Web-Based Learning: Theory, Research, and Practice* (2006), *Assessment of Problem Solving Using Simulations* (2008), and *Computer Games and Team and Adult Learning* (2008).

Pablo Pirnay-Dummer (pablo@pirnay-dummer.de) is an Assistant Professor at the Department of Educational Science at the Albert-Ludwigs-University of Freiburg, Germany. His research and publications are located in the area of

cognition, learning, and technology. He developed, implemented, and validated the language-oriented model assessment methodology MITOCAR (Model Inspection Trace of Concepts and Relations) which is built to assess, analyze, and compare individual and group models of expertise. Pirnay-Dummer also developed the web-based training software Empirix, including new approaches of automated evaluation and automated tasks synthesis algorithms.

Roger W. Schvaneveldt (schvan@asu.edu) is a Professor of Applied Psychology at Arizona State University, Polytechnic in Mesa, AZ. His best known work includes semantic priming with David Meyer and Pathfinder Network Scaling with Frank Durso and Don Dearholt. He has also investigated attention, memory, and implicit learning. He is currently working in aviation psychology, abductive reasoning, knowledge discovery, and network scaling.

Norbert M. Seel (seel@ezw.uni-freiburg.de) is chair of the Department of Educational Science at the Albert-Ludwigs-University of Freiburg, Germany. As a cognitive scientist, he is concerned with mental model research, instructional design, and media research. Dr. Seel's work is rooted in quantitative empirical research methods. He published several books and articles in these fields.

Valerie J. Shute (shute@mail.coe.fsu.edu) is an Associate Professor in the Department of Educational Psychology and Learning Systems at Florida State University where she teaches graduate students in the Instructional Systems Program. Before coming to FSU, Val was a principal research scientist at Educational Testing Service. An example of current research involves using immersive games with stealth assessment to support the acquisition of important key competencies (e.g., systems thinking, creative problem solving, perspective taking, and teamwork).

Eric G. Sikorski (egs04@fsu.edu) is a Project Manager with the Florida State University Learning Systems Institute. In this role, he has researched, designed, and developed training and other performance improvement solutions for the US Department of Defense. Dr. Sikorski's research focuses on team shared mental models (SMM) in the academic environment. The practical application of this research is to enhance student team SMM and ultimately performance through the introduction of knowledge-sharing interventions.

J. Michael Spector (mspector@uga.edu) is a Professor of Educational Psychology and Learning Systems and Associate Director of the Learning Systems Institute at Florida State University. His research pertains to learning in complex domains and has focused in recent years on how to assess progress of learning with regard to ill-structured problem solving; he also conducts research on the design of technology facilitated learning environments.

Anna Strasser (anna.strasser@t-online.de) is working as a researcher at the Berlin Mind and Brain School/Institute of Philosophy at the Humboldt-University Berlin, Germany. In her Ph.D. she focused on action theories and artificial systems. The

present research project of Dr. Strasser is on necessary and sufficient conditions of the development of self-consciousness.

Peter Tittmann (peter@htwm.de) is a Professor of Mathematics at the University of Applied Sciences Mittweida, Germany. He researches in graph theory and enumerative combinatorics. His current research interests include network reliability analysis and mathematical models for social networks.

John Vrettaros (jvr@iit.demokritos.gr) is a Professor of physics and informatics and a pedagogist. He is an associate researcher at the Institute of Informatics & Telecommunications of N.C.S.R. Demokritos, Greece and General secretariat of adult education. He is the coordinator of e-learning projects within Net Media Lab and has published more than 40 international & national articles in ICTs, 4 books, 25 educational CD-ROMs, and several patents.

Chen-Yen Wang (cw06k@fsu.edu) is a doctoral student in the Instructional Systems program at Florida State University. His primary research interests surround collaborative learning and game-based assessments. He is currently engaged in the development and modeling of an assessment on perspective taking. He plans to embed this stealth perspective-taking assessment within an immersive, collaborative game.

Michael Yacci (may@it.rit.edu) is a Professor in Information Sciences and Technology at Rochester Institute of Technology in Rochester, NY, USA. Dr. Yacci's research interests are in the areas of computer-based instructional design, computer-human interaction, evaluation, accelerated learning, and intelligent agents. He has recently published a model for interactivity in computer-based learning.

Reviewers

Alexa Breuing University Bielefeld, Bielefeld, Germany, abreuing@techfak.uni-bielefeld.de

Martin Brösamle Albert-Ludwigs-University, Freiburg, Germany, martinb@cognition.uni-freiburg.de

Simon J. Büchner Albert-Ludwigs-University, Freiburg, Germany, buechner@cognition.uni-freiburg.de

Ian Douglas Florida State University, Tallahassee, FL, USA, idouglas@lpg.fsu.edu

Dirk Ifenthaler Albert-Ludwigs-University, Freiburg, Germany, ifenthaler@ezw.uni-freiburg.de

Pablo Pirnay-Dummer Albert-Ludwigs-University, Freiburg, Germany, pablo@pirnay-dummer.de

Norbert M. Seel Albert-Ludwigs-University, Freiburg, Germany, seel@ezw.uni-freiburg.de

Anna Strasser Humboldt-University, Berlin, Germany, anna.strasser@t-online.de

Kay Wijekumar The Pennsylvania State University, Beaver, PA, USA, kxw190@psu.edu

Part I
Elicitation of Knowledge

Intermezzo 1 – To Be Moved by Knowledge: Moving Knowledge Moves Knowledge About Knowing

Pablo Pirnay-Dummer and Dirk Ifenthaler

As one of the most central aspects of the learning sciences, cognitive psychology, and technologies, knowledge is also one of the most complex psychological constructs to address. But why are theories, applications, research methods, and interpretations of results always so diverse or even sometimes inconclusive? The reason can be found in the construct itself, which is both large and complex in nature. Maybe it is even the most complex field of all to investigate. There are surely *things* which are more complex in themselves. But *investigating* knowledge means using knowledge models to investigate knowledge models – which leads to an infinite recursion. The investigation may not even be possible to conduct. However, a position like this does not suit empiricists well. So, we will skip it for now. Aside from the fact that knowledge is internal and cannot be observed directly, four different aspects play an almost equal part in the complexity of describing knowledge – and they interact dynamically. One lies in a conceptual change in the field of cognition, one in the understanding of what learning is, one in better insight into dynamics and complexity, and last but not least, one may be described as a methodological opportunity: New technologies now allow us to come closer to the construct itself – by coming from as many "directions" as possible. The influence and, to be more precise, interconnectedness of knowledge, decision making, and problem solving is formidable and the differences between classic categories, e.g., between procedural and declarative kinds of knowledge entities, converge in the face of new theoretical models. Thus, the concept of knowledge – and even more of knowing – has come a long way from a static set of entities to processes containing construction, composition, and dynamic change, most of it highly individual. As the construct changed in this way, its contingencies for learning also evolved. From a simplified perspective, knowledge was once the start and the goal of learning while learning was the transition in between those states. Now the process models of knowledge show that the individual changes and transitions between the internal "world," which we call "the mind," and the external world, which we, well, at least do not call "the mind," bring knowledge far closer to the process of learning and a little further away from memory. In the end, this also leads to the simple conclusion that learning does

not have an endpoint at all. And this again may be one of the most important theoretical rediscoveries within the field of learning and instruction in the twentieth century. Only if knowledge can be dynamic, complex, and flexible – and therefore also heuristic – may it help us navigate complex and dynamic worlds. Thankfully, the world is not at all well-structured, well-defined, or in any way linear. Keeping all this in mind, it seems only logical that knowledge will have to be addressed in different methodological ways. Despite the fact that knowledge is only very rarely still approached with a classical computer metaphor and thus "stored" and "retrieved," such misconceptions are still common in all kinds of fields of "attached" research which use the construct in one way or another. In such cases, a strong bias holds the construct theoretically close to memory, and knowledge is often considered to be operationalizable by means of memory performance alone. There is, of course, no doubt about the interconnection between knowledge and memory. However, they are different constructs and should be measured differently. For pragmatic reasons, knowledge is classically addressed by measures of its absence: If a subject does not exhibit the expected performance, then he or she lacks the knowledge to produce this certain behavior (e.g., to give an answer to a question). In other words, the individual performance which cannot be produced determines whether somebody supposedly possesses sufficient knowledge. The same paradigm holds for the contrasting approach in studies of expertise: All (!) performance which an expert is able to exhibit and nonexperts cannot produce is considered to be part of the specific expertise in that field and is therefore very often identified as a suitable learning goal for novices. The identification of this gap is also often referred to as "what the learner does not know." Gap-oriented assessments may provide an idea about what someone else may need in general, but for the learner (needs assessment) it does not really cover what is already there – within his or her mind. Unsurprisingly, this is one of the key problems in knowledge management. On the other hand, there is no doubt that prior knowledge has a great (if not the main) effect on learning and understanding. If we really want to find out what somebody knows, then we always need methodology which addresses *available* behavior as opposed to comparing gaps standing in the way of *desired* behavior. New and deeper understanding about the construct of knowledge and its part in all cognitive processes and new methodological and technological advancements allow newer and better approaches to general and specific knowledge. In the first part of this book, the authors provide multiple perspectives on new ways to assess and analyze knowledge. The part discusses the limits and the accessibility of representations and the fundamental functions of externalization, which we sometimes call re-representations: They are external representations of internal representations of objects and relations in the world. Both practitioners and researchers may focus their knowledge about knowledge on knowledge.

Chapter 1
Essentials of Computer-Based Diagnostics of Learning and Cognition

Norbert M. Seel

1.1 Introduction

As a result of the rapid progress of computer technology in recent decades, researchers from different areas have adopted artificial intelligence to develop computer-aided instruction systems and diagnostic tools for the assessment of learning and cognition. In recent years, computational intelligence (CI) has been used in this special area to provide solutions or alternative methodologies. One of the major research areas of CI is the modeling of human problem solving and decision making in complex domains. Some of the components of this instructional field are the acquisition of declarative and procedural knowledge, strategies of decision making under conditions of uncertainty, the introduction of new knowledge or the modification of previous knowledge in order to enable the construction of effective mental models, and, especially, the development of effective user interfaces that incorporate assessment and measurement tools. If instructional systems in the area of telematics are to provide effective environments for open and flexible learning, they must also contain a service to assess the users' predispositions and capacities to learn. Therefore, diagnosis is an important component of Intelligent Tutoring Systems (ITS), for instance to carry out an efficient and effective diagnosis of the progress of learning and interrelate it with other components of the learning environment for flexible and self-regulated learning.

Open and flexible learning with telematic systems presupposes a system-inherent analysis and feedback of the individual's learning progression in order to continuously adapt the learning environment to the relevant learner characteristics and the "responses" to the learning tasks. This means that if telematic systems are to provide appropriate environments for truly open and flexible learning, they must contain components to assess both the learners' predispositions to learn and the progression of their learning. In addition, there is a need for an adaptive tutoring component which regulates the necessary instructional interactions with the students depending

N.M. Seel (✉)
Albert-Ludwigs-University of Freiburg, Freiburg, Germany
e-mail: seel@ezw.uni-freiburg.de

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_1,

on their capability to deal effectively with the learning tasks. To be instructionally effective, this adaptive tutoring component must be able to detect when the learner's skill performance or domain-specific knowledge is not sufficient to meet the task requirements so that it can make appropriate adjustments.

As cognitive and educational psychology are becoming more specialized and complex while technology has been offering more and more possibilities for gathering data during learning, researchers are faced with the challenge of processing vast amounts of low-level data. Moreover, as the example of computer-adaptive testing demonstrates, the use of computers in the field of educational diagnosis has become a standard procedure in recent decades (cf. Hambleton & Zaal, 1991). Nevertheless, the systematic integration of diagnosis into multimedia-learning environments and settings for online learning is still in its infancy, although multimedia principally provides splendid opportunities for the assessment of cognitive skills: Multimedia-learning environments can provide test information in a variety of modalities, they can provide situational contexts for test items, and they allow multiple paths through the learners' knowledge. Furthermore, multimedia-based assessment techniques may allow learners to use the formats best suited to their learning styles, ability levels, and information needs. Therefore, multimedia-based assessment techniques provide many advantages to the learner, especially by virtue of their possibilities for adapting to individual differences and allowing the learner to control the paths of study. This largely corresponds with the assumption that multimedia-based learning environments in general can provide customized interfaces with varying levels of guidance that may increase the learners' engagement with the learning situation as they elaborate on current knowledge. However, to date, there is still a considerable gap between available computer technology and its use for an efficient and effective assessment embedded in telematic learning systems (Lajoie, 2000; Liu, Chiang, Chen, & Huang, 2007).

Therefore, we often can find a continuation of traditional and conventional measurements and assessments in the area of multimedia development and research. Often the following method is applied: After a pretest and a learning phase, which is sometimes experimentally varied, the subjects have to perform specific tasks considered indicative of successful learning. This involves a great variety of assessment procedures and measures, including traditional tests for assessing the retention of content to be learned, tasks for transfer and qualitative reasoning, questionnaires and ratings, assessment of frequency and type of errors, and other measurements such as eye fixations during task accomplishment and the time needed for learning or accomplishing test items. However, these conventional testing methods simply give students a score and do not provide them the opportunity to learn adaptively how to improve their learning performance when operating with a telematic learning system.

This volume provides the reader with some alternate computer-based solutions for the assessment of knowledge and cognition necessary to successfully perform multimedia and online learning. Consequently, the majority of the contributions center around the questions "What is knowledge?" and "How can we assess knowledge?" The following sections of this introductory chapter will focus on some

essentials of computer-based diagnostics of knowledge and cognition. First, some basic ideas of educational diagnostics and diagnoses are described, resulting in a distinction between "responsive" and "constructive" approaches to knowledge assessment. In the subsequent sections, computer-based procedures are described with regard to both approaches. They presuppose the application of external representations grounded on the semantics of natural language. The next section of this introduction focuses on computer-based and agent-based methodologies for knowledge diagnosis as a central component of automatic diagnostic systems. The final section will provide a brief preview of the major topics of this volume.

1.2 Diagnostics and Diagnosis in The Area Of Education and Instruction

When we speak about knowledge, we are speaking about a theoretical construct, i.e., something which we cannot observe but which we can measure on the basis of observable behavior or verbal statements made in the course of solving cognitive tasks. *Diagnosis* generally refers to collecting and interpreting information with the aim of determining which of a set of non-observable states is the "true state of nature." Of course, it is not possible to influence or determine what state of nature will occur; what we can do as researchers or instructors is to collect and process information in order to arrive at a probabilistic estimation of the true state. A good example of this probabilistic estimation of a state of nature is the ability to estimate through test-based measurement (van der Linden, 1991). To capture this idea, consider H as a set of possible states of nature whose specific realizations cannot be observed directly. In medical diagnosis, H represents the set of illnesses that a patient might have. Since physicians can assume mutually exclusive and exhaustive states, they should be able to attribute exactly one illness, H^*, to the patient. However, usually the doctor does not know which illness the patient actually suffers from but does have some ideas or hypotheses about the probabilities of different illnesses which we call "prior probabilities" $p_0(H_i)$. Clearly, these "prior probabilities" are not unconstrained but rather influenced by prior knowledge and evidence. For example, an experienced physician will ascertain a bone fracture on the basis of evidence. However, often this evidence will not be sufficient. Therefore, the physician will carry out further examinations, such as X-rays.

This idea of diagnosis can be transferred easily to the area of psychology and education, where there is much demand for diagnosis. In general, the ultimate goal of instruction is to initiate, facilitate, and guide learning processes. However, from a cognitive perspective, learning is considered a change in mental states (involving processes like accretion, tuning, and restructuring of knowledge; see Rumelhart & Norman, 1978) which is unobservable. Therefore, the instructional psychologist or educator is in a situation similar to the physician: The true mental states of learners will never be directly accessible, but rather the psychologist or educator must operate on the basis of subjective hypotheses about the students' mental states.

Both in medical and psychological diagnosis, there are sources of information which provide data D that might be used to modify the initial hypotheses about possible states of nature. The processing of D succeeds in transforming the "prior probabilities" into "posterior probabilities," $p_{\mathrm{m}}(H_{\mathrm{i}})$. Remaining within the example of medical diagnosis, instances of D would be blood pressure, appetite, special aches, etc. Usually, the values of such variables are obtained from the patient through tests or interviews. The more data the physician obtains through examinations and tests, the more likely a correct diagnosis will result. Basically, the same holds true for the field of psychology and instruction.

Thus, we can conclude that the process of proceeding from a relatively diffuse prior probability distribution $p_{\mathrm{o}}(H_{\mathrm{i}})$ to a more informative posterior probability distribution $p_{\mathrm{m}}(H_{\mathrm{i}})$ is also the essence of diagnosis in the field of instruction. As in the case of medical diagnosis, this transformation from $p_{\mathrm{o}}(H_{\mathrm{i}})$ to $p_{\mathrm{m}}(H_{\mathrm{i}})$ is not unconstrained. In the context of instruction, for instance, learner characteristics (abilities and skills), organizational conditions, and the curriculum may be important constraints. Informally, knowledge is information about some domain or subject area or about how to do something. Humans require and use a lot of knowledge to carry out even the most simple commonsense tasks. There are many knowledge-intensive tasks at which humans excel, such as recognizing faces in a picture, understanding natural language, or following legal argumentation. Indeed, there is an abundance of methods and procedures for the assessment of knowledge and cognition. In accordance with psychometric conventions, a basic distinction can be made between

- Responsive (reactive) diagnosis and
- Constructive (nonreactive) agent-based diagnosis.

1.3 Responsive Methods of Measurement and Assessment

Theoretical constructs like knowledge, learning, and cognition are often difficult to talk about because they are not clear-cut or concrete. Knowledge, for instance, is a fairly abstract concept, and a number of theories have been developed to explain how people store information in such a way that it can be retrieved when needed. Although these theories describe the basic function of memory as enabling the retention of information and personal experiences and the recall of that information and those experiences, there is no universally accepted model of human memory (Seel, 2008). Rather, models of human memory are often based on commonsense assumptions about how information is processed and stored. Moreover, the related experimental research is regularly grounded on the assumption that it is possible to infer the quality of mental states and cognitive processes from performance in specific tasks. In order to illustrate this, I will refer now to the measurement of reaction times, which can be realized easily with computers.

Since the early nineteenth century, psychologists have thought that human reaction time provides clues about mental processes and the organization of the mind (Brebner & Welford, 1980). Following the essay "On the speed of mental processes"

by Donders (1868/1969), psychologists measured the time required by subjects to perform various more or less complex tasks. These reaction times and the changes in RT under different experimental manipulations have been used as evidence for or against models of cognitive architecture and for testing hypotheses about processes and structures of the human mind. For example, measurements of RT have been used to distinguish between serial and parallel processing. Furthermore, the RT methodology has also been applied to other research issues, such as attention control, information flow, the acquisition of skills, and so on. Clearly, measurements of RT can also be applied to the investigation of automatic versus controlled information processing. Automatic processing is fast, not conscious, but rigid, requires almost no resources or attention, and can be performed parallel to other activities. Automation follows frequent, consistent practice and is based on *schemata*. It is the activation of schemata that allows automatic processing and thus minimizes the load of working memory. This argumentation is at the heart of cognitive load theory, which stresses that *schemata* must be activated in order to bypass the limitations of working memory. Skilled performance develops through the construction of an increasing number of ever more complex and abstract schemata (Sweller, 1994) that allow automatic information processing (cf. Clark, 2006). The notion of a working memory refers to computational mechanisms that maintain and provide access to information (= knowledge) to be retrieved during the performance of a task. Any computational system must support such functionality because computation is inherently a process that requires the temporary storage of information. However, this does not imply that schemata are considered as explicit entities; rather, they are implicit in our knowledge and are created by our interaction with the environment. Rumelhart, Smolensky, McClelland, and Hinton (1986) state that nothing stored corresponds closely to a schema. What is stored in memory is a set of connection strengths which, when activated, have implicitly in them the ability to generate states that correspond to an instantiated schema. If we accept this argumentation, reaction times can be considered as indicators for the instantiation of a schema which emerges at the moment it is needed to accomplish a cognitive task.

With regard to reactive measurements (reaction time and conventional multiple-choice tests), the most understated risk for a valid interpretation is the error produced by the respondent in accomplishing the tasks. Even when the subject is well intentioned and cooperative, several errors must be taken into account which reduce the reliability and validity of the measure: awareness of being tested, role selection, measurement as a change agent, and response sets (cf. Overman, 1988). However, what is considered a risk for responsive measurements can be considered an advantage for the nonreactive procedures of knowledge diagnosis.

1.4 Constructive Methods of Knowledge Diagnosis

Virtually all learning takes place by talk and text. Accordingly, discourse is an eminently important mediator of learning and thinking (especially in the context of instruction) and language can be considered as one of the most important windows

to the mind. Verbal communication and discourse is what subjects use to mediate their ideas, thoughts, feelings, and knowledge. Therefore, individual knowledge regularly becomes accessible on the basis of verbal communication. Consequently, the knowledge that is manifest in an organization or in society is also usually transmitted by language and discourse. Once someone has expressed something in a language, reasoning about it is symbol manipulation: Knowledge is then "the symbolic representation of aspects of some named universe of discourse" (Frost, 1986, p. 11). This universe may be the actual or a fictional world, one now or in the future, or one which only exists in someone's beliefs.

Taking into account the extraordinary importance of language for human communication about knowledge, thoughts, and cognitive artifacts such as mental models, it is nearly self-evident that different methods of verbalization play a central role in the diagnosis of individual knowledge. Indeed, psychologists and educational researchers often consider verbalizations and think-aloud protocols as appropriate qualitative methods to assess mental states. Besides this "direct communication" of thoughts and ideas by means of verbalizations, more extensive verbal explanations, inferences, hypotheses, speculations, and justifications are considered effective means to assess knowledge and cognitive artifacts. In spite of their indisputable ecological validity, verbal data and protocols have been criticized by some authors (e.g., Nisbett & Wilson, 1977) for their deficiencies with regard to psychometric standards of reliability and validity. It is not the place here to discuss this issue in more detail (for more details, see Seel, 1999), and it is not necessary either since direct verbalizations and natural discourse transcend the possibilities of computers we regularly work with in the field of instruction. Actually, at the moment any computer-based analysis of verbal protocols and explanations would go far beyond what we can do with conventional PCs. However, there are several suitable computer-based assessment approaches which presuppose the use of physical representations, discussed in terms of externalization of knowledge by means of specific representational formats. Such approaches may be based on the semantics of natural language (Helbig, 2006) or the measurement of associative strengths between concepts (see, e.g., the chapters by Schvaneveldt & Cohen; Pirnay-Dummer & Ifenthaler in this volume).

1.5 The Role Of External Representations

An important side effect of implementing computer technologies in instructional settings which is closely related to the computer-based measurement of reaction times in accomplishing cognitive tasks consists of the computer's potential to assess "online" protocols of learning tasks. However, it is not easy to analyze and interpret the resulting low-level data (Seel, 1995). Alternatively, computer-based approaches of cognitive modeling which use specific external representations are being considered increasingly as sound methods for assessing conceptions, ideas, and thoughts about particular subject domains.

Rumelhart et al. (1986) suggest that *external representations* play a crucial role in thought since our experience with them involves imagination but we can solve them mentally. Thus, the idea that we reason with mental models is a powerful one precisely because it is about this process of imagining an external representation and operating on it. Most of what we know is based on our experience developing and refining external representations for particular things and events.

Mislevy et al. (2007) have pointed out that several properties of external representations are highly relevant for assessment purposes. One property is that an external representation does not include everything that can be represented about a subject but rather only certain facts or entities and relationships between them. Mislevy et al. call this the *ontology of the external representation*, whereas I consider external representations as externalizations of mental models that highlight relevant entities and relationships and allow us to think about, talk about, and work with them. The velocity of a falling body is represented by the (mathematical model or) expression $v0 + g\ t$, no matter whether the body is a cannonball or a feather, whether it is falling in Austin or Tokyo. Constructing an external representation does not imply that all attributes of the object will be represented, but rather that some attributes will be considered irrelevant and will therefore not be included in the external representation.

In accordance with Goodman (1968), it can be argued that the act of representing something implies that an individual takes one object intentionally as a specific sign in order to represent another object as something. This means a sign is selected as mediator from a certain repertoire of signs that is accessible and relevant. Any sign which is taken intentionally for representation purposes is a triadic schema that involves a representing medium M, a represented object O, and an explanatory and contextual interpretation I or "meaning" of the object (cf. Seel & Winn, 1997). From a semiotic point of view, external representations are a means of communicating knowledge that involves a transmission of signs distinguished by the object-related entities "index," "icon," and "symbol." Accordingly, Aebli states that "when I want to represent a fact, an experience, or a structural relationship, I have to realize it by acting, perceiving, imagining, or speaking: Otherwise, it cannot exist in my mind" (Aebli, 1981, p. 290). In consequence, it can be argued that individual knowledge only can be assessed if it is communicated by means of actions, pictures, and/or language.

Based on this assumption and in accordance with Bruner's (1990) precept that the interpretative paradigm of symbolic interactionism focuses "on the *symbolic activities* that human beings employed in constructing and in making sense not only of the world, but of themselves," we can find different types (or formats) of representations in the field of cognitive psychology. Historically, mental representations have been interpreted by analogy with physical (or external) representations, or in other words, descriptions and classifications devised for physical representations have been applied to mental representations (Paivio, 1986). Physical representations can be picture-like or language-like. Similarly, we can find a distinction between mental images and propositions as major types of mental representations (Anderson, 1983; Markman, 1998) since the early days of cognitive psychology. Based on the

assumption of a continuous interplay between internalization and externalization, theorists made a distinction between a representation system and a communication system that share the use of picture-like and language-like formats of representation.

The *communication system* regularly maps external forms (such as speech sounds or pictorial signs) to meanings by means of internalization into the mental "space" in order to create meanings which are mostly but not exclusively related to external objects, events, or situations. A communication system is typically public, shared by many individuals, and thus presupposes the use and comprehension of shared communication modalities. The representation system may lack this immediate relatedness to the external world, but there would be no practical advantage in having a representation system which is not in some way related to the world outside of the mind processing it. Although there are parts and elements within a representation system that are not parts and elements within the communication system – and vice versa – both systems are so closely related with each other that it is often difficult to decide what is representation and what is communication. This especially holds true with respect to the assessment and diagnosis of knowledge.

There may exist so-called inert knowledge, i.e., knowledge which cannot be externalized at the moment it is needed, but in most cases it makes sense to assume that knowledge is not a sleeping copy of former experiences; rather, it is constructed at the moment it is needed to accomplish a task. This means that information stored in long-term memory is activated (retrieved) in order to meet the requirements of a particular situation. There is sufficient evidence from research on mental models and schema activation that supports this basic assumption of situation-dependent (re-)construction of knowledge (cf. Seel, Ifenthaler, & Pirnay-Dummer, 2009). We certainly do not process (and store) all of the information we perceive every day but rather only a small part of it. Many contents and processes represented in neural mechanisms are simply not retrievable and available to mental representations and thus to communication. They may be beyond awareness and inaccessible (Searle, 1992). Knowledge as we discuss it in this volume belongs to the conscious and accessible part of the human mind. It can be retrieved and communicated.

A basic assumption of cognitive psychology is that humans represent their knowledge by means of concepts (e.g., Prinz, 1983) that can be expressed through natural language. The central idea of this assumption is that the world can be described in terms of individuals (things) and relationships among them. Accordingly, in psychology a distinction has been made between within-concept relations and between-concept relations. This conception results in using concept maps and semantic networks as means of external representation of knowledge. The assumption that the world and its representations can be described in terms of conceptual structures is not a strong assumption but is rather grounded on the observation that individuals can be anything nameable, whether the named thing is concrete or abstract. Clearly, what is a "thing" is a concept of a person and a property of the world. Accordingly, for each domain or task, the specific individuals and relations must be identified in order to express what is true in the world or at least consistent with what we know or believe about the world (cf. Seel, 1991). This profoundly affects human ability to solve problems in a given domain.

In order to use knowledge and reason with it, intelligent agents (i.e., humans or robots) need a *representation and reasoning system*. Such a system is composed of a *language* to communicate and to assign meaning and procedures to compute answers in the language. The language can be a natural language, such as English, but in the context of assessing knowledge it is more convenient to use special tools like semantic networks and concept maps for the assessment of structural knowledge. Clearly, natural language is very expressive – probably everything can be expressed in natural language. However, semantic networks (and concept maps) better correspond to the idea of assessing structured knowledge which reflects the structure of the part of the world being represented.

Semantic networks and comparable tools (e.g., causal diagrams) for the assessment of knowledge provide the user with some important advantages: They make the relevant objects and relations explicit and expose natural constraints with regard to "causality," i.e., how an object or relation may influence another one. Semantic networks allow one to map the relevant structure at one time and to suppress irrelevant details; they are sufficiently transparent, concise, and computable. Therefore, semantic networks (and comparable tools) have emerged in recent years as the most important tools for diagnosing knowledge about the world (Gordon & Jupp, 1989). What is common to all semantic networks is a declarative graphic representation that can be used either to represent knowledge or to support automated systems for reasoning about knowledge. Some versions are highly informal, but others are formally defined systems of logic. Beyond the chapters of this volume which focus on the semantics of natural language for representational purposes, most of the contributions in the following chapters focus on semantic nets and their purposeful use for psychological and educational diagnostics. However, the use of semantic nets is not only restricted to the assessment of human knowledge, but also provides effective tools for Intelligent Tutorial Systems and Computational Intelligence, where the systematic assessment of knowledge is a requisite for adapting the system to individual needs and automating problem-solving tasks. Indeed, applications of computational intelligence are diverse at the moment. They include medical diagnosis, the scheduling of factory processes, robot learning within the realm of hazardous environments and natural language translation systems, and cooperative systems that presuppose intelligent reasoning and action.

1.6 Computer-Based and Agent-Based Knowledge Diagnosis

In assessing and/or predicting the impact of technological innovations in general and particularly on diagnostics, it is crucial to go beyond a narrow focus on intentional subject-matter learning to a broader examination of how utilizing telematic systems changes the content, processes, and contexts of the learning of intelligent agents in general. Clearly, any student is an intelligent agent, but an agent might also be a computational engine coupled with physical actuators and sensors, in other words a *robot*. It also might be an advice-providing computer (e.g., an expert system) coupled with a human who provides the perceptual task and accomplishes it. Finally, an

agent might even be a program which acts in a purely computational environment and presupposes an automatic diagnostic system (Sarbadhikari, 2004).

However, at all times the agent must be able to operate effectively with prior knowledge about the world, past experiences from which it can learn, goals to be achieved, observation about the current environment and interactions with it, and actions aiming at changes in the environment. In consequence, these sophisticated intelligent applications of telematic systems and computers presuppose a strong diagnostic component or subsystem which has to complete several assignments. It should be *diagnostic* inasmuch as it helps to assess bugs in the agent's knowledge and skill performance; corrective in order to erase those bugs; evaluative with regard to the agent's relevant needs, goals, and plans in particular learning situations; predictive with regard to the agent's likely responses to tutorial actions; and strategic in order to initiate significant changes in the agent's actions. To meet these requirements, a "diagnostic assistant" of the kind currently under discussion in connection with the future of the internet may be helpful. In this development, known as Web 3.0, the notion of the semantic web (and related solutions, such as semantic nets) plays a central role (Fig. 1.1).

Actually, it can be argued that Web 3.0 (discussed in terms of coupling the Internet with Artificial Intelligence) is nothing other than the realization and further

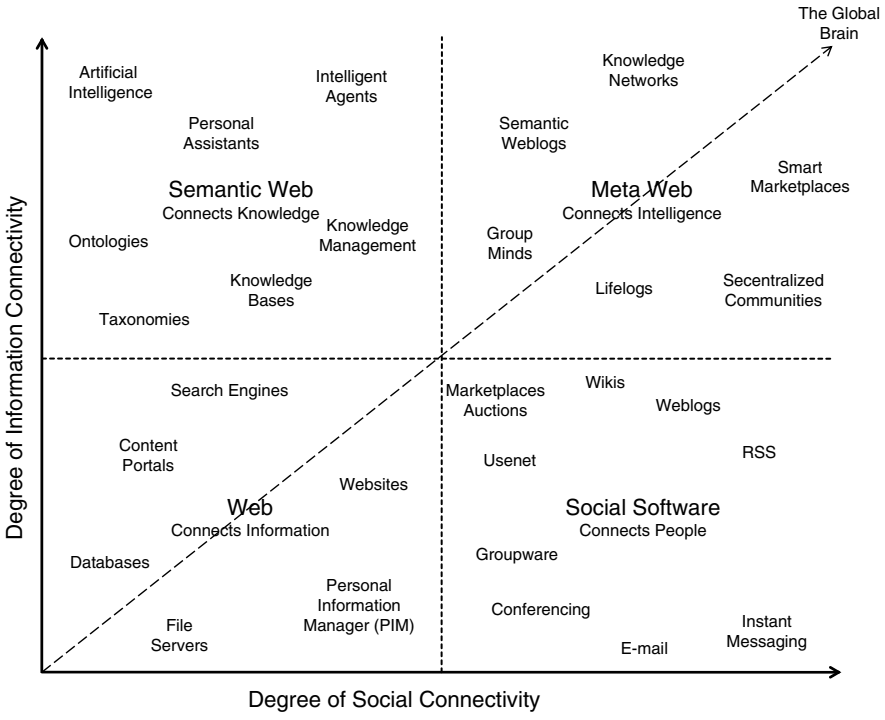


Fig. 1.1 Web 3.0 (see Howell, 2009; Seel & Ifenthaler, 2009)

elaboration of the traditional concept of semantic networks (cf. Larissa & Hendler, 2007). Strongly associated with this is the idea of allocating effective computational intelligence to intelligent agents in order to help them conduct logical reasoning.

1.7 Preview

The following chapters of this volume can be considered as a description of both the current and the future methodology of psychological and educational diagnostics of knowledge necessary in natural and artificial learning environments. In accordance with the introductory chapter, the issues described and discussed will include the representation of knowledge and methodologies for a systematic assessment of knowledge. The emphasis will be on computer-based procedures of knowledge assessment, and the central idea is to integrate diagnostic tools such as semantic nets and concept maps into telematic systems which provide computational intelligence.

References

Aebli, H. (1981). *Denken: Das ordnen des Tuns. Band II: Denkprozesse*. Stuttgart: Klett-Cotta. [Thinking: Sequencing of doing]

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Brebner, J. M. T., & Welford, A. T. (1980). Introduction: An historical background sketch. In A. T. Welford (Ed.), *Reaction times* (pp. 1–23). New York: Academic Press.

Bruner, J. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.

Clark, R. E. (2006). Not knowing what we don't know: Reframing the importance of automated knowledge for educational research. In G. Clarebout & J. Elen (Eds.), *Avoiding simplicity, confronting complexity. Advances in studying and designing (computer-based) powerful learning environments* (pp. 3–14). Rotterdam, The Netherlands: Sense Publishers.

Donders, F. C. (1869). On the speed of mental processes. In W. G. Koster (Ed.) (1969), /Attention a Performance II. *Acta Psychologica, 30/*, 412–431. (Original work published in 1868.)

Frost, R. A. (1986). *Introduction to knowledge base systems*. London: Collins.

Goodman, N. (1968). *Languages of art. An approach to a theory of symbols*. Indianapolis, IN: Bobbs-Merrill Comp.

Gordon, A. D., & Jupp, P. E. (1989). The construction and assessment of mental maps. *British Journal of Mathematical and Statistical Psychology, 42*, 169–182.

Hambleton, R. K., & Zaal, J. N. (Eds.). (1991). *Advances in educational and psychological testing*. Boston: Kluwer Academic Publishers.

Helbig, H. (2006). *Knowledge representation and the semantics of natural language*. Berlin and New York: Springer.

Howell, L. (2009). *Web 3.0 has long since passed Web 2.0 in education for 2020*. Retrieved 06-09-2009, from http://tagcblog.edublogs.org

Lajoie, S. (Ed.). (2000). *Computers as cognitive tools: No more walls, II*. Mahwah, NJ: Lawrence Erlbaum Associates.

Larissa, O., & Hendler, J. (2007). Embracing "Web 3.0". *IEEE Internet Computing* (May–June), 90–93.

Liu, Y. C., Chiang, M. C., Chen, S. C., & Huang, T. H. (2007). An online system using dynamic assessment and adaptive material. *Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference*, October 10–13, 2007, Milwaukee, WI.

Markman, A. B. (1998). *Knowledge representation*. Mahwah, NJ: Erlbaum.

Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., et al. (2007). *On the roles of external knowledge representations in assessment design. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Graduate School of Education and Information Studies*. University of California, Los Angeles.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Overman, S. (Ed.). (1988). *Methodology and epistemology for social science. Selected papers from Donald T. Campbell*. Chicago: University of Chicago Press.

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York and Oxford: Oxford University Press.

Prinz, W. (1983). *Wahrnehmung und Tätigkeitssteuerung*. Heidelberg: Springer. [Perception and action regulation]

Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three models of learning. In J. U. Cotton & R. L. Klatzky (Eds.), *Semantic facts in cognition* (pp. 37–54). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & The PDP research group (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 7–57). Cambridge, MA: MIT Press.

Sarbadhikari, S. N. (2004). Automated diagnostic systems. *Indian Journal of Medical Informatics, 1*(1), 25–28.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Seel, N. M. (1991). *Mentale modelle und weltwissen*. Göttingen: Hogrefe.

Seel, N. M. (1995). Mental models, knowledge transfer, and teaching strategies. *Journal of Structural Learning and Intelligent Systems, 12*(3), 197–213.

Seel, N. M. (1999). Educational semiotics: School learning reconsidered. *Journal of Structural Learning and Intelligent Systems, 14*(1), 11–28.

Seel, N. M. (2008). Empirical perspectives on memory and motivation. In J. M. Spector, M. P. Driscoll, M. D. Merrill & J. van Merriënboer (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 39–54). Mahwah, NJ. Lawrence Erlbaum.

Seel, N. M., & Ifenthaler, D. (2009). *Online lernen und lehren*. München: Rheinhardt Verlag.

Seel, N. M., Ifenthaler, D., & Pirnay-Dummer, P. (2009). Mental models and problem solving: Technological solutions for measurement and assessment of the development of expertise. In P. Blumschein, W. Hung, D. H. Jonassen, & J. Strobel (Eds.), *Model-based approaches to learning: Using systems models and simulations to improve understanding and problem solving in complex domains* (pp. 17–40). Rotterdam: Sense Publishers.

Seel, N. M., & Winn, W. D. (1997). Research on media and learning: Distributed cognition and semiotics. In R. D. Tennyson, F. Schott, S. Dijkstra, & N. M. Seel (Eds.), *Instructional design international perspectives. Volume 1: Theories and models of instructional design* (pp. 293–326). Hillsdale, NJ: Lawrence Erlbaum.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312.

van der Linden, W. J. (1991). Applications of decision theory to test-based decision making. In R. K. Hambleton, & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 129–156). Boston, MA: Kluwer.

Chapter 2
A Functional View Toward Mental Representations

Anna Strasser

2.1 Representation in General

Representation is a notion used in many different areas; this may be a reason why its meaning is quite ambiguous. Regarding the philosophical tradition, we can refer at least to four essential meanings of "representation" (Cp. Ritter, Gründer, & Gabriel, 2007, vol. 8, p. 1384).

1. Any mental state with cognitive content ("imagination" in a wide sense)
2. Mental states which refer to earlier mental states like memory ("imagination" in a narrow sense)
3. Any structure-preserving presentation like pictures, symbols, or signs
4. A substitution of something

To use the notion of representation for experimental studies in learning sciences or in any other empirical science, it would be desirable to have a more clearly defined notion. This chapter offers a preliminary definition which can be adapted to special domains. Ideally, this definition will facilitate the process of operationalization and the search for the right indicators to measure mental representations.

An object used as a representation can be described as an object standing for or referring to something. This object might be a material thing, a sign, a process, or a state. A representation has the role of substituting for something else. Usually, representations are not detailed copies of the object they represent. They rather picture something as something for somebody; only the important information is presented and interpreted depending on the situation. (For a critical discussion, Cp. Goodman, 1969.) An important distinction is to be drawn between two types of representations: the mental (internal) and the external ones.

A. Strasser (✉)
Institute of Philosophy, Humboldt-University, Berlin, Germany
e-mail: anna.strasser@t-online.de

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_2,

The notion of a mental representation is introduced as a theoretical construct to explain ongoing information processing. Mental representations and their relations are used to explain how humans are able to respond in a flexible way to one and the same input instead of being obliged to react in a rigid manner. Depending on this inner state, humans are able to show different reactions to the same input. Any other cognitive process like believing, anticipating, expecting, and memorizing is explained by mental representations as we will discuss in Section 2.1. Although the notion of a mental representation is a theoretical notion to explain what is going on in the "black box", in some future it might be possible to describe neurological states and processes in the brain which realize the role of mental representations. Questions concerning the possible neurological realization of mental states will not be discussed in this chapter. To describe the essential relations of a representation, there are several theoretical positions, e.g., the causal theory of representation, the functional theory, the theory of similarities, and the so-called theory of structural similarity (Cp. Dretske, 1981, Fodor, 1987, Millikan, 1984). We will return to those theories later on.

External representations are different and cannot be described within the same theory. Any object in the world can be used as a representation for another object or even for a mental representation. For example, a picture of somebody can represent this person, or a traffic sign can represent information, or an architectural model may represent a building. External representations presuppose mental representations; they cannot refer to something without somebody understanding this reference. Only if there is an interpreter of a sign can it be regarded as a representation; otherwise it is just a simple trace or an arbitrary copy of something. The necessary conditions under which something can be considered a representation are not to be found in the object itself but in the relations between the representation, the represented object, and the subject. Those representations will not be discussed in detail. But we should keep in mind that they are dependent on mental representations and interrelated with communication.

As we are dealing especially with the topic of learning and instruction, we will focus on mental representations. The structure of this chapter will be the following: To begin with we will discuss mental representations, clarifying the used terminology as there are the notions of a vehicle, a representandum, a subject, and a triple-digit relation. In the second part we will examine related notions of mental representations. And last but not least, we will have a short look at questions on operating on representations and what it means that some representations are reportable.

2.1.1 Mental Representations

Having a mental representation could be described as entertaining a mental image or an imagination. The word "image" might be misleading because it is suggesting a picture, whereas a representation can be any structural item that stands for the represented object. You may compare it to the binary code of a computer. This statement

does not claim anything about how something is represented. Typical examples would be conceptual representations like thoughts, beliefs, or wishes as well as sensory, non-conceptual representations like visual or auditory representations and any other sensation (pain, hunger, etc.). This distinction is analog to the following distinction in psychological literature: language-like representation and picture-like representation (images/imagery). The essential feature of a representation is its relation to something in the world or to another mental state – there is always a reference.

To be precise, following terminology will be introduced: Talking about mental representation you can distinguish between the vehicle (the medium, the representation itself), the representandum (the content, the represented object in the world or in the head), the subject, and the relation between those components. This relation can be described as a triple-digit relation: the representation has a meaning to the subject and refers by this meaning to the representandum (Fig. 2.1.). To quote Peirce (1931): "To stand for, that is, to be in such a relation to another that for certain purposes it is treated by some mind as if it were that other".

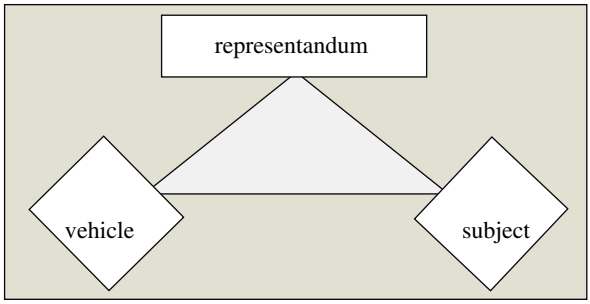


Fig. 2.1 The triple-digit relation: R<s, v, r>

2.1.2 The Vehicle

Concerning the vehicle of a mental representation, it is useful to consider the function of theoretical constructs and their possible neurological realizations. First, a theoretical construct does not imply any ontological claim, but the further development of neuroscience might show that there are corresponding states and processes. At this point, we do not need to worry about whether neuroscience will successfully determine certain states or processes as indicators of mental representations or not. We may just assume that there is "something" that is realizing the vehicle of a representation and ask ourselves what functional features go along with it.

2.1.3 The Representandum

The representation is referring to the representandum which may be an object, a fact, or another mental state. To specify the representandum, we can use the notion of

content. The content of a representation can be true, accurate, appropriate, adequate, or consistent; that means it has semantic properties. But this is only true if there is an interpreter, in the case of mental representations the subject employing the representation. The content of a mental representation could, for example, consist in knowing something about facts in the world. Language-dependent representations have a content which can only be understood if you take the language community into account as well.

2.1.4 The Subject

The relation between the representation and the representandum must be created by a person, the language-community, and the circumstances. Being a representation is not a property of something itself; being a representation is dependent on the fact that somebody is using it as a representation. Otherwise it would just be something like a natural sign (spoor or trace) and not a representation. If something is used as a representation, it has a function to the user. Besides that, the attitude of the subject toward the representandum plays an important role; you can wish, believe, or hope something.

2.1.5 The Triple-Digit Relation

As we have stated before, a representation is not just a copy of the referred-to object. Its relation to the representandum differs, depending on what kind of representation you consider. As we mentioned above, there are different theories of representations but each theory on its own seems not to be able to cover all kinds of representations. This has not been really successful because the description of the adequate relation of a representation depends on the type of representation you refer to. In this chapter we will just deal with two types of mental representations: the sensory (non-conceptual) and the conceptual ones.

Concerning the general case of sensory, non-conceptual representations (comparable to imagery/pictorial representations), there is something like a causal link to the representandum: A stimulus evokes a reaction/response of the sensory systems (organs) and together with some sort of information processing a representation evolves. Concerning visual stimuli, you can observe information processing before anything like a representation is accessible for the human being. A causal theory of representation seems to be adequate to describe the triple-digit relation to the representandum concerning sensory representations. An exception would be a hallucination, which is a representation of a non-existent object and therefore without a direct causal link to the representandum. A possible objection is the fact that hallucinations also presuppose a memory of sensory sensations. Taking this into account, you might talk about an indirect relation to a representandum. But it should be called a misrepresentation (see below) as well.

On the other hand, conceptual representations (comparable to non-pictorial, language-like representations) have structure-preserving relations like isomorphic as well as partial isomorphic relations. Concerning the former, we postulate that some essential structures are chosen and represented, but to distinguish a simple trace from a representation we have to claim a relation to the subject employing the mental representation as mentioned above. Concerning conceptual representation, a theory of structural similarity would be one way to describe the relation of the vehicle to the representandum. But taking into account that content can only have a meaning if it is playing a functional role for an interpreter, a functional theory might be a way to describe the relation of the subject to the representandum.

An interesting approach to describe representations has been developed by Gottfried Vosgerau (2008): Going back to the idea of the automata theory, the behavior of an automaton is described by a function, defining the mapping of a state of the automaton plus the input and the consequent internal state plus the output.

This approach is related to the symbol-system hypothesis saying reasoning is symbol manipulation and the so-called Church–Turing thesis claiming that any symbol manipulation can be carried out on a Turing machine. The consequence of these hypotheses implying that any symbol manipulation can be carried out on a large enough deterministic computer will not be discussed here.

To describe the behavior of human beings, we can refer to inputs (sensory sensations, internal states, mental representations), outputs (behavior), and following states. The claim of an inner state can explain why humans are able to react in a flexible way to one and the same input instead of being obliged to react in a rigid manner. A human being can react differently depending on his internal state. Just like a soda machine reacts differently depending on whether money has been inserted or not which is equivalent to two distinct states. The fact of money having been inserted or not is represented through the inner state of the machine, and this state plays a functional role for the further behavior of the machine. In the same way, internal states have a functional role for human behavior. How this function is best described will be a question for individual sciences to answer. The main idea here is to see that internal states can represent facts or substitute facts in a way so that they play a functional role to further processing.

2.1.6 Types of Mental Representation

To come to an even clearer notion of mental representation, we will have a look at further distinctions. There seem to be many kinds of representations which differ in certain ways. The terminology differs not only regarding the different disciplines. Looking, for example, into the field of imagery (Cp. Kosslyn, 1980) – a topic mainly treated by psychologists – you will find the distinction of pictorial and non-pictorial representation. The non-pictorial representations are seen as discrete or digital. The

pictorial ones are described as continuous and analog. Of course, you can also think of a hybrid form having pictorial and discrete elements.

In cognitive science, you will find several terminological notions for different formats of mental representations, just mentioning mental models (Johnson-Laird, 1983), retinal arrays, primal sketches and 2½-D sketches (Marr, 1982), frames (Minsky, 1974), sub-symbolic structures (Smolensky, 1989), quasi-pictures (Kosslyn, 1980), and interpreted symbol-filled arrays (Tye, 1991).

In the philosophical discussion as introduced above, you will find the distinction of two or three types of mental representations. The first one is called non-conceptual representation and its characteristic properties are phenomenal features. The notion of phenomenal feature involves, roughly, sensory representations, experiences, and image-like representations. The second group is called conceptual representation which is normally seen as non-phenomenal but rather abstract. The conceptual representations might be described as being in a language-like medium. But this is not accurate, because there are positions claiming that there are conceptual representations which do not depend on language. For example, very young children are able to distinguish between living and nonliving things. They have no linguistic abilities but they have an idea, a concept of living things. The third one is again a hybrid type, a representation with phenomenal features and conceptual elements.

Looking at the relation to the subject, the content of a representation can be conscious or unconscious. Unconscious representations will play no role in this chapter because our focus is on conscious representations. This is owed to the fact that diagnosis of knowledge is related conscious knowledge. There are several possible attitudes a subject can have concerning the content (representandum) of a representation, like believing, regretting, hoping, fearing. This differentiation will not be analyzed in detail because it does not play an important role in studies in the field of learning sciences.

Figure 2.2 shows you what types of representations we have already referred to:

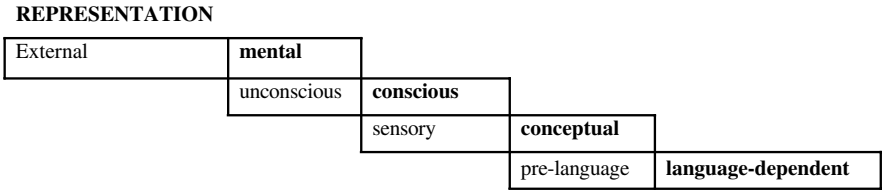


Fig. 2.2 Kinds of representations

So far we can state that the mental representations referred to in this book seem to be mainly mental, conscious, conceptual, and language-dependent representations. These are the very representations we will focus on from now on. After this general introduction, we will first examine what is not falling under this concept to describe the demarcation of the notion of representation to other nearby notions. The aim is to come near to a clear-cut discrimination.

2.2 Related Notions

Seeing mental representations as a component of information processing with semantic properties and being described by a triple-digit relation, you have to describe the relationship to other nearby notions of mental representations. The aim of this section is to prevent that those notions are mistaken with the notion of a representation. First, we will give some examples where psychological processes are explained by representations, then we will refer to notions explaining how representations can be structured, and last but not least we will discuss what a misrepresentation is meant to be.

2.2.1 "Explained by . . ."

In many descriptions of mental processes, representations play a role; they are used as cognitive building blocks. Memory, imagination, thinking, anticipation, expectation, and substitution are explained by the function of representations. Having a memory means to have a representation that is a reconstruction of an earlier mental representation. You can classify them via their content: There can be – just to mention some – perceptual, conceptual, and episodic contents, and there are more categorical representations using abstract schemata and several forms of hypothetical representations. The theoretical construct of a representation is used to describe processes in the so-called black box in order to explain cognitive abilities. Analog other cognitive abilities are explained by using representations.

2.2.2 "Structured by . . ."

In the following section, some structuring notions are shortly introduced to describe their relation to representations. Notions like schema, scripts, and mental models have in common that they describe how representations can be organized. Focusing on the meaning of the notion of a schema concerning cognitive science, you can state a schema is meant to be a hierarchically ordered structure of knowledge, evoked by recurring experience, for example, by repeated episodes of actions. Components of schemata have the role of variables which assures the flexible use of those structures (Cp. Ritter et al., 2007). A schema is a structure that can be used to organize representations. A nearby notion is the notion of scripts; they seem to be a little bit more language dependent but this impression is not backed up by any definition; it is just the casual usage of this notion. Schemata and scripts are seldom used in philosophical discussions, a similar structuring function is here fulfilled by so-called mental models (Cp. Johnson-Laird, 1983); they are understood as a conceptual framework of representations of knowledge. This knowledge can be related to the person itself (self-model), to parts of the world (world-model), or to abstract correlations. It is worth pointing out that the above description differs from the meaning of "model" in learning sciences where a model is an ad hoc construction with no duration. In philosophy, however, a model has a lasting structure. Those different meanings

should be discussed to avoid misunderstandings before any interdisciplinary work can start. Many other notions have been introduced to describe structuring features of representations. Just by looking at the papers collected in this anthology, you will find notions like semantic maps, concept maps, or an individual's knowledge structure described as a data association array.

Reflecting these structuring notions, you can state that representations which should represent complex knowledge have to be structured and connected to other representations. To be able to have a conceptual representation, you must be able to develop the ability to structure information. And this might even be a necessary condition for the possibility to "externalize" representations as we will discuss later on. The accurateness of a conceptual representation seems to depend on the right structuring framework.

2.2.3 Misrepresentation

To have a clear notion of representation, we should have an idea about what constitutes a misrepresentation. To define what is meant by a misrepresentation, we go back to our first general explanation: a representation is understood as a triple-digit relation. A misrepresentation is a representation that fails to refer to a representandum. This is the case if the representandum doesn't exist or if the represented properties do not belong to the related representandum (Drestke, 1994). A relation can go wrong if it is not adequate; this means we have to define what an adequate relation is. First, we state that any mental representation has an intentional character; the vehicle is used by somebody to refer to something. Second, we have to analyze which criteria have to be fulfilled so that this "referring" can be judged as successful. As we mentioned above, there are different positions about the nature of this relation. Keeping in mind that there are several kinds of representations, it might be reasonable to assume that this relation differs depending on what kind of representation is involved. There may be no unifying theory of representation defining the criteria of an adequate relation for every type of representation. Consequently, it is reasonable to look for a theory about particular types of representation as a first step toward an overall theory (Cp. Vosgerau, 2008.). For example, the so-called sensory representations can be described with a causal theory; stimuli evoke those representations. A misrepresentation would be a representation of a stimulus without the existence of the stimulus, for example, a hallucination. (Possible objections were discussed above.) This is still a representation; there is a relation to a representandum but this relation is not adequate concerning a causal link. In this case, it is a misrepresentation because there is no causal link to a stimulus. Concerning conceptual representations, it seems reasonable to refer to structure preserving similarities to judge whether a representation is adequate or not.

2.3 How to Operate on Representations

The detailed description of how one operates on representations depends on the special type of a representation. In this chapter, we focused on internal, conceptual,

and language-dependent representations. Just being able to represent something mentally is not enough for gaining knowledge. You have to be able to operate on representations as well. You need a form, for example, a schemata, a script, or a mental model to provide a structure which is able to picture relations between different representations. Given such a structure, you can simulate or anticipate possible changes; those structures make it possible to enrich the knowledge base.

The description of what it means to have a concept of something will lead us to a deeper understanding of the operations necessary to the ascription of a conceptual representation to a system. A concept requires abilities of differentiation, classification, abstraction, and generalization. For example, to have a concept of the color red, you should be able to ascribe this property to different objects, and you will need an idea of something being colorful.

Having a concept of something can be understood as analogue to the expression of having a mental (conceptual) representation of it. Of course, there are different theories about concepts, like concepts as abilities or as abstract objects (Cp. Margolis & Laurence, 2008). In this chapter, we will restrict ourselves to the understanding of concepts as a specific kind or form of conceptual representations. This idea goes back to Fodor (1987) and his representational theory of the mind (RTM). Concepts are deemed to psychological entities. As we described above, the relation between a subject and a mental representation is presented as taking a belief or any other propositional attitude.

A classic contemporary view (Fodor, 2003) postulates that representations have a language-like syntax and a compositional semantic. Besides this view, some claim that representation involves more pictorial structures. But if you remember the different types of representations, it is easy to imagine that pictorial structures are more suited to sensory representations than conceptual ones. Here we are concerned with the conceptual, language-dependent representations, and for those the above language-like description will be appropriate. Having a language-like syntax means that you will find a subject/predicate form including logical devices like variables or quantifiers. This mental representation view of concepts can be found in the work of Pinker (1994), Carruthers (2000), Fodor (2003), and Laurence and Margolis (1999). It seems to be a widespread position.

2.3.1 How Do You Know that Someone Is Able to Generate and Use Conceptual Representations?

From a philosophical point of view, you will ask yourself under which criteria you will be inclined to ascribe an internal conceptual, language-dependent representation to a person. This could lead to a discussion about dispositions and abilities which seem to be a consequence of representations. Given that a certain conceptual representation is explained by the knowledge about a certain fact, for example, the simple fact that if it is raining then the person with this representation has the knowledge, and under normal circumstances this person is able to express this knowledge with the sentence: It is raining. The person knows what it means for the statement

to be true and is able to judge whether incoming information about the outer world fulfills the truth conditions of this statement.

In interdisciplinary fields, this question of sufficient and necessary conditions should not only be understood as a pure conceptual question; the findings of the empirical sciences are to be included in determining the conditions for conceptual representation. There are indicators which tell us that somebody has the ability to generate and use conceptual representations. In learning sciences, the diagnosis of knowledge is done by means of several methods; on the one hand you can try to test what somebody does not know and on the other hand you can try to find out what somebody knows. To examine complex knowledge structures, concept mapping is a widespread method. Concerning the topic of text understanding, test persons are asked to give a written summary of a given text and then they are asked to develop a concept map with the main notions in boxes and pointers with annotations. Conceptual representations are not described here; all you can claim is that those tests indicate that somebody has some knowledge and this knowledge is represented somehow. An external representation like a concept map refers to mental representations in two ways. First, the external representation bears relations to the individual conceptual representations, for example, relations among items on the map are assumed to be isomorphic to relations among concepts. Second, there must be a relation to mental representations of other members of the language community, otherwise we could not talk of a successful expression of knowledge. The important point is that external representations have to be interpretable by other persons as well.

The process of transforming a mental representation into an external one is often described as externalization but this may not be the accurate way to describe this. In my opinion, a mental representation cannot be externalized but an external representation can refer to a mental representation. Given that someone has a conceptual representation about some facts in the world and he has several possibilities to express himself like talking, writing, or drawing, then he can be asked to show that he has this knowledge and that he is able to express it in an adequate way. Language is in the normal case the best means to create external representations referring to the content of an internal, conceptual, language-dependent representation.

References

Block, N. (1996). Mental paint and mental latex. In E. Villanueva (Ed.), *Philosophical issues 7: Perception* (pp. 19–49). Atascadero, CA: Ridgeview.

Block, N. (2003). Mental paint. In M. Hahn & B. Ramberg (Eds.), *Reflections and replies: Essays on the philosophy of tyler burge*. Cambridge, MA: MIT Press.

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. New York: Cambridge University Press.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. I. (1994). Misrepresentation. In St. Stich & T.A. Warfield, (Eds.), *Mental Representation* (pp. 157–173). Oxford: Blackwell.

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. (2003). *Hume variations*. Oxford: Oxford University Press.
Goodman, N. (1969). *Languages of art*. London: Oxford University Press.
Johnson-Laird, P. N. (1983). *Mental models.* Cambridge, MA: Harvard University Press.
Kosslyn, S. M. (1980). *Image and mind.* Cambridge, MA: Harvard University Press.
Laurence, S. & Margolis, E. (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.
Margolis, E., & Laurence, S. (2008). Concepts. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, URL = <http://plato.stanford.edu/archives/fall2008/entries/concepts/>.
Marr, D. (1982). *Vision*. New York: W.H. Freeman and Company.
Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.
Minsky, M. (1974). *A framework for representing knowledge.*Cambridge, MA: MIT-AI Laboratory Memo 306 June.
Peirce, C. S. (1931). *Collected papers of Charles Sanders Peirce.* Harvard University Press. Cambridge.
Ritter, J. Gründer, K., & Gabriel, G. (2007). *Historisches Wörterbuch der Philosophie*. Basel: Schwabe AG Verlag.
Smolensky, P. (1989). Connectionist modeling: Neural computation/mental connections. In L. Nadel, L. A. Cooper, P. Culicover, & R. M. Harnish (Eds.), *Neural connections, mental computation*(pp. 49–67). Cambridge, MA: MIT Press.
Tye, M. (1991). *The imagery debate*. Cambridge, MA: MIT Press.
Tye, M. (2000). *Consciousness, color, and content*, Cambridge, MA: MIT Press.
Vosgerau, G. (2008). Adäquatheit und Arten mentaler Repräsentationen. *Facta Philosophica.* Accepted.

Chapter 3
Mental Representations and Their Analysis: An Epistemological Perspective

J. Michael Spector

3.1 Introduction

The focus in this chapter is on the analysis and formative assessment of mental representations, especially mental models, which are created in response to challenging problem situations. It is widely accepted that internal representations are critical for effective problem solving and decision making, especially with regard to complex and ill-structured problems (see, e.g., Jonassen, 2000). The ability to provide timely and meaningful assessments of external representations of these internal mental structures is, therefore, critical for effective learning support. The perspective taken in this chapter is based on a naturalistic epistemology. The point of departure is learning. Then learning in complex domains is discussed as a precursor for the discussion of mental models and their assessment.

3.1.1 The Nature of Learning

Learning is fundamentally about change. A claim that an individual has learned something is a claim that the individual now believes, knows, or is capable of doing something not previously possible for that individual to the degree or extent now possible. Individual learning involves changes in attitudes, beliefs, capabilities, knowledge, predispositions, skill levels, and often more. In many cases, it is relatively simple to determine that learning has occurred. Learning how to ride a bicycle is a task that is observable. One can see the child falter and fall in the beginning. With time, the child might begin to ride without falling and only falter occasionally. Eventually, the child masters the task and might even try riding without holding on to the handle bars, which might be an unintended outcome of training a child to ride a bicycle. In any case, the changes are observable. Moreover, changes persist

J.M. Spector (✉)
Learning and Performance Support Laboratory, University of Georgia, Athens, GA, USA
e-mail: mspector@uga.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_3,

over time. The bicycle riding skill is not easily extinguished. An individual who has learned but not ridden for years is very likely to be able to still ride with ease.

Assessing progress of learning with such simple skills is not problematic. One can tell by looking if and how well (roughly) an individual has learned to ride. Riding without using the handle bars may exemplify mastery of balance and be an indicator of an advanced skill level. However, the typical pragmatic test of bicycle riding is whether an individual can ride, turn, stop, and start again without falling; establishing the validity of a test is not at all problematic with regard to bicycle riding. The performance is observable as is the relative level of performance. Multiple observers are likely to report the same results (e.g., cannot ride, rides with difficulty, rides with ease), so reliability of performance ratings is easily established. Near-transfer tests are easily constructed (e.g., "try riding the neighbour's mountain bicycle"). Far-transfer tests do not seem relevant as bicycle riding is not a building-block skill used in other or in more advanced tasks – typically, bicycle riding is an end-in-itself pursued simply for convenience or pleasure.

There are many different kinds of things that individuals can and do learn. Educational researchers have classified these things in different ways (cf., Bloom, 1956; Gagné, 1985; Merrill, 1993). Things to be learned might be characterized by the types of skills involved (e.g., psychomotor skills vs. intellectual skills) or they might be characterized by discrete aspects of the content (e.g., concepts, facts, procedures, rules). These classifications might be regarded as learning ontologies – things that can be learned – and they are useful to many instructional designers who are in the business of constructing effective, efficient, and engaging materials and activities to support learning.

However, individuals are likely to view the world in terms of other than such learning ontologies, however useful those may be for the design of instruction. An individual is more likely to view the world in terms of tasks and problems ("enterprises" in Gagné & Merrill, 1990, and "whole tasks" in van Merriënboer & Dijkstra, 1997). Moreover, individual views are often colored by interests, moods, prior experiences, and more.

The emphasis in this chapter is on how an individual comes to know and understand something (epistemology) – not on the things that can be known and understood (ontology). I acknowledge that the nature of that which is to be learned does have implications for the design of effective, efficient, and engaging instruction (Spector & Merrill, 2008). However, this volume is about computer-based diagnosis and analysis of knowledge, so the emphasis herein is on epistemology (a learning focus) rather than ontology (a design focus). In a subsequent section, an epistemological perspective believed to be appropriate for assessing learning will be described. First, however, the focus of this chapter is further constrained to complex problems and tasks.

3.1.2 Learning Complex Problem-Solving Tasks and Skills

Determining whether or not learning has occurred and assessing progress of learning is not always as simple as the bicycle riding case. Of particular interest in this

chapter are complex tasks and problems that lend themselves to multiple acceptable approaches and solutions. The reason for this focus is twofold. First, it is reasonably well established how to assess learning simple tasks and discrete knowledge items such as concepts, facts, and rules. It is also quite straightforward how to determine whether an individual has mastered a simple procedural task or knows specific concepts, facts, and rules. In cases involving simple tasks and declarative knowledge, computer-based diagnostic methods are available; these methods can be and have been successfully applied on a large scale outside experimental contexts, and the validity and reliability of particular methods have been adequately documented. In short, we know how to tell if someone can consistently perform a simple task well; we know how to find out if someone remembers specific facts, uses a concept appropriately, or applies a rule correctly. Most of the relevant things to include in assessments are readily identified and easily observed in relatively simple cases involving well-defined knowledge and task domains.

Second, how to assess learning in complex task domains is much less well understood (Spector, 2004, 2006; Spector & Koszalka, 2004). In this chapter, complex tasks and problems are those (a) that involve multiple, related, and interacting components, and (b) that are open to alternative acceptable solution approaches and solutions. Examples of such problems include (a) designing a bridge across the River Kwai, (b) developing a policy to control the salmon population in the fjords of Norway, (c) diagnosing a patient who is reporting stomach pains, and (d) planning a family ski vacation. Such problems share the characteristic of being incompletely or vaguely defined in one way or another. In some cases, the incompleteness and vagueness cannot be resolved prior to engaging in a problem-solving activity of some kind. Such problems also share the characteristic of having multiple acceptable (and unacceptable) solutions and solution approaches. As a result, such problems are often referred to as ill-structured or wicked problems (Jonassen, 2000; Klein, Orsanu, Calderwood, & Zsambok, 1993; Rittel & Webber, 1973).

As it happens, such ill-structured problem-solving and decision-making tasks pervade our lives. School-based learning often simplifies and constrains such problems so as to help individuals build up basic knowledge and gradually develop expertise. This approach to learning was perhaps appropriate for the industrial age, but it is increasingly acknowledged that knowledge workers in the information age need to develop flexible thinking and reasoning skills (Bransford, Brown, & Cocking, 2000; Olson & Loucks-Horsley, 2000). Indeed, many of the problems confronting society in the early twenty-first century are ill-structured. New curricula are appearing to address this situation. For example, there are relatively new curricula in environmental planning, conflict resolution, social health policy, and more that did not exist 10 years ago.

Curricula aimed at developing knowledge and expertise in complex domains introduces an interesting problem. How does one determine whether and to what extent an individual is improving his or her skill in solving ill-structured problems? If this obvious and apparently simple question is not answered, the efficacy and the impact of curricula aimed at complex and ill-structured problem solving are difficult to judge. One challenge pertains to far-transfer. The specific cases dealt

with in a curriculum for environmental planning, for example, are not likely to be encountered in the working life of an environmentalist; actual cases encountered on the job are likely to vary significantly from those cases discussed and analyzed in the curriculum. Many complex and ill-structured problems are unique. Just because one can design an adequate bridge over the River Kwai does not provide any guarantee that the same individual would be able to design an adequate bridge over the Tacoma Narrows – this is not a near-transfer task. The fewer specific characteristics shared between the learning task in the curriculum and the target task outside the school setting, the further the degree of transfer. Educators generally place much higher confidence in the ability of a learner who has mastered the school task to transfer that learning to a nearly identical or somewhat similar situation outside the school context. Near-transfer is generally not so problematic, although a near-transfer test is a good idea to include in many instructional sequences.

A problem arises with far-transfer tasks, however. First, it is difficult to predict whether someone who has performed well on a school-based learning task will be able to effectively transfer that knowledge and skill to a task that varies considerably and in significant ways from the school task. Establishing the content validity of a test or measurement that could support inferences with regard to far-transfer tasks is, as a result, quite difficult.

Second, when a person who did perform well on an initial task fails to do well on a far-transfer task, it is often a challenge to identify why and isolate the critical difference(s) that impeded transfer. This means that construct validity is also difficult to establish. More importantly, such information is critical for providing timely and meaningful formative feedback, which is an acknowledged requisite for effective instruction.

Third, when someone does well on both a near- and far-transfer task, it is difficult to attribute that person's success to their prior knowledge and skill, to their training or education, or to other factors. One might be tempted to say that superior performance on far-transfer tasks is not teachable – a few people do seem to develop such expertise but it is a mystery how. This is akin to the notion that the ability to perform at a very high level on a variety of very different tasks within a particular ill-structured domain (intuitive expertise) is a mysterious ability (Dreyfus & Dreyfus, 1986).

One could simply adopt the view that the development of expertise in complex and ill-structured problem-solving domains is fundamentally mysterious. The best one can expect from a school-based curriculum would be to develop basic knowledge and hope that expertise develops on the job or in the field. One consequence of such acceptance is that assessing programs and curricula becomes simple and easy, but also quite subjective and arbitrary. A second consequence is that something that might be subject to empirical investigation is pushed behind the curtain of mystery.

On the other hand, one could decide to take up this challenge and figure out a way to measure relative levels of knowledge and performance on complex, ill-structured tasks and use such measures for purposes of feedback, assessment, and evaluation. The challenge of finding valid and reliable and scalable measurements and technologies appropriate for learning and performance in complex domains is

one of the threads running through this volume, and it is the central issue of this chapter.

We have accepted the challenge. How, then, shall we proceed? One does not have to give up the things we already know how to do. That is to say that we can still measure basic declarative knowledge and performance on simple precursor tasks. Those may, in some cases, turn out to be reliable predictors of performance on more complex tasks, although one should not expect that outcome nor rely solely on simple measures.

More fundamentally, can we measure the development of expertise in complex domains? Shall we observe performance on representative (commonly encountered or frequently occurring) tasks? That might be possible although it would also require much time and effort. However, tasks vary so much in some of these domains that it would be difficult to identify representative tasks; even if there were a small set of representative tasks, there would be little guarantee that performance on those tasks would be predictive of performance on a new, far-transfer task.

We could gather performance data over time and possibly find a correlation between performance on a few so-called representative tasks and performance on a new, far-transfer task, and perhaps some other indicators of expertise (formal recognition by a professional association, certification, etc.). Someone could and should be doing that, and some are. Several researchers who are investigating highly superior task performance use performance on standard tasks, time as a skilled professional, and think-aloud protocol analysis while solving a representative task to determine levels of expertise (Ericsson, 2001; Ericsson & Smith, 1991). Ericsson has found that highly skilled performance depends to a large extent on certain cognitive and metacognitive skills (e.g., being able to identify a specific thing one needs to improve, being able to focus on how that aspect of one's performance is or is not improving). Ericsson's (2001) studies have two limitations, however. First, they do not scale up for practical use in training and education – his method of analysis involves think-aloud and retrospective protocol analysis (see also Ericsson & Simon, 1993). Second, all of his studies have been with regard to well-defined tasks that have easily identified and validated near- and far-transfer correlates and well-established, reliable indicators of levels of expertise (e.g., chess grand master). The focus here is on ill-structured tasks that lack established indicators of levels of expertise and on measurement methods that are scalable and useful in a real-time, training, or educational context (e.g., computer-based diagnostics).

One thing to carry forward from Ericsson's studies, though, is the critical role of cognition in problem solving and task performance. The traditional distinction between cognitive, affective, and psychomotor tasks (Bloom, 1956) is somewhat misleading when one takes a naturalistic perspective. That is to say that what naturally occurs with regard to how a person thinks and solves a problem does not necessarily divide into three different parts (cognitive, affective, and psychomotor). For example, how a person is feeling one day may easily impact how well they perform on an intellectual task such as solving quadratic equations. A psychomotor task such as riding a bicycle may not be easily mastered by a person with a deep fear of falling. Cognitively oriented remarks may well contribute to more efficient

mastery of a psychomotor task (e.g., "riding a bicycle is a skill that all kinds of people have mastered," "there is no such thing as a riding-the-bicycle gene – anyone is capable of learning this," or "when the front and rear wheels are aligned, balance is maintained by sitting upright and not leaning either to the left or right – try this going down this incline but do so without pedaling – put your feet out if you feel like you might fall over, but try to sit up straight and keep the wheels aligned").

In the course of a think-loud or retrospective protocol analysis, one might easily discover that there was a cognitive skill involved in the performance of an apparently psychomotor task. I recall such a discussion with Robert Gagné in the early 1990s. We were watching a video of an Air Force police sergeant teaching trainees how to handcuff someone. The person being handcuffed was a trained accomplice in this instructional sequence. The trainees had already read a training manual and seen several demonstrations of standard handcuffing procedures; there were in fact two different procedures – a procedure for most cases and a different procedure for persons who resisted arrest. As we watched the video, it became clear that the key element in successful task performance was determining whether or not the person was likely to resist arrest. Missing the cues for likely resistance nearly always resulted in problems and failure to secure the prisoner. As it happens, the primarily psychomotor task of handcuffing involved a critical cognitive component – recognizing resistance cues.

Most tasks, even those performed with apparent automaticity, involve cognitive aspects. While some very simple tasks may lack cognitive components, it seems quite likely that cognition plays a vital role in ill-structured problem solving. One can distinguish recurrent from nonrecurrent tasks (van Merriënboer & Dijkstra, 1997; van Merriënboer & Kirschner, 2007). Nonrecurrent tasks are those which are not performed the same when there are variations in the problem situation or circumstances surrounding the problem (such as the handcuffing task with the variant being resistance to the procedure). Nonrecurrent tasks require the problem solver to diagnose the situation and devise an appropriate solution path. A psychologist such as Seel (2001) would probably say that such problems require the problem solver to develop or modify an appropriate mental model to guide the resolution.

It is what the person is thinking and how that person is thinking about the problem situation that is very likely correlated with the quality of the solution that is developed and implemented. Looking at and assessing the solution to a complex problem requires the solution to be fully developed and implemented; this requires a great deal of time. Finding a completely worked out and known acceptable solution also takes time, and, with regard to ill-structured problems, there are often multiple acceptable solutions, so having a standard solution for use in an evaluation is not always feasible or appropriate. In cases where a wide variety of complex problems are involved and there is an opportunity to only base a judgment of learning progress on a few problem-solving tasks, the assessment focus shifts from the solution (or task performance) to what the person is thinking in terms of developing an acceptable solution to the problem. The focus shifts to mental models.

3.2 Mental Models

In this chapter, the term "mental model" is used to refer to any internal mental representation brought to bear in a problem-solving situation. The usage is somewhat broader than that found in Johnson-Laird (1983) as it includes what some cognitive scientists might call schema, scripts, and other hypothetical cognitive structures. Internal representations are not directly or immediately observable. We do not *perceive* mental models. We *construct* mental models in order to make sense of our experiences. This has been the prevailing view of cognitive psychologists for more than 25 years (Anderson, 2007; Johnson-Laird, 1983). The view that we construct internal representations to make sense of our experiences has a much longer history if one takes into account such philosophers as Wittgenstein. In the *Tractatus Logico-Philosophicus*, Wittgenstein (1922) remarked that we picture facts to ourselves. We create internal representations in order to make sense of things we experience. Accepting this mental representational aspect of human nature leads to a basic problem concerning the relationship between that which is internal and that which is external. As noted earlier, things that are directly or indirectly observable (i.e., external things, including external representations of internal mental structures) have a different status than things that are internal and unobservable (e.g., mental models). We make claims about external things – states of affairs – facts. We then have two kinds of things in the external world – a piece of language and a piece of reality. How do we come to understand their relationship? Wittgenstein's early attempt at this problem involved one-to-one mapping between the claim and the reality. When we can create, justify, and validate such a one-to-one mapping, then we know the alleged fact is true. Unfortunately, there are problems with this account. What counts as a one-to-one mapping is not necessarily easily resolved. More seriously, in constructing such maps, we are again engaging in thinking; we are again creating internal representations to make sense of something external – in this case, a piece of language is involved as one of the external, observable things.

As noted earlier, some cognitive psychologists distinguish mental models, schema, and scripts, all of which are hypothetical internal mental representations introduced to explain various aspects of human behavior. The discussion herein does not depend on such distinctions and is aimed generally at internal mental representations. The focus is on mental models as these are generally regarded as transitory and created just when needed to solve a challenging problem or make sense of an unusual situation (nonrecurrent problems); that is to say that mental models are a particular focus for complex and ill-structured problem solving. Schema and scripts are generally believed to be more stable and persistent internal representations that enable a person to react to a familiar situation (such as a recurrent problem) with ease. Mental models, schema, and scripts might all be involved in an individual's thinking about and responding to a challenging problem situation. We proceed without further discussion of the various kinds of internal representations.

Wittgenstein (1953) wrestled with the problem of linking internal representations to what we say and do; this struggle can be seen in his notes published after his death (see the remarks in *Philosophical Investigations* on language games for example).

Not only is it in our nature to picture facts to ourselves, but it is in our nature to talk about them – both the externally observable facts and our internal representations. The complexity of the problem of assessing what we know, especially about complex things, is now emerging. There are ways to confirm or repudiate claims about external things, even if we abandon the one-to-one correspondence approach and adopt a more practical approach involving a consensus of observers (a reliability indicator). We are aware of our ability to create internal representations, and we accept the role that such representations play in coming to understand and in developing knowledge and skill. However, one cannot observe someone else's internal representation nor can one observe one's own internal mental structures and processes. Nevertheless, we believe that we construct mental models and use them to make sense of what we experience. Is this not problematic?

Because mental models are hidden internal representations, our knowledge about them is necessarily incomplete and tentative. Why, then, introduce such hypothetical entities into discourse about learning, instruction, and performance? The answer is that it is difficult to explain how a person develops understanding and expertise without introducing internal representations. Mental models are perhaps most obvious in explaining how a person continues to make a particular kind of mistake. For example, suppose someone is entering text into a computer file using a word-processing program. The person wants the text to stay within the margins of the printed page. When the person nears the end of a line, that person hits the enter key rather than continuing to type on the keyboard. One might say such a person is thinking of the computer as a sophisticated typewriter. We might even find that this is in fact the case by asking why the person is using the enter key to move to the next line of text. What we have, then, is an external representation of that person's thinking or mental model. We still do not have the mental model itself. Nonetheless, we are now able to diagnose the problem and can advise the person to let the text wrap around to the next line, ignoring worries about the margin. The point is that mental models can be used to explain behavior; more importantly, identifying mental models (through external representations such as an explanation) can be useful in improving learning, instruction, and performance. Mental models are not idle hypothetical constructs; they are quite useful, even if they can only be known incompletely and partially through external representations.

3.3 Mental Model Assessments and Learning Progress

The challenge is to find valid ways to reliably and efficiently assess learning in complex domains. Valid tests of declarative knowledge relevant to the problem domain can be conducted reliably and efficiently. However, such tests may not predict task performance in complex domains; they may not be valid or reliable indicators of complex problem solving. Including task performance measures is, therefore, desirable in assessing learning in complex domains. With regard to ill-structured tasks, however, performance on one or two specific tasks may also not reflect how well

a person will perform on a wide variety of complex tasks in that domain. The role that mental model assessments might play is to supplement existing methods with an additional indicator that might reveal misconceptions or misunderstandings that might not influence one task performance but that could influence others. Moreover, mental model assessments might also reveal how a person's thinking is evolving through instruction and experience.

First, we ought to review some issues pertaining to validity and reliability with regard to measuring complex problem-solving skills. Cronbach and Meehl (1955) identified four types of validity with regard to psychological tests: (a) predictive, (b) concurrent, (c) content, and (d) construct. The first two (predictive and concurrent) are basically criterion-oriented indicators and useful when such accepted criteria are available (as is the case with the bicycle riding example discussed earlier). Such criteria are difficult to establish for ill-structured problems, although the notion of concurrent validity is useful when there exists a solution or problem conceptualization from an acknowledged expert that can be used as a measure of relative level of performance. As it happens, this situation can be and has been created with regard to problem conceptualizations for ill-structured problems using a mental model assessment set of tools called HIMATT (Highly Integrated Model-based Assessment Tools and Technologies) as discussed elsewhere in this volume.

What about the other two types of validity – content and construct validity? Content validity generally refers to representative nature of the items involved – namely, that the items fairly represent the problem space involved (as in "we are measuring the right things"). As suggested several times, it is very challenging to establish content validity with regard to nonrecurrent and ill-structured problems. Suppose, however, that the level of analysis for purposes of content validity involves particular aspects of a measure, such as the collection of critical nodes in a concept map (i.e., those nodes with more than four or five links). Combining the notion of concurrent validity based on multiple expert external representations (concept maps) with the notion of critical nodes for a particular problem scenario, one might be able to argue that those critical nodes serve as indicators of content validity (see Spector & Koszalka, 2004).

Construct validity is perhaps the most challenging measure to consider with regard to assessments of mental models; it is arguably the most important from the point of view of providing formative feedback and facilitating the development of expertise. Construct validity refers to an attribute or quality that is allegedly being measured by a particular test or item or procedure, and it is especially important when there is no definitive criterion measure available (Cronbach & Meehl, 1955). With regard to the kinds of assessments of mental models available in HIMATT and discussed in this volume, construct validity can be discussed at two levels (possibly there are others). First, a construct such as *misconception-A* might be introduced with regard to those whose annotated concept maps are missing a particular node found in expert representations. This level would be the concept level and other such misconceptions might be identified and subsequently validated through questionnaires and problem-solving activities.

A second level of construct validity with regard to assessing mental models might involve the general structure of external representations (i.e., annotated concept maps; see Spector & Koszalka, 2004). Several recent dissertation studies confirm that expert representations are typically highly interconnected whereas nonexpert representations exhibit fewer interconnections (Kim, 2008; Lee, 2008; McKeown, 2008). The construct in this case might be called *expert-like-structural-thinking*, and the indicator would be a measure of connectedness of the nodes in a concept map or graph (Ifenthaler, 2007).

The notion of reliability concerns the stability and consistency of the things measured. For example, we expect to get the same or very similar results when giving the same test problem to the same person from day 1 to the next. When multiple raters are involved, we expect to see high degrees of agreement among raters. When multiple problems and problem scenarios are involved from one test or person to another, we want to see high degrees of agreement across problems believed to be of the same general level of difficulty. Because the assessments in HIMATT are automated, we need not worry about interrater reliability. The concern shifts to the reliability of measurements made from 1 day to the next and across variations in problems and scenarios. Because the assessments are automated, it is relatively straightforward and simple to establish reliability.

To make this brief discussion of validity and reliability more concrete, an overview of one mental model assessment methodology is elaborated here; more detail on this methodology and on related methodologies can be found elsewhere in this volume. With regard to measures of problem-solving thinking and expertise, what can be observed are external representations of mental models that influence the problem solver's thinking. These external representations include annotated concept maps depicting how an individual conceptualizes the problem space (Spector & Koszalka, 2004) and text created in response to a problem scenario describing key factors influencing the situation (Pirnay-Dummer, 2007). External representations can be compared one to another to see how similar they are (Taricani & Clariana, 2006; Ifenthaler, 2007). The similarity metrics that have been automated in HIMATT (Highly Integrated Model-based Assessment Tools and Technologies) (Pirnay-Dummer, Ifenthaler, & Spector, in press) allow two concept maps to be compared with regard to structural features (e.g., interconnectedness of nodes, ratio of nodes and links) and semantic features (e.g., concept and propositional similarity). Protocol analysis (Ericsson & Simon, 1993) can of course be used to analyze semantic aspect of external problem representations, but protocol analysis does not lend itself to computer-based diagnostic methods.

Spector and Koszalka (2004) elicited external representations in the form of annotated concept maps. Respondents were presented with a problem scenario and asked to identify and describe the key factors influencing the situation and then the key relationships that exist among these factors. Expert representations were created as a kind of target or reference model. One finding was that it was very easy to tell by looking only at the concept map structure, which were created by experts; expert maps were highly interconnected.

Pirnay-Dummer (2007) and Ifenthaler (2007) created specific metrics for interconnectedness and other measures and these have been validated and automated in the HIMATT system (Pirnay-Dummer et al., in press). The validation studies included measures of a respondent at multiple points in time and comparisons with reference models. In summary, it has been demonstrated that it is possible to elicit external representations of what a problem solver is thinking and to analyze these in terms of progress against a previous model or a reference model. The various studies conducted using these computer-based tools have consistently shown connectedness measures and concept similarity to be the most useful in terms of distinguished highly experienced from less experience persons (Kim, 2008; Lee, 2008; McKeown, 2008). It is also possible to identify misconceptions using these methods, although they were designed for research rather than for teaching or formative feedback.

It is important to remember that mental model assessments are assessments of external representations of mental models (e.g., annotated concept maps); they are not direct measures of mental models. Such assessments are especially useful in supplementing more traditional measures of declarative knowledge and performance. The significance of these tools is that they are appropriate for use in ill-structured problem-solving domains, and they have been validated and automated. In short, computer-based diagnostic methods and tools provide new opportunities to investigate and support learning in complex domains.

3.4 An Epistemological Perspective

Epistemology is concerned with the nature of knowledge and how knowledge develops. As such, epistemology is an area of investigation for philosophers and psychologists, but of course there implications for instructional designers and teachers. One might characterize a core enterprise of philosophy as an ongoing attempt to explore the boundaries between sense and nonsense, whereas one might characterize a core enterprise of psychology as an ongoing attempt to describe how people develop understanding. In both disciplines, there has been a turn toward naturalistic approaches in the last 50 years. That is to say that philosophers and psychologists are less inclined to rely on a priori categories and deductive methods. Rather, the inclination is to describe and study people in their natural settings. Rather than argue, for example, that there are fixed, discrete and permanent categories of the mind, a naturalistic perspective is inclined to describe how people learn, make decisions, and solve problems in terms of what they actually do (performance behavior) and how they actually talk about what they do (language behavior – a kind of meta-performance behavior).

Observed performance is certainly a good indicator of understanding – especially when the task is simple and straightforward, such as riding a bicycle. As tasks increase in complexity and involve reasoning over time with regard to a number of interrelated components, task performance alone becomes more and more time consuming and a less reliable indicator due to the uniqueness of problems and issues pertaining to transfer.

Additional indicators are desirable. The argument is that if a problem solver is thinking about the problem in a manner that is indicative of expertise or more sophisticated than how that person used to reason, then there can be increased confidence in the assessment. One can, for example, elicit a representation of how a person thinks about a word-processing program. One can create a representative reference or expert model from a composite of several expert models. One can then see if a person is still thinking about the word processor as a sophisticated typewriter or whether that person is thinking about the word processor in a way that resembles how an expert might think. Granted that a word-processing problem is not so characteristic of the complex, ill-structured problems mentioned earlier (e.g., environmental planning, engineering design, medical diagnosis). The point here is that assessments of mental models can be made by eliciting external representations and then comparing those representations with reference models. These measures, when combined with measures of declarative knowledge and task performance (when feasible), together provide generally reliable indicators of relative level of expertise or understanding. Moreover, when proper care is taken, such measures can be both valid and reliable.

Because claims about mental models are necessarily tentative, one might conclude that mental model assessments are better used in a formative assessment or formative feedback context rather than in a summative assessment context. For example, since the reference model is already in the computer-based system, it is possible for the system to provide the respondent with the reference model and then ask reflective questions about differences in the response and the reference model. The system might ask a respondent to identify differences first and then to explain why something is in the reference model that does not appear in the respondent model. Such questions can be automatically generated and can serve as one kind of formative feedback. Indeed, using automated assessments of mental models created in response to problem scenarios as the basis for formative feedback is likely to prove to be one of the most powerful and productive kinds of support that can be provided for acquiring expertise in solving complex and ill-structured problems.

3.5 Concluding Remarks

We explored these questions in this chapter: (a) What can we know about mental representations? (b) Can mental models be reliably assessed? and, (c) How useful are mental model measures in facilitating the development of knowledge and expertise? The argument here has been that we can only know about mental models indirectly through external representations. The link between the external representation and the internal representation cannot be established with a high degree of confidence. We have limited knowledge of mental models and of the various mechanisms that might influence their development and structure. We can only know about external representations and then make tentative inferences about mental models.

Nonetheless, such knowledge about external representations and the associated inferences about internal representations are useful in determining how well a

person understands a family of ill-structured problems or whether and to what extent a person is making progress in developing expertise in a complex problem-solving domain. The various tools and methods described in this volume demonstrate that this is a maturing area of investigation, and, further, that it is important to making progress in understanding how knowledge is developed.

The third question has been less explored than the other two; there is the suggestion that mental model assessments (as known through external representations) can be useful in promoting learning and helping learners develop knowledge and understanding (Kim, 2008; Lee, 2008; McKeown, 2008). However, much more research remains to be done in this area, which might be considered the focus for the next generation of researchers interested in mental model assessments. In spite of all that we have come to understand about mental models and their assessments, we still know very little. Further, we know less than we are inclined to believe.

References

Anderson, J. L. (2007). *How can the human mind occur in the physical world*? New York: Oxford University Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, Handbook 1: The cognitive domain*. New York: McKay.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York: Free Press.

Ericsson, K. A. (2001). Attaining excellence through deliberate practice: Insights from the study of expert performance. In M. Ferrari (Ed.), *The pursuit of excellence in education* (pp. 21–55). Mahwah, NJ: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1–38). Cambridge: Cambridge University Press.

Gagné, R. M. (1985). *The conditions of learning* (4th ed.). New York: Holt, Rinehart, and Winston.

Gagné, R. M., & Merrill, M. D. (1990). Integrative goals for instructional design. *Educational Technology Research and Development, 38*(1), 23–40.

Ifenthaler, D. (2007, October). *Relational, structural, and semantic analysis of graphical representations and concept maps. Introducing the SMD-technology*. Paper presented at the Annual Meeting of the Association for Educational Communications and Technology, Anaheim, CA.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research & Development, 48*, 63–85.

Klein, G. A., Orsanu, J., Calderwood, R., & Zsambok, C. E. (1993). *Decision making in action: Models and methods*. Norwood, NJ: Ablex.

Kim, H. (2008). *An investigation of the effects of model-centered instruction in individual and collaborative contexts: The case of acquiring instructional design expertise*. Unpublished dissertation, Florida State University, Tallahassee, FL.

Lee, J. (2008). *Effects of model-centered instruction and levels of expertise on ill-structured problem solving*. Unpublished dissertation, Florida State University, Tallahassee, FL.

McKeown, J. (2008). *Using annotated concept map assessments as predictors of performance and understanding of complex problems for teacher technology integration*. Unpublished dissertation, Florida State University, Tallahassee, FL.

Merrill, M. D. (1993). An integrated model for automating instructional design and delivery. In J. M. Spector, M. C. Polson, & D. J. Muraida (Eds.), *Automating instructional design: Concepts and issues* (pp. 147–190). Englewood Cliffs, NJ: Educational Technology.

Olson, S., & Loucks-Horsley, S. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academies Press.

Pirnay-Dummer, P. (2007, April). *Model inspection trace of concepts and relations. A heuristic approach to language-oriented model assessment*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2009). Highly integrated model assessment technology and tools. *Educational Technology Research and Development*, *57*(6).

Rittel, H., & Webber, M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*, 155–169.

Seel, N. M. (2001). Epistemology, situated cognition, and mental models: "Like a bridge over troubled water". *Instructional Science*, *29*(4–5), 403–427.

Spector, J. M. (2004). Current issues in new learning. In K. Morgan, C. A. Brebbia, J. Sanchez, & A. Voiskounsky (Eds.), *Human perspectives in the Internet society: Culture, psychology and gender* (pp. 429–440). Southampton, UK: WIT Press.

Spector, J. M. (2006). Introduction to the special issue on models, simulations and learning in complex domains. *Technology, Instruction, Cognition and Learning, 3*(3–4), 199–204.

Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning incomplex domains* (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building). Syracuse, NY: Syracuse University.

Spector, J. M., & Merrill, M. D. (2008). Editorial: Effective, efficient and engaging (E^3) learning in the digital age. *Distance Education, 29*(2), 123–126.

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Eduational Technology Research and Development, 54*(1), 65–82.

van Merriënboer, J. J. G., & Dijkstra, S. (1997). The four-component instructional-design model for training complex cognitive skills. In R. D. Tennyson, F. Schott, N. Seel, & S. Dijkstra (Eds.), *Instructional design: International perspectives* (Vol. 1, pp. 427–446). Mahwah, NJ: Lawrence Erlbaum.

van Merriënboer, J. J. G., & Kirschner, P. A. (2007). *Ten steps to complex learning*. Mahwah, NJ: Lawrence Erlbaum.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus* (C. K. Ogden Trans.). London: Routledge & Kegan Paul.

Wittgenstein, L. (1953). *Philosophical investigations*. London: Blackwell.

Chapter 4
Multi-decision Approaches for Eliciting Knowledge Structure

Roy B. Clariana

4.1 Knowledge and Knowledge Structure

What is the nature of knowledge and of expertise and how can it be measured? Investigators' beliefs about the nature of knowledge determine the set of tools that they use to consider this issue and the tools they use inevitably alter their understanding of what knowledge is and is not (Sternberg & Preiss, 2005). Our connectionist bias follows Anderson (1984) that structure is the essence of knowledge (p. 5), and knowledge structure refers to how information elements are organized. Knowledge structure may be a facet of declarative knowledge (Mitchell & Chi, 1984), but Jonassen, Beissner, and Yacci (1993) go further to suggest that structural knowledge is a distinct type of knowledge, and that apposite structural knowledge is a critical go-between for declarative and procedural knowledge that facilitates the application of procedural knowledge (p. 4). Amalgamating their ideas, we propose that knowledge structure is the precursor of meaningful expression and is the underpinning of thought; said differently, knowledge structure is the mental lexicon that consists of weighted associations between knowledge elements (Rumlehart, Smolensky, McClelland, & Hinton, 1986).

Further, knowledge structure can be externally maintained and propagated through actions and artifacts such as this volume, and these are a residue of the actor's/author's knowledge structure. The implication is that knowledge structure can be intentionally elicited from an individual in various ways, but it can also be derived from existing artifacts, for example, from essays and concept maps (see Chapter 7, this volume). We hold a reductionist view that although an artifact must be interpreted or reconstructed by the reader, unless the information in the artifact is unintelligible, incoherent, or ambiguous, the reader will likely recapture to a greater or lesser extent the original author's knowledge structure.

Representations of knowledge structures derived from existing artifacts are probably highly constrained by the task and purpose of the artifact. Along these lines,

R.B. Clariana (✉)
The Pennsylvania State University, University Park, PA, USA
e-mail: rclariana@psu.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_4,

most text signals such as headings, underlining, and highlighting emphasize and thus convey the structure of the text and this almost certainly must augment the artifact's knowledge structure in the reader. Thus text signals and other devices are features of an artifact that deserve attention because of their likely direct effect on readers' knowledge structure.

Conceptually, "structure of knowledge" implies patterns of relationships that are complex to express or investigate. Learning simply involves modifying the connection weight between units in the network, but these few changes may have unexpected or hard to predict effects on the entire set of relationships involved. If so, from a measurement viewpoint, this means that if knowledge structure can be measured, it should be a fairly stable set of associations that primarily change incrementally through experiences, reflection, and consolidation, although occasionally it can dramatically shift conformation (e.g., a flash of insight). Currently, these patterns of relationship can be most easily represented visually and mathematically as networks. Several classes of weighted association networks align very well with this conception of knowledge structure and so provide a ready and well-established toolset for capturing, combining, analyzing, representing, and comparing knowledge structures. Any tools liberate but also constrain our musings. After careful consideration, we have focused on *Pathfinder* analysis as the optimal tool at this time (Schvaneveldt, 1990).

However, the *Pathfinder* approach for eliciting relatedness data depends on a pairwise rating approach that uses many single decisions and so is tedious, tiring, and time consuming, especially when more than 20 terms are compared. This chapter describes two computer-based multi-decision approaches for eliciting relatedness raw data that were specifically designed to complement *Pathfinder* analysis in revealing the salient associations as a path of nodes. These two "new" elicitation approaches (there is really nothing new under the sun; see Shavelson's, 1972, 1974 approaches) were developed to increase the efficiency and possibly the accuracy of relatedness data relative to the pairwise approach. Serendipitously, we determined while preparing this chapter that combing these two multi-decision approaches for eliciting relatedness data may provide an even better measure of knowledge structure that appears to capture both local and global associations better than the pairwise approach.

4.2 Relatedness Data and Its Analysis and Representation

There are several traditional approaches used to elicit concept relatedness, for example, free word association, similarity ratings of pairs of terms, and card sorting (Jonassen et al., 1993). And there are a few ways to analyze and represent that data, for example, hierarchical cluster analysis, multidimensional scaling (MDS), and *Pathfinder* Network (*PFNET*) analysis. But different elicitation and also different representation approaches obtain different measures of knowledge structure. For example, free word association is one of the earliest approaches for eliciting concept relatedness and is considered by some as the most valid approach (Jonassen

et al., 1993, p. 32). Because it is nearly context free, free word association most likely tends to elicit a general knowledge structure (versus an actuated knowledge structure discussed later), unless intentionally or unintentionally biased by the task directions, context, situation, or other factors. For example, as one of many free association investigations that he conducted, Deese (1965, p. 49) asked 50 undergraduates to free associate to a list of related terms such as "moth," "insect," and "wing." During free association, each word from a list is presented one at a time and the participant responds with the first term that comes to mind. When given the list word "moth," participants responded with terms such as "fly" (10 respondents), "light" (4 respondents), "wings" (2 respondents), and "summer" (2 respondents). This data set (Deese, 1965, p. 56) obtains related but different representations when analyzed and displayed by MDS or as a *PFNET* (see Fig. 4.1).

Besides clustering the terms in a different visual way, and that there are links between terms in the *PFNET* but not in the MDS representation, the two representations associate many of the terms similarly, for example, bug–insect–fly–bird–wing, moth–butterfly, and blue–sky–color–yellow. But not all clusters are the same in both, for example, bees–cocoon and butterfly–flower. The same data set produces different representations. Does it make sense to ask "Is MDS better than *Pathfinder* analysis?" "Are both representations correct?" or "Is one representation better for some purposes than the other?"

Relatedness raw data may have a large or small intrinsic dimensionality; but it is hard to visualize and think about high-dimensional relations (i.e., above three dimensions). Both MDS techniques and *PFNET* scaling are data reduction and representations approaches, but as pointed out in Fig. 4.1, the algorithms used in MDS and *PFNET* reduce the raw data dimensionality in different ways with different results. The central issue in MDS representation of relatedness data is to obtain a reduced dimensional display, usually two-dimensional for our benefit, which intends to preserve the nearness or orderings of *all* terms in the raw data. The reported stress value (a common measure of fit) indicates how well MDS was at representing the higher dimensional raw data in fewer dimensions. By convention, high stress (>0.15) means a poor representation and low stress (<0.1) indicates an adequate fit. So unless an MDS representation has zero stress, some distances among terms are distorted, and the greater the stress value observed, the greater the distortion. In general, longer distances between terms are more accurately represented than are shorter distances because the MDS algorithm uses all raw data values but magnifies the effect of large values thus giving more consideration to low relatedness raw data (Roske-Hofstrand & Paap, 1990, p. 63) at the expense of improved accuracy of the high relatedness raw data. If the stress is not too large, global clustering is likely to be good but local clustering less so, and the MDS distances between terms within a tight cluster of terms are more likely to misrepresent the relatedness raw data.

The *Pathfinder* analysis approach is nearly the opposite; closeness counts in horse shoes, hand grenades, and *PFNETs*. *PFNETs* are graphs where terms or other entities (called nodes) are joined by links (called edges) to indicate strong relationship between those terms. The "path" part of *Pathfinder* refers to the objective of the

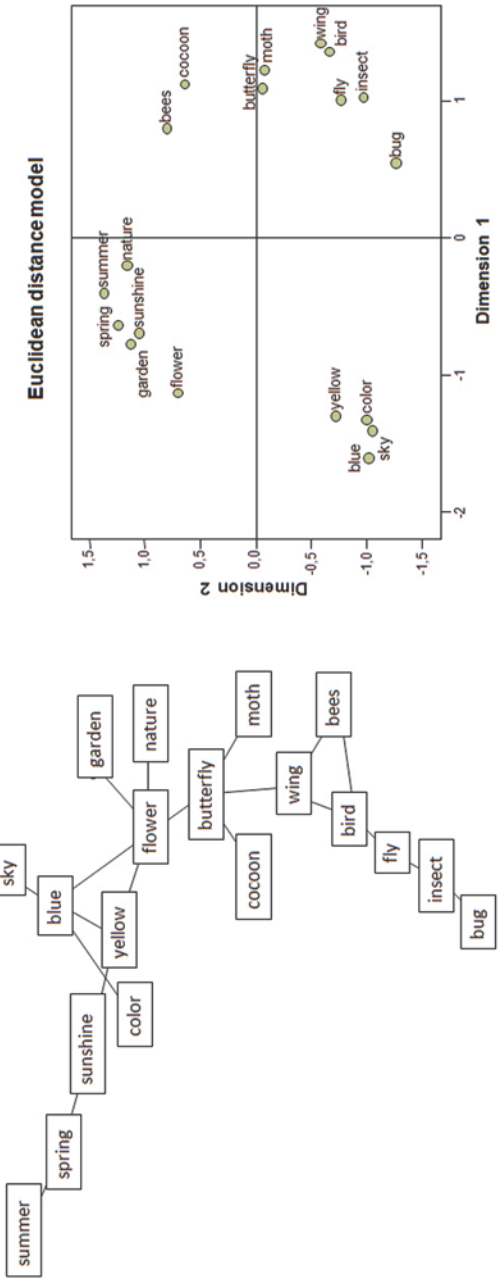


Fig. 4.1 *Pathfinder* Network representation (*left*) and multidimensional scaling (*MDS*, *right*) of the intersection coefficient association data from Deese (1965, p. 56)

approach to determine a least weighted path connecting all of the terms, thus forming a connected graph where there is a path of links to connect any node to any other node in the network. Establishing the path among all terms primarily depends on the high relatedness raw data elements (i.e., nearness), and as a result, most of the low relatedness raw data elements are disregarded in the analysis (Roske-Hofstrand & Paap, 1990). Jonassen et al. (1993, p. 74) says that *PFNETs* represent local comparisons between terms in a domain but not global information. Chen (1999, p. 408) says that *Pathfinder* analysis provides "a fuller representation of the salient semantic structures than minimal spanning trees, but also a more accurate representation of local structures than multidimensional scaling techniques." Note that the relatedness raw data by itself has concurrent validity for some cognitive outcomes, for example, predicting the order of recall of a list of terms. But *Pathfinder* analysis extracts additional psychologically valid information about the structure of memory beyond that in the original relatedness raw data while MDS does not (Cooke, Durso, & Schvaneveldt, 1986, p. 548). *Pathfinder* capability to reduce a large relatedness data set while highlighting and representing just the most critical information makes it an attractive analysis and visual representation alternative.

4.3 Alternative Approaches to Elicit Relatedness Data

Probably the most common approach used to elicit psychological relatedness data for *Pathfinder* Network analysis is pairwise comparison. To do this, participants are shown pairs of terms and are asked to indicate how related the two terms are, usually on a 1–9 scale with 1 being lowest and 9 highest. This approach obviously directs the participants' moment-to-moment decision making to the local level, which is very appropriate for follow-up *Pathfinder* analysis that focuses on these local relationships. However, the pairwise approach is tedious especially if many terms are used. For example, 15 terms require 105 comparisons while 30 terms require 435 comparisons; the number of decisions required is $n(n-1)/2$, where "n" is the number of terms considered. Note that a direct relationship between the number of terms compared and the concurrent validity of *Pathfinder* Network analysis has been reported (Goldsmith, Johnson, & Acton, 1991); more terms mean better validity.

Because there is often a need for utilizing many terms in an investigation, an elicitation approach was needed that is at least as valid as the pairwise approach but that is more efficient. Two multi-decision approaches were designed specifically to support and complement *Pathfinder* analysis data reduction and analysis: one is a listwise comparison approach and the other is a term-sorting task (Clariana, 2002; Taricani & Clariana, 2003, 2006). However, multi-decision approaches must provide more information on the screen, and this extra information will tend to establish a specific context that may influence the relatedness raw data. So the next section begins with a discussion of the likely role of context when eliciting relatedness raw data, then describes the listwise and sorting approaches, then reports two experimental investigations that utilized these multi-decision approaches,

and finally proposes combining the raw data sets from the two approaches as an optimal elicitation approach (with experimental data provided to support this idea).

4.3.1 The Role of Context

When relatedness data is elicited, we argue that the perceived context likely influences participants' relatedness responses by increasing the activation state of some ensembles (e.g., concepts, terms) and inhibiting the activation of others. This "actuated" knowledge structure arises on the fly as a subset of the participant's full internal knowledge structure in response to the purpose, task, and setting. With increased context information, some terms that may otherwise be peripheral or absent may take on a central role in the actuated knowledge structure conformation due to the influence of context. From a measurement point of view then, the way relatedness or similarity data (Gentner & Rattermann, 1991) is elicited could be goal free and context free or not, the former obtaining a fuller knowledge structure and the latter an actuated conformation (i.e., a subset of the general structure). Since context likely influences what is actually captured during an elicitation task, the context that is set by the elicitation task probably matters a great deal in the structure of knowledge that is obtained.

As far as we can tell, the role of context when eliciting relatedness data has not been previously considered nor has the likely effect of context on the structure of knowledge that is obtained been specifically examined. If context does influence the resulting knowledge structure, then context must be controlled when eliciting relatedness data.

One way to control context when eliciting relatedness data would be to include more information in the prompt, for example, by including a summary of the lesson or course content associated with the task, a purpose for the task, a case, a problem-based scenario, the list of terms to be rated, or perhaps even a story narrative. Probably in most past investigations, the elicitation tasks have used a prompt that does not intentionally set the context. For example, the Rate program provided with *Pathfinder* KNOT (2008) software says something such as "Your task is to judge the relatedness of pairs of concepts. . . . Our concern is to obtain your initial impression of overall relatedness. Therefore, please base your rating on your first impression of relatedness." The Rate program does then show the complete set of terms once at the beginning of the task, thus setting the linguistic context (Charles, 2000, p. 507), but then the list of terms is hidden as the participant completes each separate pairwise relatedness judgment for each pair of terms from the list.

Sometimes a story provides the context. In an investigation of the influence of knowledge structure on insight as measured using *Pathfinder* analysis with pairwise relatedness raw data (Dayton, Durso, & Shepard, 1990), the eliciting prompt stated, "A man walks into a bar and asks for a glass of water. The bartender pulls a shotgun on the man. The man says 'thank you' and walks out. What missing piece of

information would cause the puzzle to make sense?" (p. 269). To elicit the actuated knowledge structure for this situation, the following 14 terms (i.e., requiring 91 pairwise relatedness judgments) were presented: bar, bartender, friendly, glass of water, loaded, man, paper bag, pretzels, relieved, remedy, shotgun, surprise, thank you, and TV. There were four treatment groups that interacted passively or actively with different levels of context: "story only" who read the story (but also could not solve the puzzle) and immediately rated the 14 terms, "active nonsolvers" who read the story and then asked yes–no type questions for up to 2 h (but also could not solve the puzzle in that time) and then rated the 14 terms, "passive nonsolvers" who read the story and then listened to tape recordings of an active nonsolver asking yes–no type questions for up to 2 h (but also could not solve the puzzle), and "solvers" who read the story and then asked yes–no type questions until they solved the puzzle and then rated the 14 terms.

The results showed that only the active nonsolvers and the passive nonsolvers had strongly related but incorrect *PFNETs*; spending 2 h asking or just listening to someone else asking yes–no questions resulted in knowledge structures that were more alike. However, the solvers who also had asked yes–no questions were fairly unlike any of the other groups; solving the puzzle resulted in a very different knowledge structure (or vice versa). Specifically, for all three of the nonsolver groups, the terms "man" and "shotgun" were central high-degree nodes (with four links) while for the solver group, the term "remedy" was the central high-degree node (with four links). The solver group had correctly concluded that the man had hiccups that were then cured by fear of the bartender's shotgun. We believe that the solver group knowledge structure was initially like that of the active nonsolver and passive nonsolver groups up to the moment of solution; at that moment the solvers' knowledge structure radically shifted with this insight. Insight is a "flash of illumination" (Metcalfe, 1986, p. 239) or an "aha" experience, the dramatic and rapid reorganization of knowledge structure to fit the problem context (Dayton et al., 1990). This begs the question, once solved, is the new knowledge structure conformation fixed and fairly strongly locked in from that point on?

So in that investigation, the puzzle narrative and the list of terms were insufficient to drive a particular common knowledge conformation; thus the knowledge structure observed is each individual's own representation of the puzzle. But spending more time with yes–no questions was sufficient to drive a more similar conformation probably incrementally. Then solving the puzzle suddenly altered that specific conformation into a new and different specific conformation (or alternately, maybe the conformation shift allowed the solution to pop out).

So both too much and too little information in the prompt can influence the knowledge structure obtained. Because of the likely effects and influence of the prompt on the relatedness ratings, it seems critical to optimize the prompt with enough information to properly frame the task but not too much information that would bias the results. This could be accomplished by saying something such as "recall that the following terms were part of the lesson on ___ in order to ___" or some other such statement or story. Also, in order to establish and maintain the

linguistic context, we suggest that the list of terms to be compared should be displayed initially but should also be constantly available during the task to maintain the linguistic context. The likely influence of context was an important consideration in the design of the listwise and sorting approaches described in the next section.

4.3.2 Computer-Based Listwise and Sorting Multi-decision Approaches

The listwise approach was developed as an alternative to the pairwise approach to more efficiently elicit local relatedness data (*KU-Mapper*, 2005). Here, participants are shown a target term on the left of the computer screen and a list of all other terms on the right and are asked to pick one term from the list that is most related to the target term. Then a second term from the list becomes the target term and so on until every term has been compared to the list of terms. The raw data output for follow-up analysis is an array of "1s" (links) and "0s" (no links). As with the pairwise approach, the listwise approach obviously focuses only on local relatedness, but the listwise approach is far more efficient than the pairwise approach; 15 terms require 15 decisions and 30 terms require 30 decisions and so on. Thus if the listwise approach is valid, it should be very useful with long lists of terms (see the left panel of Fig. 4.2).

The sorting task approach was also developed to quickly elicit local and global relatedness data. Here, participants are shown all of the terms from a list randomly arranged on a computer screen and are asked to drag related terms closer together and unrelated terms farther apart, with no time limit. Essentially, the participant is asked to represent the local and global relatedness of the terms as distances. The raw data output for follow-up analysis is an array of the distances between all of the terms (see the right panel of Fig. 4.2).

4.3.3 The Effects of Headings on Knowledge Structure

Clariana and Marker (2007) used these listwise and sorting task approaches to measure the effects of learner-generated lesson headings on knowledge structure. They proposed that memory of related lesson topics would be more like the lesson topic structure for participants who generate lesson headings relative to those who do not. Generating headings during instruction influenced structural knowledge as measured by the listwise and sorting tasks in a predictable way. However, the sorting task and the listwise task obtained different *PFNET* representations, so which approach is better?

Because the lesson structure was finite and known, the lesson structure could be compared with the participants' data. To do this, two referent arrays were created, a linear referent that specified the linear order of 15 subtopics in the lesson and a nonlinear referent that specified all the possible nonlinear links between subtopics within each topic area. Based on these two referents, the listwise task was a bit better

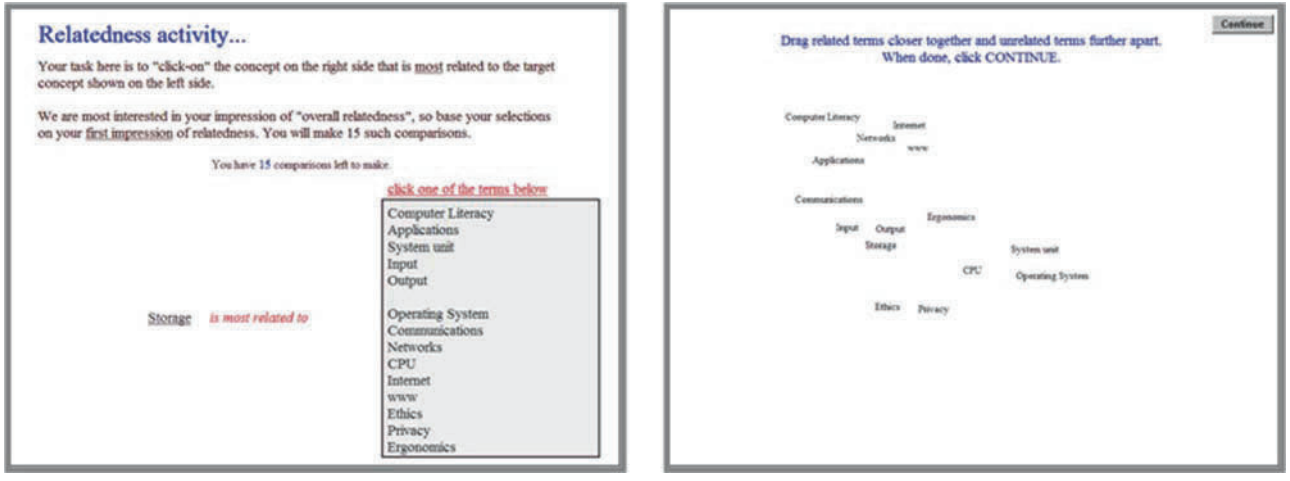


Fig. 4.2 The listwise approach (*left*) and the sorting approach (*right*)

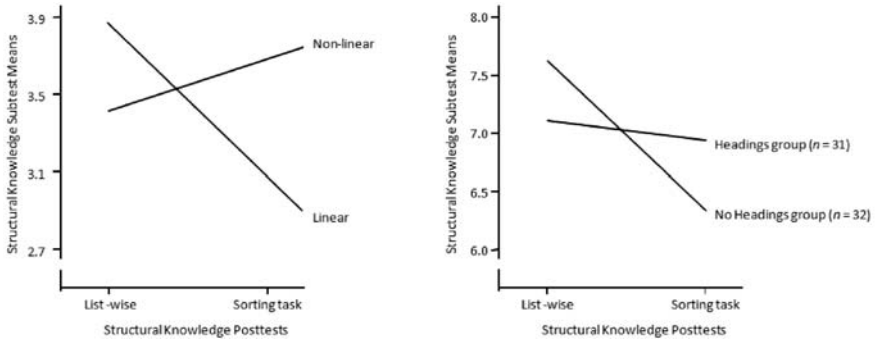


Fig. 4.3 The significant interaction of relatedness task and knowledge structure (*left*) and of the Headings treatment with knowledge structure posttests (*right*), from Clariana and Marker (2007, pp. 186–187)

at eliciting the linear subtopic structure and the sorting task was better at eliciting the nonlinear topic structure (see the left panel of Fig. 4.3).

The authors also reported a significant interaction of the Headings and No Headings treatments with the listwise and sorting tasks posttests. The No Headings group listwise task mean was significantly greater than its sorting task mean while the Headings group obtained nearly identical sorting and rating task means (see the right panel of Fig. 4.3). To account for this finding, we speculate that the text without headers treatment tends to establish a very linear knowledge structure that is more accurately measured by the listwise task, while the text with headers treatment establishes a less linear, clustered knowledge structure that is more accurately measured by the sorting task.

Besides these findings, previously unreported correlation data from Clariana and Marker (2007) presented here now (see Table 4.1) show that the listwise linear knowledge structure measure of the No Headings group correlated more with the constructed response verbatim declarative knowledge posttest (i.e., CR Posttest) than did the sorting task measure ($r = 0.62$ compared to $r < 0.24$), while for the Headings group both the sorting task and listwise measures correlated with the constructed response verbatim declarative knowledge posttest.

4.3.4 Listwise and Sorting Approaches Compared to the Pairwise Approach

Clariana and Wallace (2009) directly compared the multi-decision listwise and sorting task approaches to the more traditional pairwise approach. Undergraduate students ($N = 84$) in an introductory business course completed the three approaches in random order after taking the final examination for the course. All three of the tasks used the same 15 important terms that were covered during the course. Results indicate that the three approaches obtain knowledge structural representations by

Table 4.1 The No Headers and Headers treatment group correlations (from Clariana & Marker, 2007)

	A	B	C	D	E
No Header treatment group (N = 32)					
A. CR Posttest (15 max.)	1				
B. Sorting task (linear)	0.24	1			
C. Sorting task (nonlinear)	–0.02	–0.37*	1		
D. Listwise task (linear)	0.62**	0.30	–0.21	1	
E. Listwise task (nonlinear)	0.08	0.04	0.20	0.00	1
Header Treatment group (N = 31)					
A. CR Posttest (15 max.)	1				
B. Sorting task (linear)	0.22	1			
C. Sorting task (nonlinear)	0.49**	0.09	1		
D. Listwise task (linear)	0.44*	0.36*	0.39*	1	
E. Listwise task (nonlinear)	0.37*	0.30	0.30	0.04	1

*$p < 0.05$; **$p < 0.01$.

Pathfinder analysis that substantially overlap but are differently sensitive to linear and nonlinear knowledge structure.

First, it was reported that the two multi-decision approaches were faster than the pairwise approach, but not as fast as might be expected. The pairwise approach on average required 447.4 s (SD = 140.6), the listwise approach required 193.3 s (SD = 79.6), and the sorting task approach required 115.5 s (SD = 62.7). Next, the raw relatedness data from each task were averaged together to obtain a total group representation for each of the three approaches, pairwise, listwise, and sorting task (see Table 4.2). For total group averaged relatedness data, the listwise and pairwise approaches were most alike (71% links in common) and then listwise and sorting were next most alike (64% links in common), while the pairwise and sorting task approaches were relatively least alike (57% links in common). A linear and a nonlinear referent were created to reflect the actual structure of the 15 lesson topics as taught during the course, and the group average representations were compared to these referents. As in the earlier study, analysis of the similarity data showed that the listwise task was most sensitive to linear knowledge structure. The listwise approach

Table 4.2 Links in common (above the diagonal) and percentage of total links (below the diagonal) for each group average *PFNET* and the linear and nonlinear referents with the maximum number of links shown on the diagonal

	P	L	S	Lin	Non
Pairwise (P)	(14)	10	8	5	1
Listwise (L)	71%	(14)	9	6	1
Sorting (C)	57%	64%	(14)	5	1
Linear referent (Lin)	36%	43%	36%	(14)	1
Nonlinear referent (Non)	7%	7%	7%	7%	(11)

provided the best reflection of the actual course structure (43% of the linear referent and 7% of the nonlinear referent) and the pairwise and sorting tasks were equally reflective of the actual course structure (36 and 7%).

The *PFNET* visual representations of the group averaged data from Clariana and Wallace (2009) are shown in Fig. 4.4; the numbers beside the terms in the figure indicate the chronological order of presentation of these concepts during the course (e.g., 2a and 2b indicate that these terms were taught in the same lesson with 2a taught just before 2b). Note that the term "Internet" (e.g., 2a) was a central high-degree node (i.e., an important concept) for the pairwise and listwise *PFNET* representations (see the top panels of Fig. 4.4) but the sorting task *PFNET* has no high-degree nodes at all (see the bottom panel of Fig. 4.4). Although the listwise and sorting average group *PFNETs* were structurally quite like the pairwise average group *PFNET* (e.g., 71 and 57% overlap), inspection of *individual* participants' listwise and sorting *PFNETs* indicates that these were quite different structurally than those students' pairwise *PFNETs*. *Individual* participants' pairwise *PFNETs* were all connected graphs (i.e., there is a path from any node to any other node in the graph) with from 14 to 17 links (mode 14), with most having branching (e.g., similar in appearance to the pairwise group *PFNET* in the top left panel of Fig. 4.4). In slight contrast, nearly every *individual* participant's sorting *PFNET* is a connected graph with 14 links and no branches; all nodes have two links except for the beginning and ending node (e.g., similar in appearance to the sorting group *PFNET* in the bottom panel of Fig. 4.4). But in stark contrast, *individual* participant's listwise *PFNETs* were typically *not* connected graphs; all have exactly 15 links, and most have branching (contrast the listwise group representation in the top right of Fig. 4.4 to the *individual* representation on the left side of Fig. 4.5).

With the listwise approach, a worst case scenario for establishing a path between all nodes would be that all of the rating decisions resulted in a highly disconnected graph as the participant selects terms only within clusters of concepts, for example, selecting A–B, then B–C, then C–A. This would obtain a highly disconnected representation consisting of five separate cycle graphs and yet it *perfectly* represents the nonlinear structure of the course content and strongly represents its linear structure (see the right side of Fig. 4.5). This pattern of "clustering" was present in some participants' listwise *PFNETs* but was not common (see, for example, 10a–10b–10c configuration on the left side of Fig. 4.5), possibly because participants tend to associate terms linearly (Meyer & McConkie, 1973, p. 113) and so the listwise approach may especially elicit and capture linear associations. One problem with disconnected graphs is that it is not clear which clusters of terms are more related to what other clusters. For example, in the right panel of Fig. 4.5, it is not clear whether cluster 1–2a–2b is closer to 3–4a–4b or 10a–10b–10c; all clusters are equally disconnected from each other.

So to summarize, it seems that for individual participant's, the sorting and the listwise approaches capture separate aspects of the pairwise approach (i.e., a connected path and branching) but neither the sorting nor listwise approach fully obtains *PFNETs* that precisely resemble the individual's pairwise *PFNET*. Thus, although the sorting and especially the listwise approaches seem to be quite satisfactory

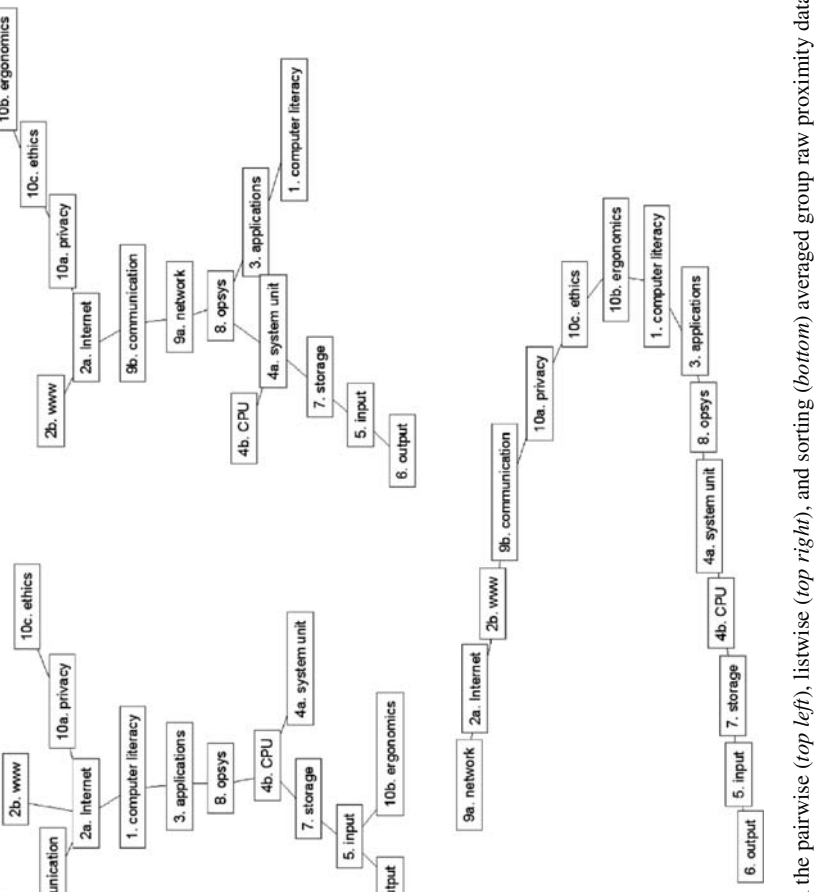


Fig. 4.4 The *PFNETs* derived from the pairwise (*top left*), listwise (*top right*), and sorting (*bottom*) averaged group raw proximity data

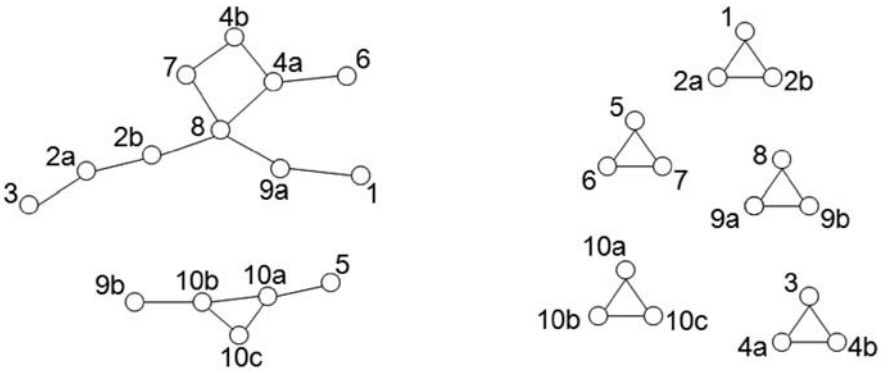


Fig. 4.5 The listwise *PFNET* disconnected graph representation of a randomly selected participant (*left*) and a hypothetical listwise *PFNET* (*right*)

for group average representations, these two approaches may be less adequate for purposes that require representations of individual's knowledge structure.

While writing this chapter, it occurred to us that combining an individual's sorting and listwise raw data should result in a *PFNET* that is a connected graph (due to the contribution of the sorting task), thus indicating which clusters go together (global), and yet maintains the most critical linear associations due to the contribution of the listwise task (local). If this approach obtains valid *PFNETs* comparable to those obtained from the pairwise approach, then for long lists of terms, this combined approach would be considerably faster to complete and thus more efficient. Thus, the relatedness data from the study by Clariana and Wallace (2009) are *reanalyzed* in order to compare the new combined sorting-plus-listwise measure to the pairwise measure; the findings of this new analysis are reported here for the first time.

4.3.5 Sorting and Listwise Combined Approach

To combine the sorting and listwise raw data, first the sorting task distance raw data for each individual were converted to a 0–1 scale by dividing each distance value in the array by the maximum distance observed in that array (e.g., a distance of 32 pixels divided by the maximum for that array of 929 is $32/929 = 0.042$) and then this scaled value is inverted by subtracting it from one ($1–0.042 = 0.958$), thus changing it from a distance/dissimilarity value where smaller values mean greater relatedness to a similarity value where larger values mean greater relatedness. Scaling and then inverting of the sorting data were necessary so that the listwise and sorting raw data would both be similarity data, and then the listwise and sorting values in each complementary cell of the two arrays can be simply added to form the new sorting-plus-listwise measure.

The participants' data from the previous investigation were reanalyzed here by randomly assigning all participants into two equivalent groups, A or B, and then Groups A and B were partitioned by median split of final course grade into high- and low-achievement groups. The pairwise and the combined relatedness raw data were averaged for each of these four groups, High Group A ($n = 20$), Low Group A ($n = 19$), High Group B ($n = 22$), and Low Group B ($n = 18$), to obtain group average *PFNETs* which were compared to each other (see Table 4.3).

Table 4.3 Percent *PFNET* overlap between high- and low-achievement Groups A and B

	Group A				Group B			
Group and approach	A	B	C	D	E	F	G	H
Group A								
A. High Group A, combined	1							
B. High Group A, pairwise	60%	1						
C. Low Group A, combined	86%	53%	1					
D. Low Group A, pairwise	83%	52%	76%	1				
Group B								
E. High Group B, combined	*93%*	60%	86%	76%	1			
F. High Group B, pairwise	79%	*53%*	64%	69%	71%	1		
G. Low Group B, combined	79%	47%	*79%*	76%	79%	57%	1	
H. Low Group B, pairwise	48%	58%	48%	*53%*	48%	48%	48%	1
Referents								
Linear referent	36%	27%	36%	35%	36%	43%	29%	35%
Nonlinear referent	24%	15%	32%	23%	24%	0%	24%	15%

The combined approach was very consistent, obtaining a 93% overlap between the group average *PFNETs* for the high achievers in Groups A and B (see Table 4.3) compared to 53% overlap for the pairwise approach and 79% overlap between the group average *PFNETs* for the low achievers in Groups A and B (see Table 4.3) compared to 53% overlap for the pairwise approach. Within-group comparisons also show higher consistency for the combined compared to the pairwise approach. Analysis of percent overlap between average group performance and the linear and nonlinear referents also indicates that the combined approach reflected the course topic coverage, both linear and nonlinear, in most cases considerably better than did the pairwise approach (about 60% compared to about 48%). To better comprehend these different data representations, the pairwise, listwise, sorting, and combined listwise-plus-sorting *PFNETs* of the student with the best course grade are presented (see Fig. 4.6).

This student's listwise *PFNET* dominated the combined *PFNET* structure, except that the links between terms 7 and 9a and between 2a and 3 in the listwise *PFNET* did not occur in the combined *PFNET*; also the two separate portions of the listwise *PFNET* were joined between terms 1 and 10b (see Fig. 4.6). This top student's

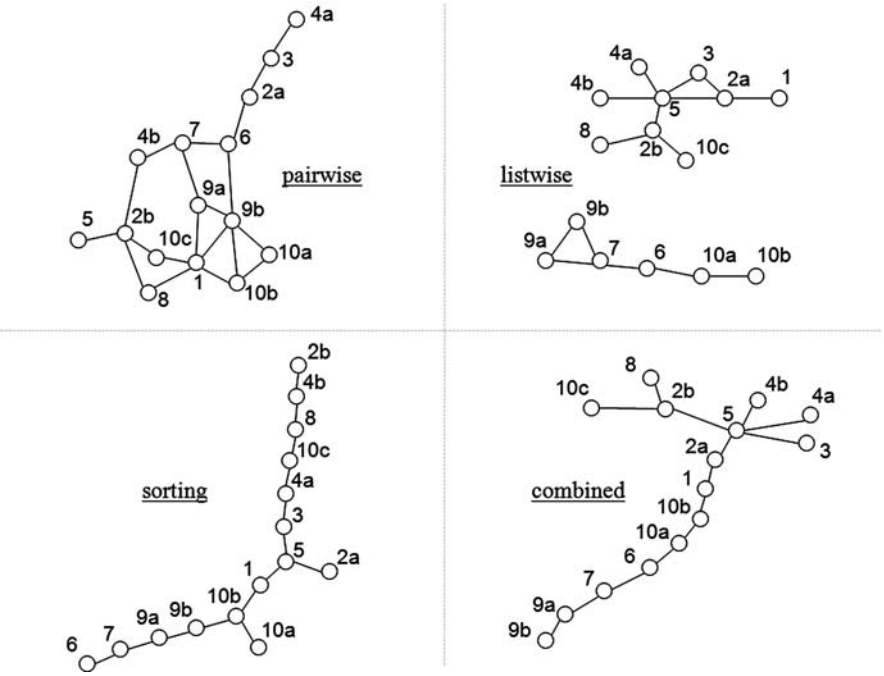


Fig. 4.6 Pairwise, listwise, sorting, and combined listwise-plus-sorting *PFNETs* of the top-performing student

pairwise *PFNET* has six links that match the linear and one that matches the nonlinear course referents and 13 links that do not match the course structure (i.e., *pairwise PFNET* course organization, 7/20 = 35% overlap), while her combined *PFNET* has five links that match the linear and one that matches the nonlinear course organization and eight links that do not match the course structure (i.e., *combined PFNET* 6/14 = 43% overlap). The pairwise *PFNET* has two central (high-degree) nodes, terms 9b and 1, while the combined *PFNET* has one central node, term 5.

Thus the combined sorting-plus-listwise approach appears to provide a better measure than the pairwise approach that is more consistent across the two groups (e.g., a measure of reliability) and has better concurrent validity. However, an alternate explanation is that two measures are better than one! Possibly using the pairwise rating approach twice and combining the two runs would result in better (more reliable and valid) pairwise comparisons too; we can't be certain. However, recall that the pairwise rating approach is notoriously time consuming for the participants to complete and the original design intent of the sorting and listwise approaches was to improve efficiency, i.e., to obtain data that are comparable to the pairwise approach but require less time to complete. This data and analysis indicates that the combined sorting-plus-listwise approach achieves this efficiency objective; but considerable further research is needed to confirm this.

4.4 Summary and Conclusion

After considering several methods of eliciting and representing knowledge, our present view of knowledge structure best aligns with the *Pathfinder* analysis approach. We developed a computer program called *KU-Mapper* to implement two multi-decision approaches based on our view of knowledge structure that complements *Pathfinder* analysis, a sorting task and a listwise task. During this development, we realized the likely influence of internal and external context on knowledge structure. Because of the sparseness of most elicitation tasks, too little internal instructions/directions tend to misdirect participants, and the knowledge structures obtained would not accurately represent their knowledge structure for the domain area, while too much internal or external context information biases the knowledge structure obtained toward those context variables rather than capturing the participants' knowledge structure (the Goldilocks' principle). To obtain an optimum, we determined that the listwise and sorting tasks should include a brief descriptive review of the task domain and that it should include all of the list terms on every screen in order to maintain the lexical context during the elicitation task.

An investigation by Clariana and Marker (2007) suggests that the listwise task is better at eliciting and representing linear knowledge structure while the sorting task better elicits and represents nonlinear (clustered) knowledge structure. A moderately strong correlation ($r = 0.62$) was noted between the listwise *PFNET* measure and the constructed response verbatim declarative posttest for the group who did not generate headings. Possibly, generating headings while reading tends to shift participants' knowledge structure from linear to nonlinear, and this shift may account for these and for some previous findings of the effects of headings on various kinds of posttest measures. Instructors and researchers should be made aware that eliciting knowledge structure probably alters knowledge structure (an intervening test effect).

An investigation by Clariana and Wallace (2009) directly compared the multi-decision listwise and sorting tasks to the traditional pairwise approach. Though individual listwise, sorting, and pairwise *PFNETs* were not strongly related; group average listwise, sorting, and pairwise *PFNETs* were strongly related, with the pairwise and listwise group average *PFNETs* sharing a 71% overlap.

Because of the likely limitations of individual participant's *PFNETs* obtained using the listwise and sorting approaches, a new approach was suggested that combined the relatedness data from both. Previous raw data from Clariana and Wallace (2009) were used to generate the combined listwise-plus-sorting data set and these new data were compared to the pairwise data using a group average approach. The combined *PFNETs* were considerably more consistent across equivalent groups than the pairwise *PFNETs*. These preliminary results suggest that the combined multi-decision approach may be an adequate substitute for the pairwise comparison approach especially when the list of terms is long (20 or more terms). These results support further research to confirm or refute this combined approach. Our hope is that these approaches will be validated and extended and that other investigators

will incorporate these ideas and methods into future software tools to advance the systematic analysis of knowledge through new multi-decision approaches.

References

Anderson, R. C. (1984). Some reflections on the acquisition of knowledge. *Educational Researcher, 13*(10), 5–10.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics, 21*, 505–524.

Chen, C. (1997). Tracking latent domain structures: An integration of pathfinder and latent semantic analysis. *Artificial Intelligence and Society, 11*, 48–62.

Clariana, R. B. (2002). *ALA-Mapper software,* version 1.01. Retrieved September 28, 2008, from http://www.personal.psu.edu/rbc4/ala.htm

Clariana, R. B., & Marker, A. (2007). Generating topic headings during reading of screen-based text facilitates learning of structural knowledge and impairs learning of lower-level knowledge. *Journal of Educational Computing Research, 37*(2), 173–191.

Clariana, R. B., & Wallace, P. E. (2009). A comparison of pairwise, listwise, and clustering approaches for eliciting structural knowledge. *International Journal of Instructional Media, 36*, 287–302.

Cooke, N. M., Durso, F. T., & Schvaneveldt, R. W. (1986). Recall and measures of memory organization. *Journal of Experimental Psychology, 12*, 538–549.

Dayton, T., Durso, F. T., & Shepard, J. D. (1990). A measure of the knowledge reorganization underlying insight. In R. W. Schvaneveldt (Ed.), *Pathfinder network associative networks: Studies in knowledge organization* (pp. 267–277). Norwood, NJ: Ablex Publishing.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, MD: The Johns Hopkins Press.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development*. London: Cambridge University Press.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*, 88–96.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

KNOT (2008). *Knowledge network organizing tool*. Retrieved December 24, 2008, from http://interlinkinc.net/Ordering.html

KU-Mapper (2005). *Knowledge unit mapper version 1.01*.Retrieved December 26, 2008, from http://www.personal.psu.edu/rbc4/KUmapper.htm

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 623–634.

Meyer, B. J. F., & McConkie, G. W. (1973). What is recalled after hearing a passage? *Journal of Educational Psychology, 65*, 100–117.

Mitchell, A. A., & Chi, M. T. (1984). Measuring knowledge within a domain. In P. Nagy (Ed.), *The representation of cognitive structure* (pp. 85–109). Toronto, Canada: Ontario Institute for Studies in Education.

Roske-Hofstrand, R. J., & Paap, K. R. (1990). Discriminating between degrees of low and high similarity: Implications for scaling techniques using semantic judgements. In R. W. Schvaneveldt (Ed.), *Pathfinder network associative networks: Studies in knowledge organization* (pp. 61–73). Norwood, NJ: Ablex Publishing.

Rumlehart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models* (pp. 7–57). Cambridge, MA: MIT Press.

Schvaneveldt, R. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex Publishing.

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology, 63*, 225–234.

Shavelson, R. J. (1974). Methods for examining representations of subject-matter structure in a student's memory. *Journal of Research for Science Teaching, 11*, 231–249.

Sternberg, R. J., & Preiss, D. D. (2005). *Intelligence and technology: The impact of tools on the nature and development of human abilities*. Mahwah, JN: Lawrence Erlebaum.

Taricani, E. M., & Clariana, R. B. (2003, October). *Semantic map automated assessment techniques*. Presented at the annual conference of the Association for Educational Communications and Technology (AECT) in Anaheim, CA.

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development, 54*, 61–78.

Chapter 5
The Problem of Knowledge Elicitation from the Expert's Point of View

J. Vrettaros, A. Leros, K. Hrissagis-Chrysagis, and A. Drigas

5.1 Introduction

The aim of the e-learning environment under study is the implementation of an educational system suitable not only for teaching but also for the evaluation of teaching the English language to deaf individuals. This chapter focuses on the assessment part where an expert system is being used. Expert systems technology is a subfield of artificial intelligence which is based on the idea that knowledge can be transmitted from a human to a computer. The actual aim of expert systems is the implementation of an e-consultant who not only will give advice but also will give explanations if necessary (Turban & Aronson, 2001). The proposed expert system aims at achieving assessment of deaf students in the context of teaching English. The significant components of this expert system are a database centralizing all questions and possible answers, a database including tutorials/lessons, an interface as well as the significant part under study, namely a neural system and/or a neurofuzzy system that allows the system to make trustworthy inferences.

Below we briefly present research conducted so far on the field of student assessment and the wider field of student modeling using artificial intelligence techniques.

Indeed, artificial intelligence has proved to be a fruitful tool when applied to current educational research streams such as student modeling, natural language dialogue (language processing for simulating human dialogues), cognitive modeling (for human thinking simulation), complete systems and evaluation, as well as authoring tools, knowledge acquisition, and development tools (Lane, 2006). Among the fields mentioned above, student modeling seems to be one of the greatest challenges for researchers since it is considered a keyword for personalized interaction between humans and a hypermedia system and consequently for adaptive learning, which has proved to be an efficient way to maximize learning results

J. Vrettaros (✉)
NCSR Demokritos, Institute of Informatics and Telecommunications, Net Media Lab, Paraskevi, Greece
e-mail: jvr@iit.demokritos.gr

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_5,

(Frias-Martinez, Magoulas, Chen, & Macredie, 2005). Student modeling consists of the student model and the diagnostic part which performs student diagnosis (Stathakopoulou, Magoulas, Grigoriadou, & Samarakou, 2004).

The student model is one of the components of an Intelligent Tutoring System which provides a description of student-related information such as his knowledge level, skills, or even preferences, while diagnosis is the inference process which in the end updates the student model.

In order for this kind of intelligence to be achieved, researchers have adopted many Artificial Intelligence methods. The most famous among them are neural networks, fuzzy logic, as well as several search methods such as genetic algorithms.

Neural networks are on the top of the researchers' choice since they provide a system with the ability to recognize patterns, to derive meaning from vague data, and to identify matching in similar cases (Frias-Martinez et al., 2005). Fuzzy set theory is widely used since it can deal in a reliable way with human uncertainty and it obtains smooth modeling of human decision making. Genetic algorithms are ideal for optimal expert knowledge representation. Finally, neurofuzzy synergism is getting more and more popular in this area since it seems to overcome obstacles that come up when each of the methods involved is solely applied (Al Hamadi & Milne, 2004). Below we present several typical examples of the application of these methods in student's diagnosis.

A neurofuzzy system has been used in order to obtain maximization of adaptability in business education tutoring. For the training of the network the student's data that come up during interaction are being used (Kinshuk, Nikov, & Patel, 2001).

Grigoriadou et al. incorporated fuzzy logic and multicriteria decision making in INSPIRE (Intelligent System for Personalized Instruction in a Remote Environment), a web-based Adaptive Educational Hypermedia System (Grigoriadou, Kornilakis, Papanikolaou, & Magoulas, 2002).

Mir Sadique and Ashok applied the architecture of the Adaptive Neuro Fuzzy Inference System (ANFIS) in the field of Intelligent Tutoring Systems. The system that came up examined learners' memory, concept understanding, and possible deficiencies and finally obtained reliable classification of their performance (Mir Sadique & Ashok, 2004).

A system implementing a Neural Network Genetic Programming method has also been proposed, aiming at creating a reliable evaluation tool substituting an e-tutor. The system was trained through data extracted from an educational project called DEDALOS and through the assessment given by an expert (Vrettaros, Pavlopoulos, Vouros, & Drigas, 2008).

5.2 Description of the System and Knowledge Elicitation

According to the structural and functional details of the e-learning procedure, the attribution factor of an expert system and therefore, of the e-learning model's synergy, is the codification and the content of the input and output of an expert system

as well as the structure, formulism and content of the questionnaires database which require further attention and adroit handling.

The evaluation procedure of teaching deaf individuals pertains to the accomplishment of ESOL (English for Speakers of Other Languages) models of levels 1 and 2. Those levels consist of five sections which with ascending order of priority are [*A*], [*B*], [*C*], [*D*], and [*E*]. Section [*A*] represents the letter recognition and alphabetical order, section [*B*]represents spelling and vocabulary, section [*C*] represents grammar and sentence structure, section [*D*] represents reading, and section [*E*] represents writing.

According to the e-learning environment specifications of ESOL, the input and output parameters of an expert system can be specified undoubtedly, while at the same time their translation is simple and direct enough.

About the input, altogether per question there are five couples of parameters, which are: $a = \{a_{\text{val}}, a_{\text{rel}}\}$, $b = \{b_{\text{val}}, b_{\text{rel}}\}$, $c = \{c_{\text{val}}, c_{\text{rel}}\}$, $d = \{d_{\text{val}}, d_{\text{rel}}\}$, and $e = \{e_{\text{val}}, e_{\text{rel}}\}$.

That is to say, each couple answers to a section of the language of a specific level. Parameter a describes the letter recognition and alphabetical order of section [*A*], parameter b correlates with spelling and vocabulary of section [*B*], parameter c represents grammar and sentence structure of section [*C*], the respective parameter for reading of section [*D*] is d, while the ability of writing of section [*E*] is quantified with parameter *e*. The index (value) represents the evaluation of the particular section according to a given answer, while the index (relevance) recognizes the grade of relevance/weight of a specific question among the contents of a section.

The evaluation values of the input parameters a_{val}, b_{val}, c_{val}, d_{val}, and e_{val} derive from the universe of discourse $S = \{-1\} \cup [0,1]$. If a section is not examined by a question of the respective parameter, the domain is defined with the value –1. An answer which is incorrect according to a section leads to a respective value zero (0), while the value of the parameter of a section is one (1) if the chosen answer is correct according to that section. Similarly, answers which are partially correct have their values lie in the interval [0,1]

On the other hand, one could claim that the relevance parameters a_{rel}, b_{rel}, c_{rel}, d_{rel}, and e_{rel} characterize the question instead of the probable answers. Although that is true, the negotiation with relative parameters as a part of a given answer is convenient and more governable from evaluation point of view (as further explained below). As a result, the relevance/weight is considered to vary in the interval [0,1], where the value zero (0) or values near zero mean low relevance, value one (1) or values near one mean high relevance and all the other values of weight similarly vary between. However, it should be underlined that the relevance parameters are common and same for all the answers to a given question.

Based on the above, for *single-select questions*, the craftiest method for information supply (records) in the input of an expert system, relative to the five sections, is the sequence in an ordered form by ten values for the parameters of the input pairs:

a_{val}	a_{rel}	b_{val}	b_{rel}	c_{val}	c_{rel}	d_{val}	d_{rel}	e_{val}	e_{rel}

For example, let's consider a question that exhibits low relevance to section *A*, high relevance to section *C*, and medium relevance to section B. Let's also suppose that the question under consideration does not contain information about sections *D* and *E*. Now, let's also consider an answer to the previous question which is correct according to section *A*, partially correct according to section *C*, incorrect according to section *B*, and obviously does not contain any information about sections *D* and *E*. Such an answer results in a sequence set of ten values in the universe of discourse $S = \{-1\} \cup [0,1]$. In addition, it is obvious that the above-mentioned sequence of the ten values can be directly coded as a numerical string similar to the following:

String of single-select questions									
a_{val}	a_{rel}	b_{val}	b_{rel}	c_{val}	c_{rel}	d_{val}	d_{rel}	e_{val}	e_{rel}
1	0.1	0	1.5	0.7	0.9	–1	0	–1	0

This way, the specific string of single-select questions can be easily imported as input to an expert system.

For *multi-single-select questions*, according to the specifications of e-learning environment by ESOL, the craftiest method for information supply in an expert system, relative to the five sections above, is the sequence of the input parameters in an arranged form, as follows:

String of multi-select questions			
Relevance values of sections	Correct answer	Evaluated learner's answers	Evaluation values of sections for every answer
$a_{rel}b_{rel}c_{rel}d_{rel}e_{rel}$	$c_1c_2c_3c_4...c_n$	$a_1a_2a_3a_4...a_n$	$a_{val}b_{val}c_{val}d_{val}e_{val}$

In the above codification of multi-select questions for information supply (records) as input in an expert system in an e-learning environment, the values in the second column (correct answer) and in the third column (learner's answers) are in binary form, i.e., "0" or "1", where "0" means "FALSE" and "1" means "TRUE". The first bit, $c_1$of the correct answer or a_1 of the learner's answer, refers to the first answer of the selected question. The second bit, c_2 or a_2, refers to the second selected answer, etc. The last bit, c_n or a_n, refers to the *n*th answer of the selected question. It is obvious that for each question with n multi-selections, the above binary codification of the correct answer is just one, but the number of all the learner's possible answers is 2^n. Among those possible answers only several are noticeable while the rest of them are considered as irrelevant, selection that is always handled with caution by an expert pedagogical.

Hence, in multi-select questions for information supply in the input of an expert system in an e-learning environment, there are records which consist of five (5) sections with the five (5) sections relevance values, of the binary codification

domains of the correct answer with n bits, which are as many as the select answers of the questions, of the binary codification domains of the learner's answers with n bits, and of n times the five (5) domains with the five (5) sections evaluation values, which are as many as the select answers of the question.

As an example, for a question with relevance values [4,0,0,1,0] respectively for the five (5) sections a_{rel}, b_{rel}, c_{rel}, d_{rel}, and e_{rel}, with six (6) select answers where the correct ones are 2, 3, and 6, i.e., $c_1c_2c_3c_4c_5c_6 = [011001]$, while the learner's choice as correct answers are 1, 2, and 3, i.e., $a_1a_2a_3a_4a_5a_6 = [111000]$, and with six (6) times the evaluation values $[0, -1, -1,1, -1]$ respectively for the five (5) sections a_{val}, b_{val}, c_{val}, d_{val}, and e_{val} for each time, the record for information supply in the input of an expert system in an e-learning environment will have 47 domains arranged in the form:

$$[4,0,0,1,0][011001][111000][0, -1, -1,1, -1][0, -1, -1,1, -1]$$
$$[0, -1, -1,1, -1][0, -1, -1,1, -1][0, -1, -1,1, -1][0, -1, -1,1, -1]$$

About the output of both types of questions, single-select questions and multi-select questions, the observation or even the monitoring on the functional and relevant characteristics of an expert system leads to the conclusion that the output parameters of the system are six (6), nominally y_1, y_2, y_3, y_4, y_5, and y_6. The first five parameters are the evaluations/estimation of the language skills per section, while the sixth parameter represents the overall evaluation of the user's overall language skills, as follows:

y_1 = letter recognition and alphabetical order skills
y_2 = spelling/vocabulary skills
y_3 = grammar/sentence structure skills
y_4 = reading skills
y_5 = writing skills
y_6 = overall language skills *(in fact, it is the weighted average of $y_1 - y_5$, representing a general evaluation of the learner's language level, as an expert pedagogical would define it in a real scenario).*

It is obvious that the output parameters are continuous. The evaluation is considered to be normalized in the continuous interval [0,1], because of the fact that the outputs of an expert system represent an estimation which is related to a specific language section. The translation of the final numerical values is simple: zero means no language skills, one means perfect language skills, whereas all other levels of language skills can be evaluated using similar numerical interferences. The output values, which are already numerically encoded, can be inputted to the e-learning environment as an arranged array of six values:

y_1	y_2	y_3	y_4	y_5	y_6

Let us suppose that the next evaluation is real for a student:

0.6 = letter recognition and alphabetical order skills
0.4 = spelling/vocabulary skills
0.2 = grammar/sentence structure skills
0.5 = reading skills
0.3 = writing skills
0.4 = overall language skills

This six-valued sequence, which is ordered and includes continuous elements, could be directly encoded as a numerical string, similar to the following:

y_1	y_2	y_3	y_4	y_5	y_6
0.6	*0.4*	*0.2*	*0.5*	*0.3*	*0.4*

This way, final outputs are directly available to the rest e-learning environment.

It is considered by pedagogical experts that a learner who selects one combination of answers could show more or less understanding than a learner who selects another. Training data values are assigned to specific combinations of answer options.

Evaluation values and training data values are the same for single-select questions as there is only one correct answer option. However, in multi-select questions evaluation values and training data values may differ. In multi-select questions more than one answer is required in order to be completely correct. The learner may still demonstrate partial understanding by selecting say two out of three correct answers.

In this example, "Which three adjectives can you use to describe a car?" the training data values are assigned to five answer option combinations. A, B, C, D, and E refer to the learning areas while OS is an overall skill value and represents the pedagogical expert's view of the learners overall language skills based on the combination of selected answers (Table 5.1).

Which three adjectives can you use to describe a car?

Even though combinations 2, 3, and 4 are not completely correct, the pedagogical expert considered that they demonstrated an understanding of the question and assigned positive values to them. If the learner selects any other combination, data values of 0 are assigned for areas that are relevant to the question and –1 for areas that are not. 0 is assigned as an overall skill value.

The above discussion, which is according to the ESOL specifications, for the adroit codification and the content of the inputs and the outputs, as well as the structure, formulism, and content of the questionnaires database, pertains to the use of neural networks and neurofuzzy technologies for modeling the input–output relation of the e-learning expert system for the automatic prediction of evaluation values of teaching the English language to deaf individuals. Indeed, neural networks and neurofuzzy models are a very fruitful choice when it comes to mining complex patterns in noisy or incomplete data (Frias-Martinez et al., 2005).

Table 5.1 Encoding of students' answers

Answer option combination code	Answer option combination	A	B	C	D	E	OS
1	New (correct) Smart (correct) Small (correct)	–1	1	1	1	–1	1
2	New (correct) Young (incorrect) Smart (correct)	1	0.6	0.6	0.6	–1	0.6
3	New (correct) Smart (correct) Happy (incorrect)	–1	0.6	0.6	0.6	–1	0.6
4	New (correct) Young (incorrect) Small (correct)	–1	0.6	0.6	0.6	–1	0.6
5	All other combinations	0	–1	–1	0	–1	0

Neural networks and neurofuzzy technologies have already been successfully applied to many prediction problems with similar inputs/outputs features (Shavlik & Eliassi, 2001). The present report studies the purpose of applying the neural networks and neurofuzzy technologies on modeling the automatic evaluation of deaf individuals' answers in questions on five sections in an e-learning environment of an expert system.

5.3 Language Skills Database

The e-learning environment's semantic context core is situated around the utilized Language Skills Database. A closer insight reveals the two constituting elements of language skills database, which namely are the Questionnaires Knowledgebase (denoted as QK) and the Tutorials/Lessons Database (denoted as TLDB). During all phases of the proposed e-learning process, questions or questionnaires are interchanged interactively with corresponding lesson/tutorial sessions.

Questionnaires knowledgebase contains the whole series of questions and possible corresponding answers, regarding all three learning phases. Questionnaires knowledgebase is associated directly to the inputs of the expert system, namely, users' answers are applied to the expert system's inputs after a trivial transformation.

Teaching sessions, skill tutorials, and language lessons comprise the tutorials/lessons database. Though the connection between the expert system's outputs and the contents of tutorials/lessons database is not directly visible, it exists and moreover proves determinant. A specific teaching/instructing session or lesson/tutorial sequences of such sessions are controlled by the output values. Suppose a pedagogical expert has defined certain thresholds that determine the basis for

language skills sections. By taking into consideration both factors (the expert system's outputs and the pedagogical expert's thresholds), the language sections whose assessment is not satisfactory need further teaching/tutoring. Consequently, the appropriate elements which are designed so as to enhance the individual's partial language knowledge are being retrieved from tutorials/lessons database and furthermore are being utilized appropriately by the e-learning environment. The structure of language skills database is presented below Fig. 5.1.

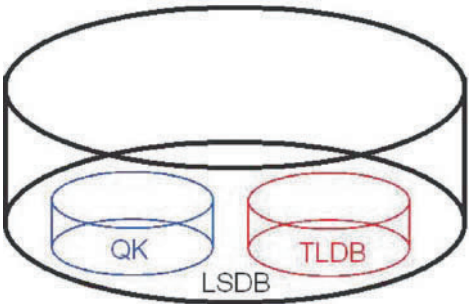


Fig. 5.1 Language skills database

5.4 Adaptive Fuzzy E-Learning Subsystems

The technology of fuzzy inference systems is a popular computing framework based on the concepts of fuzzy set theory, fuzzy if-then rules and fuzzy reasoning. A typical fuzzy inference system for knowledge processing follows three stages: fuzzification of the input data, conduction of fuzzy inference based on fuzzy data, and defuzzification of the output in order for the final outcome to be produced (Frias-Martinez et al., 2005). Fuzzy logic has found successful applications in a wide variety of fields such as control systems (Bugarin & Barro, 1998), medical diagnosis (Meesad & Yen, 2003; Sendelj & Devedzic, 2004), job matching (Drigs, Kouremenos, Vrettos, & Kouremenos, 2004), computer security (Reznik & Dabke, 2004), user modelling (Kuo & Chen, 2004; Vrettos & Stafylopatis, 2002), etc. Because of its multidisciplinary nature the fuzzy inference system is known by numerous other names, such as fuzzy-rule-based system, fuzzy expert system, fuzzy model, fuzzy associative memory, and simply fuzzy system.

The basic structure of a fuzzy inference system consists of three conceptual components: a rule base, which contains a selection of fuzzy rules; a database (or dictionary), which defines the membership functions used in the fuzzy rules; and a reasoning mechanism, which performs the inference procedure upon the rules and given facts to derive a reasonable output or conclusion.

A fuzzy inference system implements a nonlinear mapping from its input space to output space. This mapping is accomplished by a number of fuzzy if-then rules, each of which describes the local behavior of the mapping. In particular, the antecedent

of a rule defines a fuzzy region in the input space while the consequent specifies the output in the fuzzy region.

In general, the designing of a fuzzy inference system is based on the (possibly partial) known behavior of the target system. The target system under consideration is the language skill evaluation/assessment expert subsystem of the e-learning environment. The fuzzy system is then expected to be able to reproduce the behavior of the target system.

Generally speaking, the standard method for constructing a fuzzy inference system, a process usually called fuzzy modeling, has the following feature: The rule structure of a fuzzy inference system makes it easy to incorporate human expertise on the target system directly into the modeling process, namely, fuzzy modeling takes advantage of domain knowledge that might not be easily or directly employed in other modeling approaches. A key point in designing/defining the proposed system's rule base is presented in the axioms below:

Axiom1: Evaluation of a section using an answer takes place only when information regarding the specific section (possibly among other section(s)) is available.

Axiom2: Only meaningful input variables (namely, those with values other than –1) are manipulated by the expert system.

The afore-presented axioms drastically reduce the maximum number of fuzzy if-then rules which can be constructed when taking under consideration the type and amount of input and output variables. Moreover, these axioms delineate the selection of teaching/tutoring sequences, since poor performance in certain language sections is confronted only with teaching sessions (taken from tutorials/lessons database) affecting comprehension of the specific sections.

The proposed rule base which is going to be utilized by the fuzzy system employs fuzzy if-then rules. Note that, in order to accomplish the creation of such a rule base one must rely on preexisting knowledge of the e-learning environment, information provided by pedagogical experts who are familiar with the e-learning environment, or simply trial and error.

After this first stage of fuzzy modeling, the obtained rule base can more or less describe the behavior of the e-learning environment by means of linguistic terms. Further refinement of the rule base is carried out during the second stage, the identification of the deep structure. Specifically, the identification of the deep structure means refining the parameters of the inference system using regression and optimization techniques (adaptation stage).

Literally, the proposed expert system, which is part of the general e-learning environment, demonstrates functionality equivalent to adaptive fuzzy inference systems. Correspondingly, the proposed architecture/model is referred to as Adaptive Fuzzy e-Learning Subsystem (denoted as AFELS). The proposed interconnection and interrelation between the adaptive fuzzy e-learning subsystem architecture (AFELS) and the remaining e-learning environment is illustrated in the next page, mainly for demonstration and clarification purposes. Also, through Fig. 5.2 essential operating issues are presented clearly.

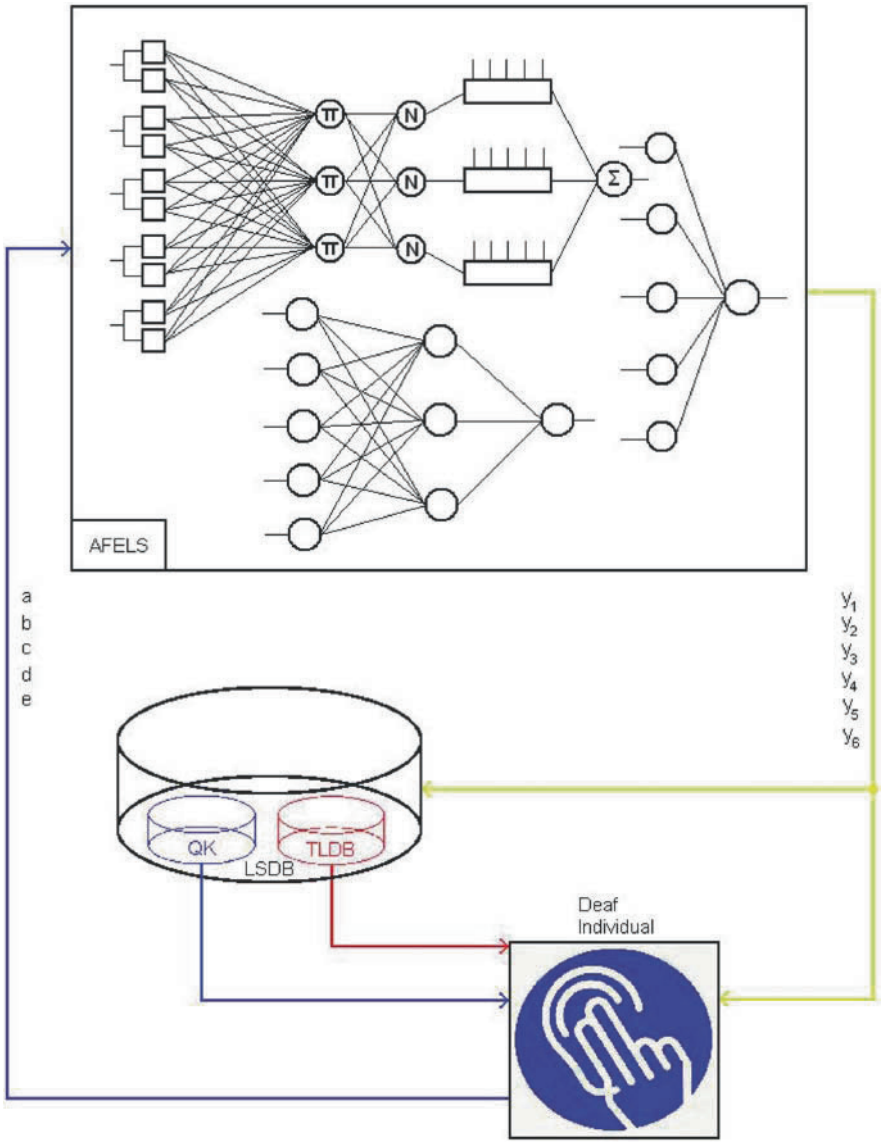


Fig. 5.2 Structure of adaptive fuzzy E-learning subsystems

5.5 Supervised Learning Schema

An adaptive network, like adaptive fuzzy e-learning subsystem, is a network structure consisting of a number of nodes connected through directional links. Each node represents a processing unit and the links between nodes specify

the causal relationship between the connected nodes. Moreover, the outputs of these nodes depend on modifiable parameters pertaining to these nodes. The learning rule specifies how these parameters should be updated to minimize a prescribed error measure, which is a mathematical expression that measures the discrepancy between the network's actual output and desired output.

Conceptually, a feed forward adaptive network is actually a mapping between its input and output spaces. A supervised learning algorithm's aim is to construct a network for achieving a desired nonlinear mapping that is regulated by a data set consisting of desired input–output pairs of the target system to be modeled. This data set is usually called the training data set, and the procedures followed in adjusting the parameters to improve the network's performance are often referred to as the learning rules or adaptation algorithms.

As already mentioned, usually, a network's performance is measured as the discrepancy between the desired output and the network's output under the same input conditions. This discrepancy is called the error measure and it can assume different forms for different applications. Generally speaking, a learning algorithm is derived by applying a specific optimization technique to a given error measure. In the proposed expert system, the scope is confined to modeling problems with desired input–output data sets, so the resulting adaptive fuzzy e-learning subsystem has adjustable parameters that are updated by a supervised learning rule. Such networks are often referred to as supervised learning or mapping networks for obvious reasons.

In order to successfully accomplish the adaptation task, as described briefly in the previous paragraphs, an extensive series of input–output pairs representing the e-learning environment is necessary. Training data could be of any form or format as long as mandatory information is included.

Let $v, v', w, w', x, x', y, y', z, z' \in \{-1\} \cup [0,1]$ and *h, i, j, k, l,* $n \in [0, 1]$ then each pattern of the training set could be similar to the following template tuple (where ø denotes an empty value):

a	*b*	*c*	*d*	*e*	
(v, v')	(w, w')	(x, x')	(y, y')	(z, z')	
y_1	y_2	y_3	y_4	y_5	y_6
h if $v \tilde{\neq} 1$ else ø	*i* if $w \tilde{\neq} 1$ else ø	*j* if $x \tilde{\neq} 1$ else ø	*k* if $y \tilde{\neq} 1$ else ø	*l* if $z \tilde{\neq} 1$ else ø	*n*

A sample training set is illustrated in a tabulated form below. It is apparent, that information encapsulated in such a training data set should be collected and preprocessed by a pedagogical expert since such an expert appears as the most suitable person for creating the afore-mentioned content (Table 5.2).

Table 5.2 Sample training set

Pattern (#)		1	2	3	4	5	6	7
a	a_{val}	0.6	0.1	1.0	0.9	0.6	0.1	0.1
	a_{rel}	0.5	0.5	0.5	0.7	0.7	0.7	0.2
b	b_{val}	0.8	0.7	0.1	–1	–1	0.8	–1
	b_{rel}	0.8	0.8	0.8	0	0	0.6	0
c	c_{val}	0.6	0.8	0.2	–1	–1	0.9	0.1
	c_{rel}	0.1	0.1	0.1	0	0	0.5	0.8
d	d_{val}	–1	–1	–1	0.9	0.1	0.6	–1
	d_{rel}	0	0	0	0.5	0.5	0.4	0
e	e_{val}	–1	–1	–1	0.6	0.8	0.2	0.9
	e_{rel}	0	0	0	0.3	0.3	0.1	0.3
	y_1	0.9	0.2	0.3	0.5	0.8	0.4	0.7
	y_2	0.7	0.8	0.2	ø	ø	0.1	ø
	y_3	0.4	0.5	0.2	ø	ø	0.1	0.5
	y_4	ø	ø	ø	0.9	0.0	0.1	ø
	y_5	ø	ø	ø	0.6	0.2	0.6	0.8
	y_6	0.1	0.5	0.7	0.1	0.2	0.9	0.7

5.6 Conclusions

The system proposed in this chapter has been applied in DEDALOS, an EU program in the framework LEONARDO DA VINCI with very encouraging results. Indeed, the use of neural and neurofuzzy technologies proved to be very fruitful when it comes to simulating the knowledge of expert if we succeed in mining the existing knowledge patterns as well as in using appropriate data codification. In the present research, the volume of the available data has been delimited so it is considered as high future priority the enrichment of the input/output data so as to achieve higher success rate and to come to more secure conclusions.

References

Al Hamadi, A. S., & Milne, R. H. (2004). A neuro – fuzzy classification approach to the assessment of student performance. *In Proceedings of the IEEE International. Conference on Fuzzy Systems, 2,* 837–841.

Bugarin, A. J., & Barro, S. (1998). Reasoning with truth values on compacted fuzzy chained rules. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 28*, 34–46.

Drigs, A., Kouremenos, S., Vrettos, S., & Kouremenos, D. (2004). An expert system for job matching of the unemployed. *Expert Systems with Applications, 26*, 217–224.

Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2005). Modeling human behavior in user – Adaptive systems: Recent advances using soft computing techniques. *Expert Systems with Applications*, *29*(2), 320–329.

Grigoriadou, M., Kornilakis, H., Papanikolaou, K., & Magoulas, G. (2002). Fuzzy inference for student diagnosis in adaptive educational hypermedia. In I. P. Vlahavas & C. D. Spyropoulos

(Eds.), *Methods and applications of artificial intelligence. Lecture notes in artificial intelligence, 2308* (pp. 191–202). Berlin: Springer-Verlag.

Kinshuk, Nikov, A., & Patel, A. (2001). Adaptive tutoring in business education using fuzzy backpropagation approach. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality*(pp. 465–468). London, Mahwah, New Jersey: Lawrence Erlbaum Associates Inc. Publishers.

Kuo, R. J., & Chen, J. A. (2004). A decision support system for order selection in electronic commerce based on fuzzy neural network supported by real coded genetic algorithm. *Expert Systems with Applications, 26*, 141–154.

Lane, H. C. (2006), Intelligent tutoring systems: Prospects for guided practice and efficient learning.

Meesad, P., & Yen, G. G. (2003). Combined numerical linguistic knowledge representation and its application to medical diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 33*, 206–222.

Mir Sadique, A., & Ashok, A. G. (2004). A neuro-fuzzy inference system for student modeling in web-based intelligent tutoring systems. *International conference on cognitive systems (ICCS), New Delphi.*

Reznik, L., & Dabke, K. P. (2004). Measurement models: Application of intelligent methods. *Measurement, 35*, 47–58.

Sendelj, R & Devedzic, V. (2004). Fuzzy systems based on component software. *Fuzzy Sets and Systems, 141*(3), 487–504.

Shavlik, J., & Eliassi, T. (2001). A system for building intelligent agents that learn to retrieve and extract information. *User Modeling and User Adapted Interaction, 13*(1–2), 35–88.

Stathakopoulou, R., Magoulas, G., Grigoriadou, M., & Samarakou, M. (2004). Neuro – Fuzzy knowledge processing in intelligent learning environments for improved student diagnosis. *Information Sciences, 170*, 273–307.

Turban, E., & Aronson, J. E. (2001). *Decision support systems and intelligent systems* (6th ed.). Hong Kong: Prentice International Hall.

Vrettaros, J., Pavlopoulos, J, Vouros, G., & Drigas, S. (2008), The development of a self-assessment system for the learners answers with the use of GPNN. *WSKS, 1*, 332–340.

Vrettos, S., & Stafylopatis, A. (2002). A fuzzy rule–based agent for web retrieval-filtering. Web intelligence: Research and development. In N. Zhong, Y. Yao, J. Liu, & S. Ohsuga (Eds.), *Lecture notes in artificial intelligence*, 2198 (pp. 448–453).

Part II
Aggregation and Classification of Knowledge

Intermezzo 2 – Artefacts of Thought: Properties and Kinds of Re-representations

Dirk Ifenthaler and Pablo Pirnay-Dummer

Knowledge is internal. Its representations are internal. External expressions about them are re-representations. Re-representations are representations of representations. Externalizations are the only available artefacts for empirical investigations. An externalization is always made by means of interpretation. But the externalization also needs interpretation for its analysis. These are two different kinds of interpretation. All kinds of features may be clustered for a description and aggregation of the artefact. Some of the interpretation is done by the learner and some of it is carried out by humans and technology. In most cases a mixture of all three interpreters will be part of the assessment. This mixture and the complexity of the construct both make it specifically difficult to trace the steps and bits of knowledge. Not all types of externalizations have the same types of properties and strengths, e.g., written language is always sequenced and has multiple dimensions at the same time (it is still impossible to trace them all), concept maps are not semantic webs most of the time due to underspecification problems and a lack of homogeneity, association networks do not have directions and propositions, causality networks can not deal with dynamics, and representations of dynamic systems are almost impossible to aggregate – nor are they supposed to be aggregable in the first place. The list is not even complete. There is no easy and no complete way to integrate any of them; and the strength of good research therefore lies, maybe more than in other research domains, in a fitting integration: Multiple perspectives on the same construct are usually needed. Only if the research questions are very specific may a single approach suffice. But this is rarely the case. Researchers and practitioners will have to carefully justify their selection alongside their research questions and goals, especially if important long-term decisions are based upon the assessments. The same care should be taken for decisions in the field. The only way to make better decisions about the kind of externalization and the type of instrument to be used on it is to know the strengths and weaknesses of the instruments. It is worth the effort to acquaint oneself with at least a representative selection of the available tools. In the following second part of the book, the authors present different approaches

and discuss them carefully and critically in relation to the underlying theories and applicable research standards. The overview and available detail of the information provided make it easier to select the proper tools and learn how and when to use them. The tools cover different approaches of aggregation and classification strategies on different externalized artefacts.

Chapter 6
Automated Knowledge Visualization and Assessment

Pablo Pirnay-Dummer and Dirk Ifenthaler

6.1 Introduction

The rapid advancement of information and communication technologies (ICT) has important implications for learning and instruction. Accordingly, remarkable repertoires of hypermedia systems, cognitive tools, learning management systems, and computer-based applications have been developed for almost every subject domain during the past decades. However, these important changes in teaching and learning through emerging technologies require new perspectives for the design and development of learning environments (see Hannafin, 1992; Ifenthaler, in press-a; Kirschner, 2004). Closely linked to the demand of new approaches for designing and developing up-to-date learning environments is the necessity of enhancing the design and delivery of assessment systems and automated computer-based diagnostics (Almond, Steinberg, & Mislevy, 2002; Ifenthaler, 2008a). These systems need to accomplish specific requirements, such as (1) adaptability to different subject domains, (2) flexibility for experimental and instructional settings, (3) management of huge amounts of data, (4) rapid analysis of specific data, (5) immediate feedback for learners and educators, and (6) generation of automated reports of results. Accordingly, an automated computational diagnostic system integrates various options which will be customized by the researcher or educator.

However, diagnosis is the systematic and theory-based collection and preparation of information with the aim of justifying, controlling, and optimizing conclusions and procedures (see Ifenthaler, 2008a). Thus, even if a computational diagnostic system is available, the researcher or educator needs to ensure that it is suitable for the research question or practical utilization (see Pirnay-Dummer, 2008). More importantly, the system should guarantee high reliability and validity (Seel, 1999). Additionally, students will benefit more from a computational diagnostic system if the results of the diagnosis are provided directly and instant feedback contributes to better comprehension (Ifenthaler, in press-b).

P. Pirnay-Dummer (✉)
Albert-Ludwigs-University of Freiburg, Freiburg, Germany
e-mail: pablo@pirnay-dummer.de

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_6,

This chapter provides insight into possible applications for computational diagnostics. The next section focuses on the challenges of automated computer-based diagnosis. We argue that despite several advantages computational diagnostics have, developers of such systems will be confronted with new structural requirements and will need to concentrate on unprecedented issues such as adaptive test forms and data security.

The following section identifies the potential of current computer technology for enhancing automated diagnostics. Multifaceted databases and network technologies serve as particularly powerful and flexible tools for multiple diagnostic purposes. Additionally, in this section we describe the technological framework of a server operating system. Next, we discuss automated tools based on the technological framework. Then, we introduce a further development of the tools: AKOVIA (Automated Knowledge Visualization and Assessment). The chapter concludes with applications and future prospects for automated computational diagnostics in different fields of research, learning, and instruction.

6.2 Applying Current Computer Technology

The latest hardware and software technology provides great potential not only for the design and development of learning environments but also for the enhancement of automated diagnostics. Advanced databases and network technologies contribute an especially wide variety of applications for an efficient diagnosis of individual and group data (Koper & Tattersall, 2004). However, numerous capabilities remain unused because standard diagnostic tools do not facilitate these technological features (Janetzko, 1999).

Regardless of the data collection technique, diagnostic purpose, and amount of collected data, a *designed server system* is the primary prerequisite. We propose that a low-cost designed server system for automated computational diagnostics can be developed in Linux using Apache, MySQL (MY Structured Query Language), PHP (PHP Hypertext Preprocessor), and Perl (Practical Extraction and Report Language). However, it is also possible to use other commercial products to build such a system. An example of a three-tiered hierarchical system architecture is shown in Fig. 6.1. The *web browser* enables researchers and subjects to log into the system. Such a system enables the researcher to (1) create tests and experiments, (2) manage tests, experiments, and subjects, and (3) analyze data from completed tests and experiments. The subjects are able to access different tests and experiments and receive immediate feedback on their performance. The *researcher* and *subject interface* builds dynamic web content for the requested application. A *transferring script* connects the subject and researcher interactions with the *database*. The specific requirements for a three-tiered hierarchical client/server model could be implemented as follows:

- *Web server*, e.g., Apache 2.2.8 (The Apache Software Foundation, 2008)
- *Database server*, e.g., MySQL 5.0.51 (MySQL AB, 2008)

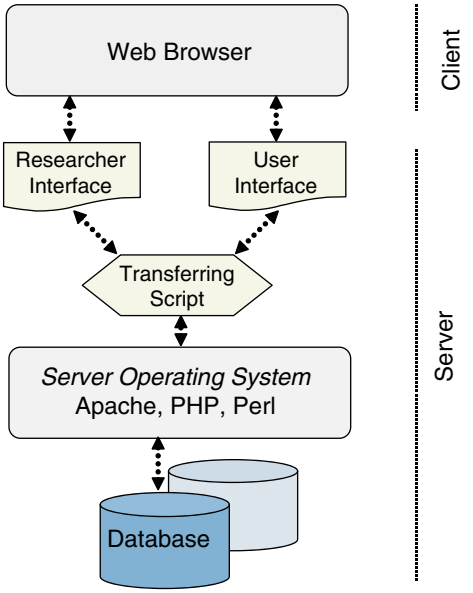


Fig. 6.1 Architecture of the *designed server system*

- *PHP* scripting language, e.g., PHP 4.4.8 (The PHP Group, 2008)
- *Perl* scripting language, e.g., Perl 5.8 (The Perl Foundation, 2008)

The *web server software* is used to serve dynamic and static web pages over HTTP (HyperText Transfer Protocol) on the World Wide Web. Usually, additional features such as server-side programming languages are implemented into the *web server software* to extend the core functionality. The *database server* provides multi-user access to specified databases. The databases are used to store and organize information. Stored information can be requested from databases by sending a query using a specific language, e.g., PHP. *PHP* scripting language is designed for creating dynamic web content. PHP can be embedded into HTML (HyperText Markup Language) and deployed on various web servers and operating systems. Many server-side commands can be realized with PHP, including queries from databases. *Perl* scripting language can be used for a wide range of applications, including network programming, data management, the creation of web content, GUI (Graphical User Interface), and system administration. The combination of these servers and scripting languages enables us to implement a framework for an automated computational diagnostics.

Most web hosts support all required servers, databases, and scripting languages by default. However, we recommend building a designed server system exclusively for automated computational diagnostics. This gives the researcher unrestricted access to manipulate all running server applications, databases, and additional preferences.

6.3 Automated Tools

Following the framework design from above, we will introduce automated tools for knowledge assessment and analysis which have helped us in many previous studies. Automation is not an end in itself. However, in many settings manual and therefore labor-intensive methods have limits, e.g., when the groups under investigation are large or practical applications do not have the resources which prototypes may have. Also, from a methodological viewpoint the automation helps in raising the objectivity of studies. Another important focus of our work is to make the applicability of the tools as broad as possible while retaining a very specific set of well-tested algorithms and methods which have shown to be reliable both in technical and empirical respects.

What the tools have in common is that they are all based strictly on the formal attainments of mental model theory. Historically, they derived from slightly different approaches since they were initially developed for specific research questions – and even for specific studies – and developed into more generally usable tools over time. The tool sets may be divided in two different sections:

1. The assessment of knowledge structures
2. The analysis and comparison of knowledge structures

Within our tools, natural language-oriented assessment plays an important part. We felt that in this domain new developments could be especially beneficial to the field – especially designs which do not include any manual coding. This does not at all mean that we limit our research to written language only approaches. In fact, we also use concept mapping and other graphical methods. It was simply not necessary to develop any more tools in this part. There are many tools available which are well established.

We will start with an introduction to the initial instruments which are still in use. Afterward, we will present recent developments which aim at a comprehensive integration of tools.

6.3.1 Mitocar

MITOCAR is a software toolset developed and introduced by Pirnay-Dummer (2006). It is based strictly on mental model theory (Seel, 1991) and has proven to deliver valid, homogeneous, and reliable results. MITOCAR is an acronym for "Model Inspection Trace of Concepts and Relations." It measures the properties of language re-representations of a realization prepared by a group. The re-representation is called the group consensus model. MITOCAR also measures whether there is sufficient agreement within the group (homogeneity).

6.3.1.1 Two Phases of Data Collection

To produce the consensus model of the graph, all the subjects need to do is go through a two-phase web-based assessment procedure which takes approximately 1.5 h for a whole group (Fig. 6.2).

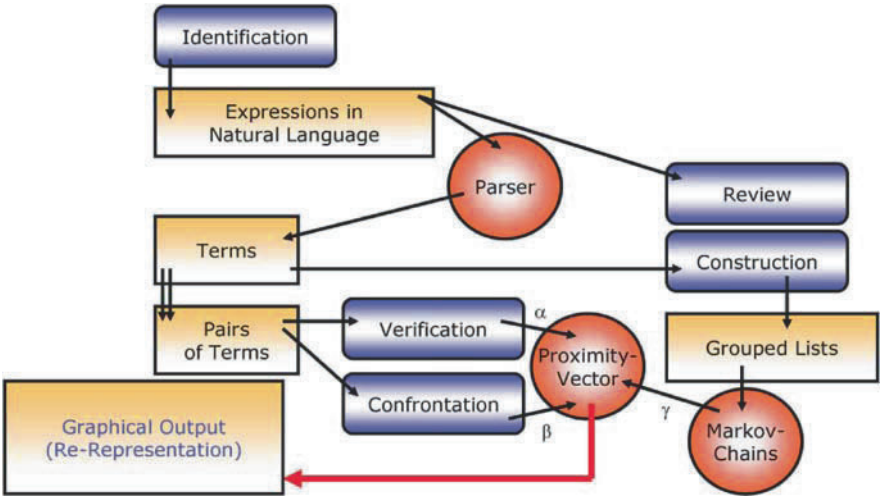


Fig. 6.2 MITOCAR model elicitation process (Pirnay-Dummer, 2006)

Identification, Review, Construction, Verification, and Confrontation are the modules which are presented separately to and used by the subjects. While Identification and Verification are mandatory for the functioning of MITOCAR, all of the other modules can be used to improve the quality of the knowledge assessment. All of the other steps are calculated automatically by the software and handled and stored on a database. The *identification* mode is the first phase of MITOCAR and is a simple collection of statements on a given subject domain. Between the first and the second phase, a *concept parser* filters nouns (with and without attached adjectives) and compiles a list of the most frequent concepts from the "mini-corpus." The second phase consists of the review, construction, verification, and confrontation. In the *review*, every group member rates all expressions of the group for plausibility and for their relatedness to the subject domain. In the *construction*, the subjects categorize concepts into groups which can be processed into model information using Markov chains: The instrument for the construction mode is based on the knowledge tracking technology as introduced by Janetzko (1996). It uses the $N = 30$ most frequent terms from the concept parsing, which are presented to the subjects as a list in randomized order. The subjects are asked to go through the list and click on concepts which they associate with a meaningful group. They do not have to name the groups. They are asked to keep the number of groups as small as possible while still being selective as regards content. Whenever a concept is clicked on it switches to the new list. A group marker can be set to identify the start of a new group. No step can be undone because we need the direct associations rather than a systematic collection. Systematic and more reactive methods are also part of MITOCAR and will be introduced in the next paragraphs. All frequencies of pairs are calculated from the grouped list: If concept B follows concept A in a list (and no group separation marker is in between), then the pair (A, B) gets +1 added to its frequency value. The matrix of term frequencies can be transformed

into basic transition probabilities which allow us to build Markov chains from the matrix (cf. Chung, 1968). The probabilities ζ can be seen as weights for the graph: $0 \leq \zeta \leq \nu$ for the relations (edges) e (ν being the highest probability within the matrix). We have $\nu < 1$ if the frequencies of more than one relation are bigger than 0: $f(e_i) > 0 \wedge f(e_i + j) > 0$ because the sum of probabilities within the matrix must be $\Sigma(\zeta) = 1$. To make sure that the matrices can be compared to others, each matrix can be standardized. The probabilities are standardized by assigning 0 to the lowest probability in the matrix and 1 to the highest one: $\min(\zeta') = 0$ and $\max(\zeta') = 1$. All other probabilities are adjusted linearly in between. *Verification* and *confrontation* are both modes for a pairwise comparison of concepts.

Pairs of these concepts are rated by the subjects in the second phase of MITOCAR for their closeness and contrast. Additionally, the subjects rate how confident they are about their rating. The three basic measures and meaningful combinations of them can be used later on for the graphical reconstruction of the model. All items are rated on a 5-point Likert scale on screen by the subjects.

1. *Closeness*: The item of closeness describes how closely related two concepts are rated as being by the subjects.
2. *Contrast*: For the item of contrast the subjects rate how different two concepts are or to what extent they exclude each other (e.g., fire and water).
3. *Combined*: This measure combines the items of closeness s and contrast k. It is calculated by $|(s-1)-(k-1)|+1 = |s-k|+1$. High contrast with low closeness and low contrast with high closeness both generate high combined values. The closer contrast and closeness become the lower the combined value will be. The scale remains the same as for closeness and contrast.
4. *Confidence*: The confidence rating ς measures how sure the subjects are of their ratings of contrast and closeness. To save space in titles and headers, all measures which are weighted by the confidence rating are designated by a (+) sign.

The MITOCAR software takes six pairwise-related model representation measures into account:

1. *Closeness*: The model is constructed only on the basis of the closeness rating s.
2. *Contrast*: The model is constructed only on the basis of the contrast rating k.
3. *Closeness+*: The model is constructed on the basis of closeness and weighted by confidence: $k \cdot \varsigma$. If the subjects rate the relation of concepts with more confidence, they will also be more likely to become a part of the model.
4. *Contrast+*: The model is constructed on the basis of contrast and weighted by confidence: $s \cdot \varsigma$.. If the subjects rate the relation of concepts with more confidence, they will also be more likely to become a part of the model.
5. *Combined*: The model is constructed on the basis of the combined measure $|(s-1)-(k-1)|$.
6. *Combined+*: The model is constructed on the basis of the combined measure and weighted by confidence: $|(s-1)-(k-1)| \cdot \varsigma$. If the subjects rate the relation of concepts with more confidence, they will also be more likely to become a part of the model.

Depending on the quality of the data (which is tested before re-representation), different measures may be used. e.g., if the combined item has too much deviance or is inhomogeneous within a group, it will be excluded from re-representation. This is automatically tested and reported by the MITOCAR software. In this study the data quality sufficed for the combined measure (all measures had a good quality). *Verification* and *confrontation* modes differ only in the pairs of terms which are rated. In the verification mode subjects rate the terms which come from their own group (utilizing their own power of language), while in the confrontation mode they rate pairs from another group (typically from a group which they are being compared to). This information is used to build (re-represent) the knowledge structure in the form of a concept map. So far, most studies conducted with MITOCAR have had sufficient data quality to use the combined+ measure.

6.3.1.2 Graphical Re-representation of the Model

The re-representations are constructed from these data on an undirected graph. For model re-representation a graph is drawn from the $N_M = 30$ strongest relations within the whole proximity matrix (in this study: combined+) using GraphViz as described by Ellson, Gansner, Koutsofios, North, and Woodhull (2003). N_M can be set within the software and can be adjusted if an upcoming change in standards or comparison to other approaches and data requires different model sizes in the future. For verification and confrontation modes this leads to a total of $N_P = 435$ rated pairs. The main graph and several different subgraphs are automatically assembled into a large report. The graph $G(V, E)$ is constructed from the concepts (which are the vertices) $v \in V$ and the strongest relations from all rated pairs (which are the edges) $e \in E$. Measures from construction, verification, and confrontation can be weighted separately to integrate them into the graph, depending on whether the research question requires finding out what different groups know, comparing the knowledge structure and its interaction within specific sets of groups, or finding out more about the non-reactive associative use of knowledge (e.g., spontaneous decision making). All pairs of terms are part of a list which constitutes the proximity vector. High-ranking pairs are considered to be close (strong relations), while low-ranking pairs are interpreted as being far apart. The report includes all parameters, MITOCAR comparison measures, analysis based on graph theory including the measures of SMD Technology (see Ifenthaler, 2008b), tests for multidimensional scaling, tests for homogeneity, and model complexity, and sorted and reviewed lists of statements from the "mini-corpus." All statistical tests in these reports provide automatically generated assistance for the interpretation process, making the reports suitable and readable for nonexperts on research methods, e.g., instructors, teachers, instructional designers, etc. Only the most significant subset of the reports generated by MITOCAR is presented in this chapter. Different research projects using some or all of the MITOCAR modules may consider different parts of the report depending on the research questions they are addressing.

6.3.1.3 Additional Descriptive Elicitation Modules

The model constructed by the data already contains all information from the elicitation and is the main object for later comparison, tracking over time, etc. So far, it is a standardized procedure for re-representing the model content of a group of learners. Especially for practical applications of the methodologies described here, additional formats can be very helpful for accessing specific information more easily than from the main representation. We will introduce a selected set of additional descriptive elicitation modules available within MITOCAR in the following paragraphs.

The *stepwise model inspection* allows users to "zoom" into a knowledge structure by subsequently adding the most associated edge of the structure. The software creates one picture file for each zoom step. At the start, the strongest edge in the graph has two concepts (vertices) attached. Each subsequent step introduces another edge to the graph (Fig. 6.3).

In other model structures this second step may only contain three concepts and may already integrate them. In this way, the stepwise model inspection builds up (1, 2, . . ., *n*) graphs which are all subgraphs of the main model. It ends with the complete model. Thus, it is possible to retrace the composition of the model one step at a time. This can be used to support feedback to the group. Showing the composition of the models is like "zooming into" the model structure, and participants tend to like this stepwise introduction because it does not show all of the (maybe complex) information of the whole model at once. Beyond this simplification, they can follow the steps new concepts take in the process of being integrated into the already introduced structure of concepts within their own knowledge organization (to zoom into the structure). This simple chain of graphical representations can be used to support classical methods of providing feedback to the group, especially when the feedback is about progress (which has already been made or which is still to be made).

In addition to the stepwise model inspection, analysis and feedback can also be focused on specific concepts within the model. The *star model inspection* of concepts can add more insight when we investigate central concepts of a domain. Its purpose is to investigate single concepts rather than whole model structures. To do so, one builds a star model for each concept of the main model (Fig. 6.4).

A star model shows the N_s strongest connections to its center concept, including those from the main model. It also contains additional (deeper) relations from the data matrix due to the fact that it always shows a number of N_s relations. Therefore, it can even contain concepts which the main model does not show at all. It is clearly not a subgraph of the main model although it always contains links (vertices) from it.

Spanning trees can be created with ease from the MITOCAR main models. They are especially interesting for practitioners who want (or need) to use group models for planning on instruction, e.g., if we want to use the knowledge structure of advanced learners to find sets of better learning sequences. Spanning trees eliminate the cycles within a graph, making it more sequential. We can almost intuitively follow the spanning tree to come up with an inductive path from the peripheral concepts to the key concepts (bottom up). We can also follow a more deductive path from the center concepts to the branches. Of course, we should keep in mind that

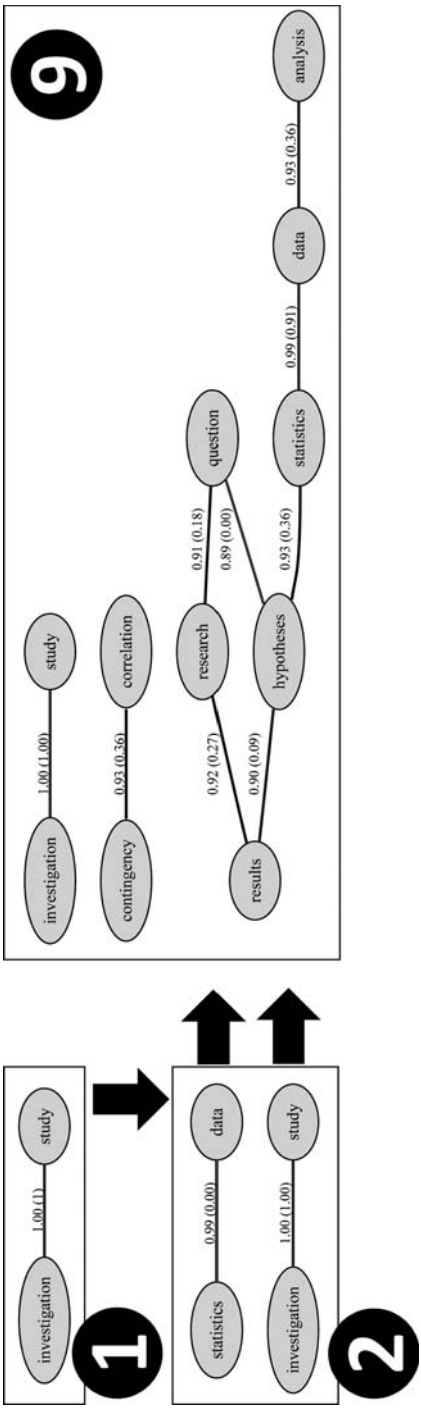


Fig. 6.3 Stepwise model inspection. Steps 1, 2, and 9 (as examples)

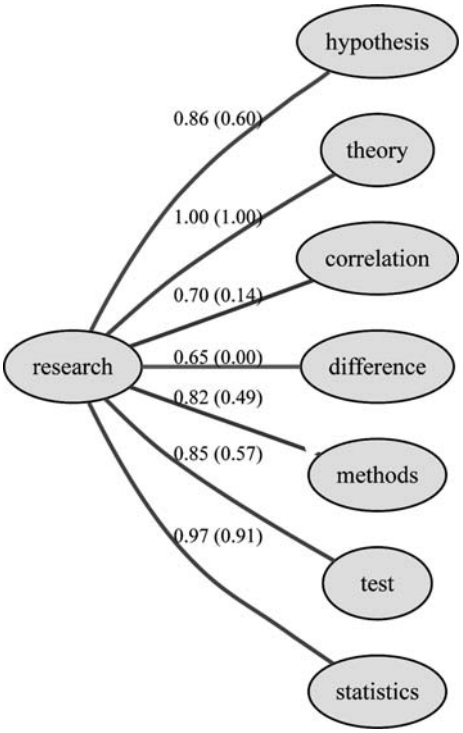


Fig. 6.4 Star model inspection around a single concept

good curriculum design is also based on many other insights (e.g., organizations, socialization, personalities, teaching strategies), but the structure of novice knowledge versus expert knowledge will certainly provide valuable information for the design.

As an addition to the graphical analysis of MITOCAR, it may be interesting, e.g., in qualitative studies, to compare the condensed graphical results to the original language expressions. Therefore, MITOCAR uses the data from the review mode to present *ranked lists of expressions*. All three ratings from the review may be used separately or combined to generate the ranked lists. This information can be especially helpful if the data is used for decisions for instructional design (e.g., comparing the knowledge of different groups for needs assessment).

6.3.1.4 Automated Report Engine

MITOCAR generates automated reports which not only display the knowledge structure in a concept map-like format but also calculate and interpret several tests, e.g., multidimensional scaling and homogeneity (within a group's knowledge), and provide additional descriptive measures and graphs which help to find answers within the knowledge structure (Pirnay-Dummer, 2006). MITOCAR can generate

two kinds of reports. Both are composed automatically from the data and output as PDF files. Reports may be generated once the following modes have been complete: identification, review, verification, and/or construction. The completion of more modes may help to improve the quality of the data. The improvement depends on various factors, such as the homogeneity of the expertise and/or the group of experts or on the number of available experts. In addition to the graphs, tables, and statistical tests of the data used for both re-representation and analysis, the reports also interpret the effects. They also contain useful hints on what follows from the data and how they can be interpreted. All the graphs and tables are also stored separately in case they are needed later on, e.g., for further analysis and publications. The first report describes a group's model in detail. It contains a preface, demographics (of the group), the model itself, a statistical evaluation of the data used for the proximity vector to create the re-representation, an evaluation of the relation strengths, a multidimensional scaling of the model including stress values, and the distribution graph. The report continues with graph-theoretical analyses of the model and homogeneity measures – for determining whether the members of the group are sufficiently aligned to consider their output a real group model. It checks for and interprets the correlations between the variables used for model construction. For more descriptive insight, *star model inspections* and the *stepwise inspection* are also generated, as are the initial expressions from the identification mode ordered by their evaluations (ratings) within the group. The second report compares different groups. It contains a more quantitative analysis than the single group report. It contains analysis of every model and adds more statistical tests, e.g., for the model complexity. It generates, visualizes, and tests every possible proximity vector from the data. After the individual analysis of each model, the comparison is carried out on the level of concept matching, structural matching, and a linear combination of both. Also, the structural complexity and the density of the models (γ) are analyzed. We will describe the comparison measures later on in detail. After each group has gone through the two phases of MITOCAR, the reports can be generated. They provide the basis for both the description of the individual group's knowledge and the comparison.

Both reports are built for research use. There is a prototype of the first type of report which uses a comprehensive subset of the analysis to be read by practitioners. In order to reach this goal, we applied general findings from readability research to change the output of the research tool – and also left research-oriented information out, e.g., the inferential statistics part. In an initial evaluation conducted in 2005 with 14 expert teachers at German high schools, practitioners already found this prototype to be understandable. The mean for readability was 4.467 (SD = 0.91) on a scale from 1 to 6. The sample size was only $N = 14$ expert teachers, so only descriptive conclusions can be made at this point. Tools and other material will still have to be built around the prototype to aid the teachers in applying the findings to practice. The actual practical use of the prototype was evaluated with a mean of 3.6 (SD = 1.2). Eleven of the experts reported independently that they found the report particularly useful in group settings.

6.3.2 T-MITOCAR

So far, MITOCAR is good for assessing the expertise of small groups. Individual assessment with the toolset is practically not feasible – although it would be theoretically possible. But a single person would have to go through too many cycles (e.g., rating pairs of terms) in order to keep variances low enough for the proximity vector. This gave us the initial motivation to start working on a different approach. The goal was to improve the availability of written text across all subject domains (in schools, in companies, in learning management systems, in forums, in chats) and of course also from qualitative research. T-MITOCAR (Text-MITOCAR) stands on the same theoretical fundament as MITOCAR. The methodology behind the assessment is very different, however. MITOCAR parses the concepts out of a text and leaves the rating of their associatedness to the subjects in the second phase. T-MITOCAR, on the other hand, tries to track the association of concepts from a text directly (within predefined boundaries). To do this, it uses a heuristic which assumes that texts contain model structures. Closer relations tend to be presented closer within a text. Please note that this does not necessarily work within single sentences, since syntax is more expressive and complex. But everyday texts which contain 350 or more words can be used to generate associative re-representations. The re-representation process is carried out in different stages. All of the stages are automated. Thus, the only data needed is a text written by a subject under investigation (e.g., an expert, a learner, a teacher). Later on, we will also present the user interface and the output of the software.

6.3.2.1 Preparing the Text

When text is pasted from unknown sources (unknown to the software), it will most of the time contain characters which could disturb the re-representation process. Thus, a specific character set is expected. All other characters are deleted. Tags are also deleted, as are other expected metadata within each text: Formatting code would be in the way if the language processing were carried out.

6.3.2.2 Tokenizing

After preparation, the text gets split into sentences and tokens. Tokens are words, punctuation marks, quotation marks, and so on. Tokenizing is somewhat language dependent, which means that we need different tokenizing methods for every language we want to use.

6.3.2.3 Tagging

Only nouns and names should be part of the final output graph. Tagging helps to find out which words are nouns or names. There are different approaches and heuristics for tagging sentences and tokens. A combination of rule-based and corpus-based tagging is most feasible when we do not know the subject domain of the content in advance. And, since T-MITOCAR should work domain independently, this is an

important factor. Tagging and the rules for it is a quite complex field of linguistic methods. An explanation of our tagging technique would go beyond what is presentable in this chapter. Please see Brill (1995) for a good discussion on mixed rule-based and corpus-based tagging.

6.3.2.4 Stemming

Different flexions of a word should always be treated as one (e.g., the singular and plural forms "door" and "doors" should appear only once in the re-representation). Stemming reduces all words to their word stems. Therefore, all words within the initial text and all words within the tagged list of nouns and names are stemmed before the re-representation.

6.3.2.5 Fetching the Most Frequent Concepts from the Text

After tagging and stemming, the most frequent noun stems are listed from the text. How many terms are fetched from the text depends on the length of the text in words and sentences. Thus, larger texts also generate larger models. There is, however, a ceiling value. In the running versions of T-MITOCAR no more than 30 single terms are fetched from a text. This value can of course be set for the software.

6.3.2.6 Sum of Distances: Determining Pairwise Associatedness

The algorithms of T-MITOCAR calculate the associatedness, which constitutes the proximity vector. This measure compares to the weight of the links in MITOCAR and is also visualized in the same way. However, it is generated quite differently. The following steps are carried out for re-representation:

1. The default length is calculated. The words are counted for each sentence. The default length is the longest sentence within the text plus 1.
2. All fetched terms are paired, so that all possible pairs of terms are in a list.
3. For each pair all sentences are investigated. If the pair appears within a sentence, the distance for the pair is the minimum number of words between the terms of the pair within the sentence: If at least one term occurs more than one time in the sentence, then the lowest possible distance is taken.
4. If a pair does not appear in a sentence (true also if only one term of the pair is in the text), then the distance will be the default length.
5. The sum of distances is determined for each pair.
6. The *N* pairs with the lowest sum of distances find their way into the re-representation. Like the list of terms, *N* depends on the number of words and sentences within the text. The exact values can be controlled by the software settings.
7. Another process automatically cuts the maximum distance from re-representation, even if pairs would normally be presented on the basis of the number of sentences and words. This prevents the algorithm from just

deriving random pairs which do not really have any association evidence within the text.

6.3.2.7 Determining the Weights

The weights are calculated from the pair distances. They are to some extent comparable to the *combined* measure of the MITOCAR toolset. All weights ($0 \leq w \leq 1$) are linearly mapped so that 1 is the pair with the lowest sum of distances and 0 is the pair with the maximum sum of distances. Please note that the pair distance values have no direct meaning. They depend on the longest sentence of the text. Therefore, only the relative measure of the weights is interpretable directly.

6.3.2.8 De-stemming

Linguistic word stems sometimes look strange to untrained viewers. Although one can still guess which words they come from, deriving the output directly from the word stems does not help one in reading the re-representations. Thus, lists of words and their stems are created during stemming for the text. After determining the associatedness and the weight, T-MITOCAR looks up this table to search which word led most frequently to the stem: If it was the plural then the plural is presented. If it was the singular, then the singular moves into its place. Thus, the final output model contains a real word in that it uses the flexion which was used most frequently in the text.

6.3.2.9 Writing the Model to the List Form

The list form is a table (see Table 6.1) which accounts for an undirected graph containing all *N* pairs. It is sorted by weight (descending).

Table 6.1 List form of the T-MITOCAR output

Term 1	Term 2	Sum of distances	Weight
Fire	Water	3428	1
Fire	Foam	5756	0.73
...	...	...	...

6.3.2.10 Example from Wikipedia Texts (Economy)

We took an example from Wikipedia to illustrate how T-MITOCAR works. The text on "philosophy of science" is easy to look up on Wikipedia (see references for the links). We took the first section of the text and ran it through the software. First, T-MITOCAR shows the most frequent terms within the text (Table 6.2):

Second, the pairs, sums of distances, and weights are calculated. The distances cannot be interpreted directly. Thus, the weight is usually the value to look at (Table 6.3).

Table 6.2 Most frequent terms within the first section of "philosophy of science" on Wikipedia

Term	Frequency
Observation	52
Theory	51
Science	36
Measurements	16
Paradigm	15
Explanation	12
Scientists	12
Problem	11
...	...

Table 6.3 Top part of the list of pairs within the first section of "philosophy of science" text on Wikipedia

Term 1	Term 2	Distances	Weight
Observation	Theory	492	1
Observation	Science	516	0.43
Ockhams	Razor	516	0.43
Observation	Paradigm	519	0.36
Observation	Scientists	519	0.36
Theory	Science	519	0.36
...	...	...	...

As already described, the list form can be transformed directly into the re-representation graph:

Figure 6.5 shows a representation of the Wikipedia entry on philosophy of science. The font size has been manually increased from the original SVG format for better readability. The main concepts are on the vertices of the graph, and the association strengths can be found at the edges (links). Like in MITOCAR, only the strongest associations are represented. In its standard settings, T-MITOCAR displays up to 25 links from the list form. Therefore, there are two measures for the association strength at the links. The value outside the brackets shows the weight from the list form. The second value inside the brackets displays the weight relative to what is actually visualized. The strongest association will be 1 and the weakest observation will be 0. Depending on whether the whole construct or the graph is being discussed, either value may be more helpful (easier) to interpret. A close look at Fig. 6.5 will also reveal that there are terms whose flexions are not unified by T-MITOCAR: both "hypothesis" and "hypotheses" appear in the graph. In those rare cases, the stemmer does not recognize the word stems properly. To account for this, T-MITOCAR has a lexicon of those cases. However, in some of the special cases, we decide not to unify the flexions because they may have completely different meanings. Figure 6.5 contains such a case which we selected to show.

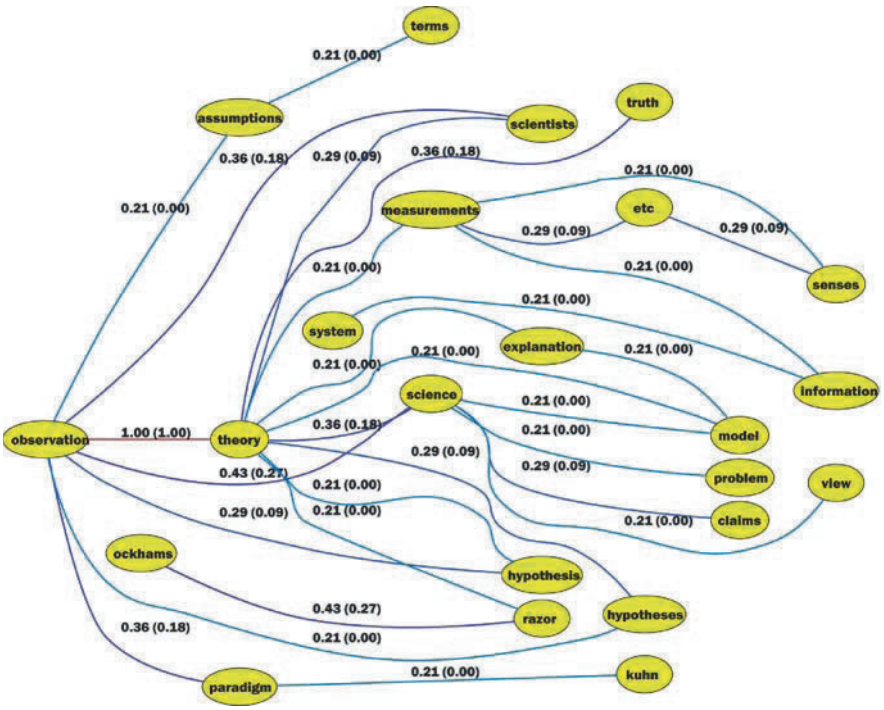


Fig. 6.5 Graphical representation of the first section from the "philosophy of science" text on Wikipedia

6.3.2.11 How to Use T-MITOCAR

Using T-MITOCAR is very easy. Just start with a copy of any text and paste it onto a form. Give the text a label so it can be found later on for analysis and comparison (Fig. 6.6).

Click on the "submit" button, and everything else is done by the system. After that, the following menu items are available on the label:

- *Terms:* Gives the user a list of the most frequent terms.
- *View:* Displays a thumbnail of the graphical model which is linked to the full size picture of the re-representation.
- *Tables:* Displays the model in the list form and generates an MS Excel® file of the list form for download and further analysis.
- *Compare:* Allows quantitative comparison of two or more models (pairwise) with seven separate measures for comparison (ranging from surface over structure to semantic analysis and comparison functions). The comparison measures will be described in more detail later on.
- *Discard:* Allows one to discard one or more models by identifying the labels.

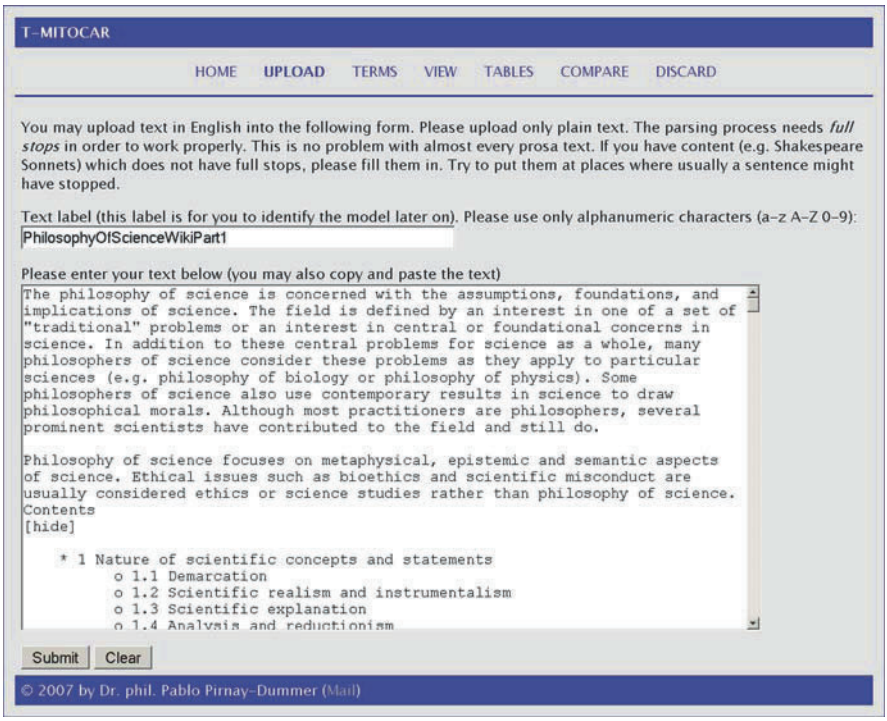


Fig. 6.6 Pasted text and a label are all T-MITOCAR needs for re-representation

6.3.2.12 Applications

T-MITOCAR helps to analyze and compare small or medium datasets and single-text models. Texts written by learners during the instructional process may be compared to any expert text, advanced learner's solution, or standard or model solutions for the task. Thus, it is also possible to track change over time – if the learners write texts at several measurement points. Also, any semantic and structural differences within or between groups may be measured on the level of the graph comparison indices. The comparison functions are also built into T-MITOCAR, making it unnecessary to transform data. We will describe the semantic and structural comparison measures later on. Teachers may also take the visual output of individual models (e.g., from assignments) to their class and discuss the knowledge structure of student solutions as part of their teaching methods. Group aggregations may also be processed with T-MITOCAR. If more than one text is pasted onto the upload field, then the software will analyze the texts as a single text. Since the re-representation process is based on associations, weights which are repeated throughout the text will become stronger. Distances are determined as a linear function: Distance is measured linearly against the longest sentence in the uploaded text. The transformation from the distance measure to the standardized weight measure is also a linear function. Therefore, if the uploaded texts originated from the

same assessment – which is generally a good idea when aggregation is intended, T-MITOCAR will stably aggregate a group model from more than one text. An aggregation will allow the comparison measures to be used to determine semantic and structural levels of alignment or coherence within a group (e.g., as compared to another group).

6.3.3 T-MITOCAR Artemis

T-MITOCAR Artemis uses the algorithms of T-MITOCAR for a different purpose. The software does not directly aim at knowledge assessment but rather uses the natural language-oriented diagnostic tools of T-MITOCAR to create knowledge maps out of larger text corpora, such as all the documents of a project, department, or even a whole company. Artemis uses multi-cluster text corpora as input. Thus, an automatically pre-clustered corpus which resembles the written knowledge of any group can be used without further formatting (e.g., tagging) and without manual work. Each cluster is graphically visualized by Artemis and becomes a "continent" on the knowledge map. Each continent is represented in a different color. The whole process takes from 2 min to several hours, including the time for clustering (usually 0.5–3 min) and for graphical representation, which takes between 1 min and several hours depending on the amount of text material. Compared to classical manual knowledge mapping techniques this is still fast, but the computing power is beyond what is feasible in real time on a web server. Therefore, in contrast to T-MITOCAR, which is an online tool, Artemis can only be run offline.

6.3.3.1 Input Formats and Interface

There are three interfaces available for T-MITOCAR Artemis at the moment. The first one is a ZIP file which contains each text cluster as a separate plain-text file. The text needs to be utf-8 encoded. This format is used if Artemis is implemented as part of a knowledge mapping methodology or if other software is integrated with it. The second interface is also a ZIP file, but it contains further ZIP files, one for each continent of the knowledge map. The further ZIP files contain at least one MS Word® format file (.doc) each. They may also contain more than one Word file. In the process, the Word files within each ZIP file are treated as one cluster (one individually colored continent). This format is suited for manual clustering from multiple sources when Word files are available. The third interface links Artemis to Wikipedia. Artemis needs at least one term as input and constructs the map using the text from Wikipedia as a corpus. Each term builds a continent on the map. This technology could easily be used to process documents from the web onto the knowledge map. Unfortunately, the APIs (Application Programming Interface) to search engines necessary for this are far from available. Some good API services have even been discontinued, e.g., the Simple Object Access Protocol, SOAP, from Google, which stopped providing new access codes in 2006.

6.3.3.2 Output Format of the Knowledge Map

As output, Artemis first generates a list form of the graph and then a bitmap in the Portable Network Graphics format (PNG) or the Scalable Vector Graphics format (SVG) using GraphViz like T-MITOCAR.

Figure 6.7 shows a miniature of a whole map. It also shows how central concepts are cross-linked more densely, thus integrating the continents into the whole map. Not all knowledge maps from Artemis are integrated like this. If the content does not connect, the knowledge map will also only have disconnected "islands." We use this output with learners and stakeholders within companies to discuss decision-making processes and procedures. The maps may be used for different applications, e.g., in meetings, workshops, and long-time learning settings. In short-term learning (up to 1.5 h), the maps do not provide a significant benefit as it takes the learners too long to get to know the maps.

Figure 6.8 shows the details of a knowledge map. Except for the colors and the missing weights at the links, this looks like the output from MITOCAR or T-MITOCAR. The weights still play the same role when the graph is generated. In order to make the map simpler to read, they are not represented. Artemis was developed only recently. There are some initial studies available, but it will still need to go through more practical applications to empirically demonstrate its benefits and limits.

6.3.4 SMD Technology

Based on the theory of mental models (Seel, 1991) and graph theory (Bonato, 1990; Chartrand, 1977; Harary, 1974; Tittmann, 2003), the computer-based and automated *SMD Technology* (Surface, Matching, Deep Structure; see Ifenthaler, 2008b) uses (a) graphical representations such as concept maps or (b) natural language expressions to analyze individual processes in persons solving complex problems at single time points or multiple intervals over time. The in-depth analysis process generates quantitative measures and standardized re-representations for qualitative analysis and feedback. The results of the SMD Technology are determined in four phases:

6.3.4.1 Phase 1: Input

Once knowledge structures have been elicited with an adequate methodology, e.g., DEEP (Spector & Koszalka, 2004), Cmap Tools (Cañas et al., 2004), or as written text, they can be described and measured with the help of the SMD Technology. Depending on the elicitation process (e.g., using the *Structure Formation Technique* [paper and pencil]; *concept mapping tools* [computer-based]; *natural language statements* [computer-based or paper and pencil]), the raw data should be stored pairwise (as propositions P_i) and include (a) the *model number* as an indicator of which model a proposition belongs to, (b) *node1* as the first node of the proposition,

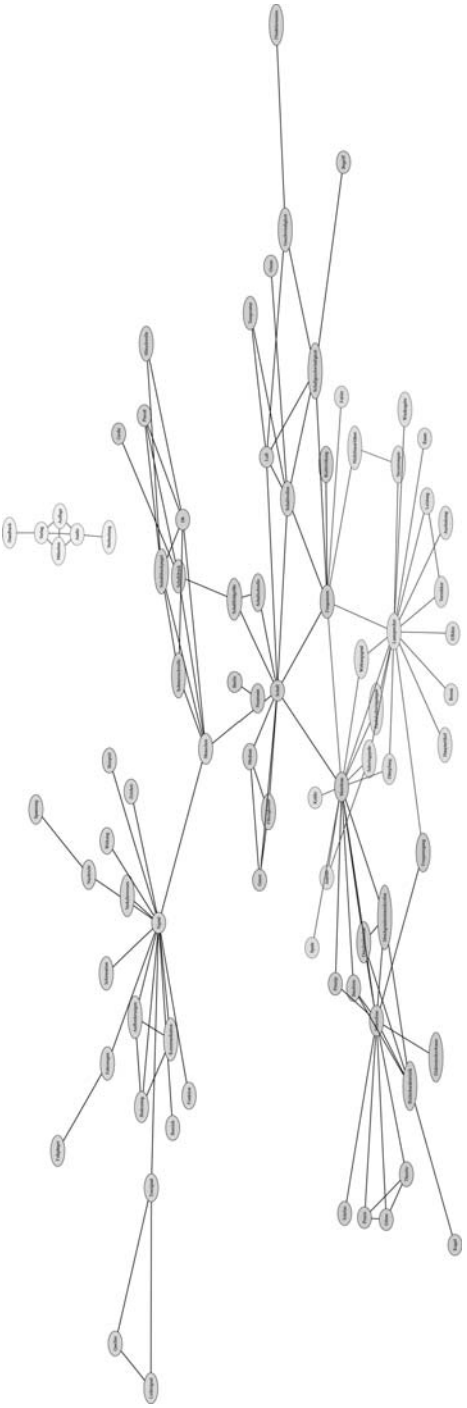


Fig. 6.7 Miniature of a whole knowledge map

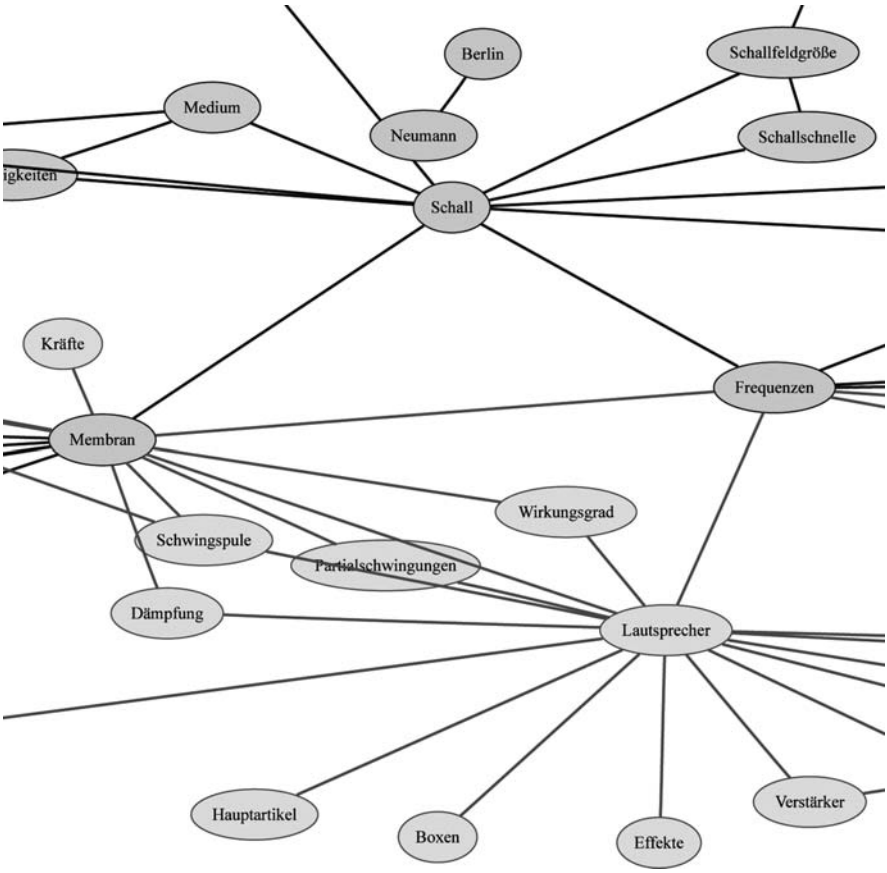


Fig. 6.8 Details of a knowledge map

(c) *node2*, which is connected to the first node, and (d) a *link* which describes the link between the two nodes (see Table 6.4).

The data structure described above should be stored as a comma-separated CSV file (comma separated values), which can be easily stored on the SQL database of the SMD Technology. This process (including only one or multiple re-representations) can be repeated as often as necessary.

Table 6.4 Input format for the SMD technology

Model number	Node1	Node2	Link	Subject number
001	Learning	Example	Through	912abz3
001	Learning	School	Takes place	912abz3
...				
007	Example	Theory	For	543sfe9

6.3.4.2 Phase 2: Analysis Specification

In the second phase, the researcher selects specific sets of knowledge representations to be analyzed from all stored data. SMD Technology allows the user to choose from different forms of knowledge representations: (1) an individual representation of a specific point in the learning process, (2) an expert representation of a single domain expert, (3) a combined expert representation of two or more domain experts, (4) a textbook or conceptual representation, and (5) a shared representation of two or more individuals or of two or more measurement points. The automated analysis process of the *SMD Technology* will be started by the researcher and will automatically calculate quantitative measures and generate standardized graphical re-representations for each individual knowledge representation (Ifenthaler, 2008b).

6.3.4.3 Phase 3: Quantitative Analysis Output

In the third phase, quantitative measures for the requested data are generated automatically. SMD calculates (1) the number of propositions in an individual representation (surface structure) and (2) the "diameter" of the spanning tree of the representation (matching structure). These measures represent the structural complexity of the data. The diameter and further indicators (e.g., number of cycles, number of submodels, ruggedness, etc.) are derived from graph theory (Harary, 1974; Ifenthaler, Masduki, & Seel, 2009). Additionally, domain dependent semantic measures are calculated with the help of the similarity measure between individual or team representations and reference representations (e.g., expert representations). SMD then calculates (3) the semantic similarity of single nodes – vertex matching (Pirnay-Dummer, 2006) and the semantic similarity of propositions (Deep Structure).

Once calculated, the indicators are stored on the SQL database. All indicators and additional information, e.g., subject number, measurement point, experimental group, etc., can be downloaded in various formats and used for further statistical calculations.

6.3.4.4 Phase 4: Standardized Graphical Output

In the last phase, our visualization feature automatically generates a standardized graphical output of all data selected in phase two. With the help of the open source graph visualization software *GraphViz* (Ellson et al., 2003), all visualizations are stored on the web server as *PNG* (Portable Network Graphics) images and named dynamically with additional information (e.g., subject number). The user can click on the thumbnails to see the actual size of the standardized graphical output (see Fig. 6.9).

The PNG image (1) represents the *subject representation*, (2) the *reference representation (e.g., expert solution)*, (3) the *similarity representation*, and (4) the *contrast representation*. The *subject representation* includes all nodes and

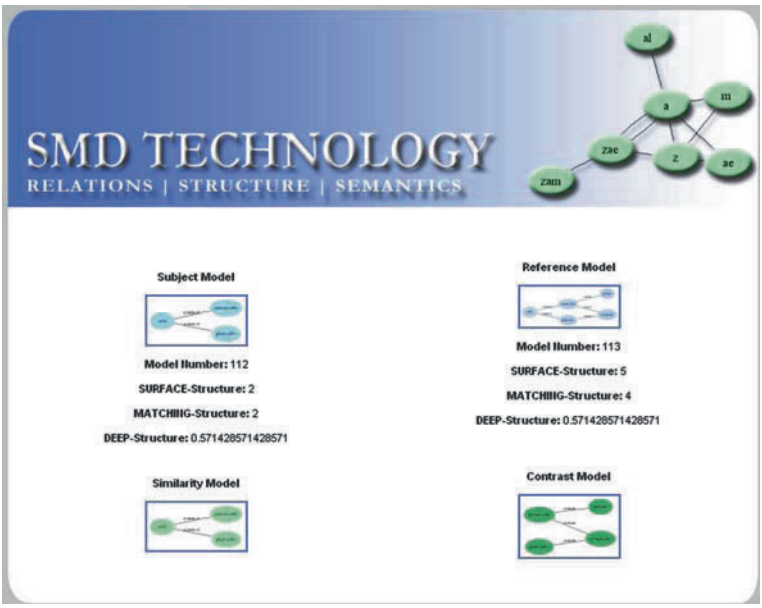


Fig. 6.9 Standardized graphical output of the SMD technology

links which were originally constructed by a subject. The *reference representation* includes all nodes and links of a subject, group, expert, or conceptual model. Within the *similarity representation*, only nodes and links which are semantically similar between the *subject* and *reference model* are included. The *contrast representation* includes all nodes and links of the *subject representation* which are not semantically similar to the *reference representation*. Depending on the size of each *subject* and/or *reference representation*, the dynamic web page is generated within one to several seconds, which also includes the whole analysis process of the back-end of the SMD Technology.

6.3.5 Model Comparison

There are seven indices for the knowledge-oriented comparison of graphical re-representations from the SMD Technology (Ifenthaler, 2006) and from MITOCAR (Pirnay-Dummer, 2006). Some of the measures count specific features of a given graph. For a given pair of frequencies f_1 *and* f_2, the similarity is generally derived by

$$s = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)}$$

which results in a measure of $0 \leq s \leq 1$, where $s = 0$ is complete exclusion and $s = 1$ is identity. The other measures collect sets of properties from the graph. In

this case, the Tversky similarity (Tversky, 1977) applies for the given sets A and B:

$$s = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A - B) + \beta \cdot f(B - A)}$$

α and β are weights for the difference quantities which separate A and B. They are usually equal ($\alpha = \beta = 0.5$) when the sources of data are equal. However, they can be used to balance different sources systematically (e.g., comparing a learner model which was constructed within 5 min to an expert model, which may be an illustration of the result of a conference or of a whole book).

Figure 6.10 shows an overview of the different measures. On the structural level there is surface matching (frequency), graphical matching (frequency), structural matching (Tversky) and gamma matching (frequency). On the semantic level comparisons may be carried out as concept matching (Tversky), propositional matching (Tversky), and balanced semantic matching. The latter may be derived from the first two.

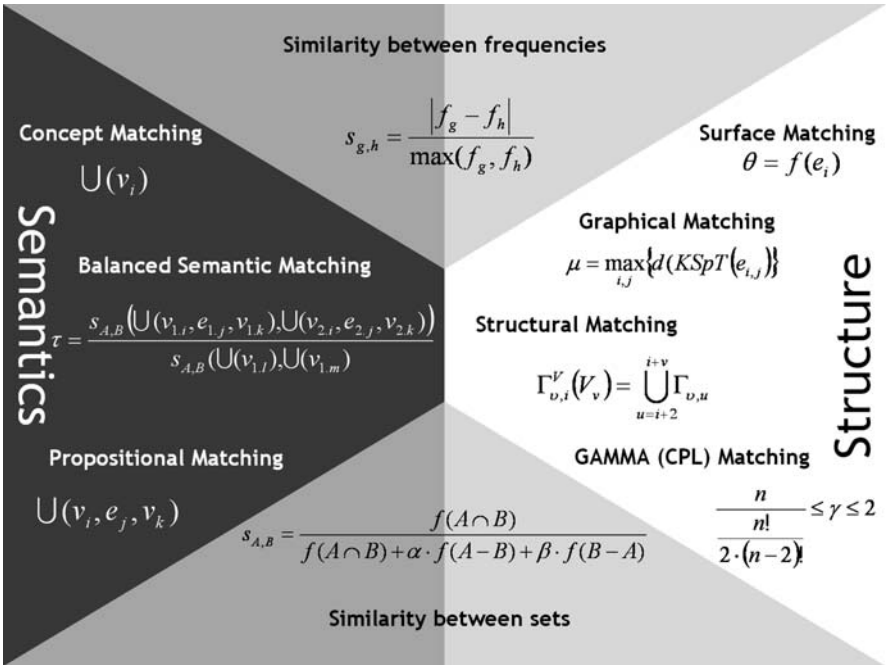


Fig. 6.10 Measures for structural and semantic comparisons

The individual comparison measures are described in detail in the following subsection.

6.3.6 Comparison Measures

The above discussed tools SMD Technology, MITOCAR, and T-MITOCAR introduced six different measures for comparing knowledge representations (see Fig. 6.10). These measures include surface, graphical, structural, and gamma matching for structural analysis of the knowledge representations. Concept and propositional measures include semantic measures.

6.3.6.1 Surface Matching

The surface structure is computed as the sum of all propositions (node-link-node) in a representation (Ifenthaler, 2008b). It is defined as a value between 0 (no proposition) and *n* (*n* propositions of the representation).

The surface matching compares the number of nodes within two knowledge representations. According to the theory of mental models (Seel, 1991), the number of nodes and links or propositions a person uses is a key indicator for the investigation of the progression of knowledge over time in the course of problem-solving processes.

6.3.6.2 Graphical Matching

The structural property of a knowledge representation is displayed in the graphical measure. It is computed as the diameter of the spanning tree of a knowledge representation and can lie between 0 (no links) and *n*. In accordance with graph theory, every representation contains a spanning tree. Spanning trees include all nodes of a representation and are acyclic (Harary, 1974; Tittmann, 2003). A diameter is defined as the quantity of links of the shortest path between the most distant nodes. For calculation of the graphical measure, the spanning tree is transformed into a distance matrix *D*.

$$D = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}$$

The graphical measure is calculated as the maximum value of all entries in the distance matrix *D*. The graphical matching compares the diameters of the spanning trees of the knowledge representations, which is an indicator for the range of conceptual knowledge. It corresponds with structural matching as it is also a measure for structural complexity only.

6.3.6.3 Structural Matching

Structural matching compares two graphs on a structural level only. Instead of using a graphical heuristic, it reconstructs the structure of a graph on specific lists. The algorithm was initially developed for MITOCAR in order to solve a specific problem. Pirnay-Dummer (2006) investigated structural hypotheses which – in a nutshell – stated that expertise can be structurally different depending on how it evolves. Due to different use of language between the groups of experts (e.g., merchants vs. economists, instructional designers vs. teachers), this could not be carried out on the semantic level. Thus, an algorithm was developed to solve this problem and investigate such hypotheses. Chapter 13 of this book is dedicated to the foundations, design, development, and evaluation of this algorithm. A similarity of $s = 1$ means that two models have identical structures (e.g., when each of them is a circle of five concepts). A similarity of $s = 0$ means that the two models do not share any structural components at all.

6.3.6.4 Gamma Matching

The density of vertices describes the quotient of edges per vertex within a graph. Since both graphs which connect every term with each other term (everything with everything) and graphs which only connect pairs of terms can be considered weak models, a medium density is expected for most good working models. Groups of experts have usually shown a mean of $s = 0.32$ (Pirnay-Dummer, 2006). The density between two models is matched by the numerical similarity. A similarity of $s = 1$ means that the two models have the exact same density. A similarity of $s = 0$ means that the densities differ completely. The latter condition can only be reached if one graph consists of only paired concepts, whereas every concept is connected to every other concept in the other graph.

6.3.6.5 Concept Matching

Concept matching compares the sets of concepts (vertices) within a graph to determine the use of terms. This measure is especially important for different groups which operate in the same domain (e.g., using the same textbook). It determines differences in language use between the models. The Tversky similarity measure is used. A similarity of $s = 1$ means that all concepts are alike while a similarity of $s = 0$ shows that no concepts match between both models.

6.3.6.6 Propositional Matching

Additionally to the concept matching measure, propositional matching compares only fully identical propositions between two knowledge representations. It is a good measure for quantifying complex semantic relations in a specific subject domain. The propositional matching measure (see Ifenthaler, 2008b) is calculated as the semantic similarity between an individual representation and a second individual or reference representation.

6.3.6.7 Balanced Semantic Matching

The balanced semantic matching measure uses both concepts and propositions to match the semantic potential between the knowledge representations (Pirnay-Dummer, Ifenthaler, & Spector, 2009).

6.3.6.8 Triangulation of Types of Expertise

MITOCAR was initially built to compare different kinds of expertise within different groups of experts. Therefore, it comes with a triangulation module which also visualizes two different groups of experts as compared to a non-expert group. To do so, it records the similarities s between the groups into distances d as $d_{i,j}=1-s_{i,j}$. Distances and similarities are calculated between the models i and j. Any similarity index is possible. Next, it transforms the three distances into the radian measure in order to present them as a triangle on a circle. For a circle C with the radius r and the center in x, y the coordinates for the points $P_i(a, b)$ are derived as follows. ξ, q_1, and q_2 are auxiliary variables to make the equations easier to read. They do not have a real meaning:

$$\xi = \sum d_{i,j}$$

$$q_1 = 2\pi \cdot \frac{d_{1,2}}{\xi}$$

$$q_2 = 2\pi \cdot \frac{d_{1,2} + d_{2,3}}{\xi}$$

$$P_1 = (x + r \cdot \sin(q_2), y + r \cdot \cos(q_2))$$

$$P_2 = (x, y + r)$$

$$P_3 = (x + r \cdot \sin(q_1), y + r \cdot \cos(q_1))$$

In the visualization of the triangle, the distances can now be read and interpreted by researchers somewhat more easily:

Figure 6.11 shows what the distance graphs look like. The examples are taken from studies on different kinds of expertise (Pirnay-Dummer, 2006). Not every set of distances may be presented as a triangle. However, in such cases the software would draw the longest distance as an arc rather than a straight line. This would lead to a slight incongruence but still show the differences between the distances.

6.3.7 HIMATT

HIMATT (Highly Integrated Model Assessment Technology and Tools) is a combined toolset which was developed to convey the benefits of various methodological approaches to one environment (Pirnay-Dummer et al., 2009). It can be used by

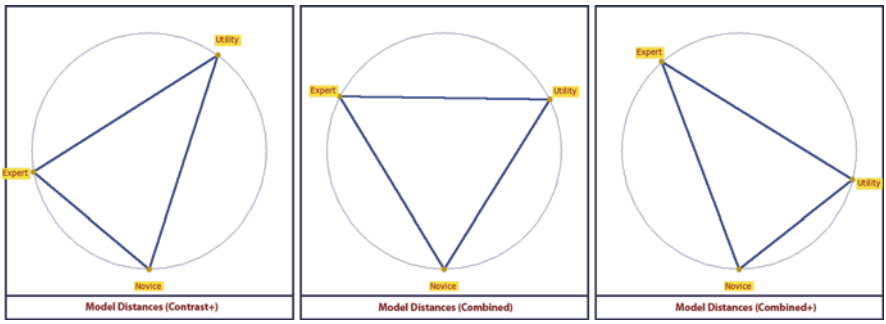


Fig. 6.11 Different model distances between two groups of experts and one group of novices

researchers with only little prior training. HIMATT is implemented to run on the Web and integrates the features of MITOCAR, T-MITOCAR, and SMD Technology discussed above.

The HIMATT architecture consists of two major platforms: (a) HIMATT Research Engine and (b) HIMATT Subject Environment. Functions for conducting and analyzing experiments are implemented within the HIMATT Research Engine. These functions include: (1) experiment management, (2) researcher management, (3) subject management, (4) view function, and (5) analysis and compare function. The HIMATT Subject Environment provides assigned experiments to individual subjects dynamically.

HIMATT was implemented and run on a Web server using Apache, MySQL (MY Structured Query Language), and PERL (Practical Extraction and Report Language), plus additional packages such as GraphViz (Ellson et al., 2003).

The core unit in HIMATT is the experiment, which can be laid out flexibly by the researcher. Experiments in HIMATT consist of three assessment modules: (1) DEEP, (2) T-MITOCAR, and (3) MITOCAR as well as an INSTRUCTION module which is used to give the subject instructions and explanations. The instructions are texts which may contain HTML code (e.g., to link pictures, videos, or other objects). Thus, they may also be used to present simple interventions to the subjects between the assessments, although this option is not very well developed.

Besides mandatory labels and names for experiments, the researcher can add meta-information about them. This allows the researcher to identify the purpose of the experiment and quickly select from a large number of experiments with the help of a search function. The number and sequence of modules and the content of all subject information can be changed any time. Once an experiment is laid out completely, subjects may be assigned to the experiments with the subject management function.

The subject management function includes multiple options. First, a researcher can add subjects to the HIMATT database. Subject information includes at least a username and a password. If a researcher wants to add a large number of subjects, HIMATT can automatically generate a specified number of subjects with individual usernames and passwords. Additionally, the user can add a prefix to all usernames

or passwords in order to more easily identify them later on during experimentation and analysis procedures.

Another important option within the subject management is the assignment of subjects to experiments. Once an experiment has been laid out completely and subjects have been added to the database, researchers can assign subjects to experiments. HIMATT also contains an export function which enables the researcher to export all assigned subjects from an experiment onto a spreadsheet. Naturally, all subject information can be deleted and changed whenever the researcher wishes.

The view function presents the knowledge graph as a picture to the researcher. This function allows the researcher to choose from specific experiments and knowledge graphs, which are then available as Portable Network Graphics (PNG) images for download. Depending on the underlying module (DEEP, T-MITOCAR, or MITOCAR), the graphs will have different features: annotations for DEEP concept maps, associative strengths at the links for T-MITOCAR, and pairwise rated strengths for MITOCAR. Essentially, the standardized re-representation is done in the same way for all three modules using the pairwise stored information from the database and GraphViz (Ellson et al., 2003).

The analysis function consists of descriptive measures to account for specific features of the knowledge structure, like interconnectedness and ruggedness. Using the compare function, researchers may compare any kind of knowledge model with standardized similarity measures (Pirnay-Dummer et al., 2009). These measures range from surface-oriented structural comparisons to integrated semantic similarity measures. The similarity indices range from 0 to 1 for better in-between comparability. Matrices of multiple models can be compared simultaneously. All of the data, regardless of how it is assessed, can be analyzed quantitatively with the same comparison functions for all built-in tools without further manual effort or recoding. Additionally, HIMATT generates standardized images of text and graphical representations.

6.4 AKOVIA

Although HIMATT has already been used by several researchers, it has two design problems worth mentioning. On the one hand, the user interface was accepted by researchers and subjects alike, and it even had a good usability (Pirnay-Dummer et al., 2009). On the other hand, it was a web service which integrated both the data collection and the analysis. Researchers understandably wanted to integrate the data collection into their experiments and studies. However, subjects needed to log into HIMATT in order to input their data as text or draw graphs. They needed to enter another login, username, and password, which might have disturbed the experimental setting in some cases. The second design problem results from the first: We were often given raw data to upload into the HIMATT system so that the researchers could use the analysis facilities on their data. After following this procedure more often than the system had been used through the “front door,” we felt it was time for a complete redesign of the blended methods.

6.4.1 Foundation and Design

Based on our experience with the HIMATT framework, we took the diagnostic toolset one step further and developed AKOVIA. We decided to concentrate our work with AKOVIA on the analysis methods. Instead of limiting the framework to a narrow set of data collection procedures, we redirected our efforts to the development of more interfaces to different methods. The core analysis in AKOVIA is a comprehensive blend of MITOCAR, T-MITOCAR, and the SMD Technology. Thus, it is also based strictly on mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991; Seel, 1991, 2003). The results of the analysis are unchanged. However, the input formats and outputs have been changed to better accommodate the needs of researches, thus allowing more applications as in the original technologies and HIMATT. Hence, more assessment strategies may be used for data collection, which is in following with a major research requirement of the field addressed by Jonassen and Cho (2008):

> After describing the various methods that have been used to manifest or assess mental models, we argued that mental models are multidimensional, so no single form of assessment can be used effectively to describe mental models. In order to manifest mental models, learners need to use computer-based modeling tools to externalize their mental models in the form of computer-based models. Because mental models are multi-dimensional, no single modeling tool can manifest the complexity of mental models. (Jonassen & Cho, 2008, p. 156)

AKOVIA follows at least a good part of this requirement in that it offers several different analysis tools which were initially developed for different purposes and integrates them into a single framework to obtain a more comprehensive perspective on the knowledge externalizations under analysis.

Figure 6.12 provides an overview on the modules of AKOVIA. There are two general input formats (text and graph). Thus, the software can be used to analyze many currently available assessment methods. A standard interface may be used for graphical methods. This interface is derived from SMD and HIMATT and uses the list form. Specific interfaces are under construction. The software can visualize, aggregate, describe in detail, and compare the models. The indices from SMD and MITOCAR are embedded and available for use, as are the text to graph algorithms from T-MITOCAR. In the following paragraphs we introduce the process from input to output in more detail. There are also examples for the AKOVIA scripting technology, which helps handle large data.

6.4.2 AKOVIA Input

The data may be collected by any means the researcher sees fit. Conceptually, AKOVIA supports two different model input formats:

1. Re-representations on graphs (e.g., list form)
2. Re-representations as text

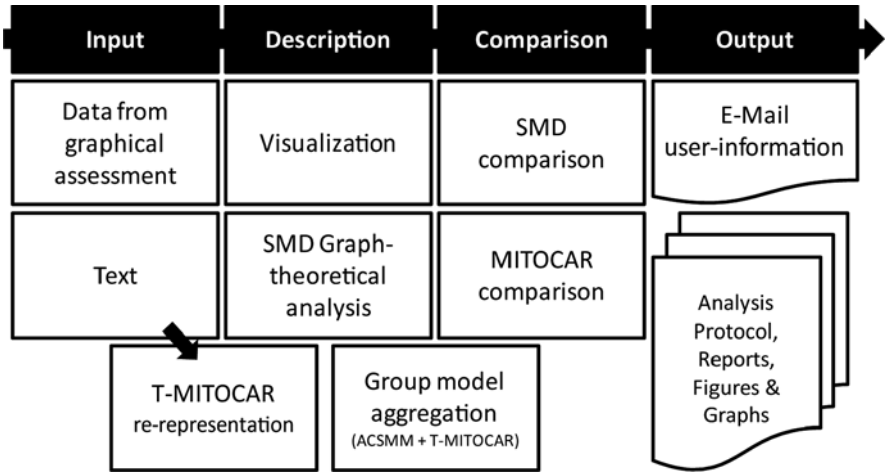


Fig. 6.12 AKOVIA framework

AKOVIA transforms the text into the list form using the corresponding modules from T-MITOCAR. For several technical reasons, MS Excel® files are used to input data into AKOVIA. Although it is unconventional and usually XML is used, we found that the Excel format has several benefits, especially when character sets in plain text sometimes raise incompatibilities. Moreover, in some methodologies the list forms of models are hand-coded and researchers find it easier to work with Open Office and/or Excel to input data. However, in the future we will also work on a stable XML input format to ensure better connectivity with other computer programs.

6.4.2.1 Input from Graphs (List Form)

When graphs are input in a list form, AKOVIA needs an Excel file, which may be uploaded to the AKOVIA server. The Excel file must be called *akovia_listinput.xls*. The first sheet of the Excel file contains the list form. The uploaded file will have to contain at least three columns: (1) model number, (2) concept 1, and (3) concept 2. Each row represents one link within the model. AKOVIA treats all links with the same model number as one model. The heading (first row) must contain the names of the variables. Any number of additional variables is possible, but they will have no effect unless they correspond to a specific method in the analysis. Special variables are reserved (reserved names), e.g., if a column is called "weight," then an association weight $0 \leq w \leq 1$ is expected and treated accordingly within analysis and output. Other reserved names refer to direction (0 = no direction, 1 = forward, 2 = backward, 3 = both), link (an annotation of the link, e.g., a proposition or causal annotation). If the second concept contains an asterisk (*), then the first concept will be integrated.

The second sheet of the Excel file describes the model. Each row contains at least the model number and the language of the model. It may also contain more variables to group the models for analysis and comparison (e.g., measurement time point, treatments, and groups). In the same way, metadata may be included (e.g., Dublin Core, ISO 15836:2009).

We are currently working on interfaces to common graphical assessment tools such as CMap (Cañas et al., 2004) or GraphViz (Ellson et al., 2003). In this case, a ZIP file with all of the individual model files will have to be uploaded alongside an Excel file (e.g., cmap.xls) which contains the second sheet as described above. Additionally, this sheet will need a column with the heading "file." The file which corresponds to the model is specified in each row at this position. The recognition of the file is case sensitive and its name must not contain any special characters (e.g., umlauts) or space characters.

6.4.2.2 Input from Text

Uploading input from text is similar. Each text model artifact has to be contained in a separate text file (acsii, uft-8 encoded). If the format is correct, no manual coding or other preparation of the text is necessary. Like in T-MITOCAR, the software does the necessary coding itself. Additionally, an Excel file needs to be named "tmitocar.xls" and included. Again, the first column contains a model number. The second column contains the name of the corresponding text file. More variables may be provided and used during analysis. All of the texts and the Excel file are collected into a ZIP file, which is then uploaded to AKOVIA.

Input from text is transformed into graphs before the analysis. If a text is too short or if it is not consistent enough to allow the T-MITOCAR algorithm to construct a graph, the user will be notified with the corresponding model number. This information is not directly available after the upload but is part of the analysis.

Both for text and graph input, the analysis refers to the model numbers, which therefore have to be exclusive. The models can be aggregated in different ways during the analysis. New model numbers are specified in the analysis script to correspond to the aggregated models (e.g., a group model).

6.4.2.3 Mixed Format Input

AKOVIA searches for all available formats within an uploaded ZIP file. Thus, multiple formats may be included for the same analysis. If the model numbers are unique in the different files, both analysis and comparison can be carried out. Although technically possible, the research question must correspond to a mixed format, too. Generally, this is the case when the knowledge is assessed on the same task but in different ways. Other usage may not be advisable, e.g., when one group is assessed only with one method and the other group with another. There are differences between the available methods (see Pirnay-Dummer et al., 2009).

6.4.3 Common Model Data Frame

AKOVIA has one common model data frame. All external formats are transformed into this data frame. An AKOVIA model is represented as a hash. The following keys are needed. Others (e.g., metadata from the models) may be included.

1. *Type*: The type of the model (e.g., t-mitocar, "list," or "CMap").
2. *Status*: Usually this is 1 or 0. 0 means that there is no model in the data frame (e.g., if the amount of text is not sufficient for a re-representation in T-MITOCAR. 1 means that the model is fully contained and readable for further analysis. Special codes are possible, e.g., a status of 2 in T-MITOCAR means that there is text but that it does not yield enough consistency to generate a model.
3. Weights_are_in: If the model does not have association weights at the links, then this value is 0. Otherwise, it points at the position (column) in the list form where the weight is found.
4. *Surface*: The number of propositions (links, associations) in the model.
5. *Language*: An indicator for the language of the model, e.g., "en" for English or "de" for German. This tag allows us to integrate other languages more easily.
6. *Model*: A reference to the list form in which the actual model data is contained.
7. *Stemmed_model*: A reference to the list form which is exactly like *model* except that it contains the word stems instead of the words. This is needed for several of the comparison and similarity measures.

Depending on the type of model, other keys might be necessary to conduct a full analysis, e.g., for T-MITOCAR, the initial raw text is also stored within the key "text."

6.4.4 Analysis and Scripting

For the model analysis and comparison, a very simple scripting technique is used. Within this system, the users refer to the models by the model number they provided in the overview Excel sheet. Another Excel file named "akovia-script.xls" provides a list of analysis and comparison steps to be performed by AKOVIA. The file has only one sheet and a table with three columns. The first column lists the AKOVIA command. The second column displays the specifications for the command, and the third column contains the arguments for the command (model numbers, most of the time). Some selected commands will be introduced in the following short paragraphs. A full list is provided by the AKOVIA reference manual. The scripting technique will gain more options as we embed more model types, analysis functions, and comparison technologies.

6.4.4.1 Visualize

The *visualize* command draws a graphical output of the model. The specifications allow control over what gets represented at the links (e.g., the weight or a variable in the list form which may contain annotations). Also, the output format (PNG, SVG) may be specified. The argument is either a single model or two models. In the first case, the model is just drawn as is. Instead of providing a single model number, the ALL argument is also allowed, meaning that simply every model in the dataset is drawn. In the second case, the notation decides whether a difference model or a union set model is output: 8–22 would draw a difference model which contains all the links of the model number 8 which are not part of the model number 22. 7.9 would draw a union set model in which only links and nodes are included which are in both model 7 and 9.

Table 6.5 shows how the examples from above look in *akovia-script.xls*. To determine the matching for the union set and difference models, one can use word stems by referring to the list form in the key *stemmed_model*. If, e.g., a union set model does not exist, nothing will be drawn and a note will be written to the report file.

Table 6.5 Scripting examples for the command "visualize"

Command	Specifications	Arguments
Visualize	Weight;SVG	5
Visualize	Weight;PNG	8–22
Visualize	Weight;PNG	7.9
Visualize	Weight;SVG	ALL

6.4.4.2 Ganalyze

With the command *ganalyze* (short for graph analysis), graph features of single models will be computed. The specification field is left empty, and the argument may be a single model or ALL. An Excel file named "ganalize.xls" is generated if at least one ganalyze command is on the list. This document contains the list of model numbers in they order in which they appear in the command list, any data (e.g., group variables) from the original model list as uploaded, and the graph theoretical indices as they are built into AKOVIA. Ifenthaler (2008b) provides an overview of the indices which are interpretable for knowledge structures.

6.4.4.3 Compare

The *compare* command carries out model comparisons. The specifications contain the similarity measures which should be applied. If all measures should be applied, then the specifications should read EVERY. The arguments contain a pair of model numbers. Thus, if an expert model is in number 1 and learner models are in numbers 2–12, then one can compare all of them to the expert model by subsequently applying the compare command (1;2), (1;3), (1;4), . . ., (1;12). If at least one comparison is carried out, the file *comparison.xls* will be created, which contains all comparisons in the order in which they appear in the command list.

6.4.4.4 Aggregate

The *aggregate* command groups individual models into an aggregated model. The specifications select the method of aggregation and how it is applied. Up to now there are two such methods available. The ACSMM aggregation method (Johnson, Ifenthaler, Pirnay-Dummer, & Spector, 2009) may be applied to every model type. If it is selected as aggregation method, the amount of matching needs to be specified in brackets (e.g., ACSMM[0.5]). With [0.5] a proposition will make it into the aggregation if it is contained in at least 50% of the models in the list. Multiple ACSMM aggregations (e.g., from 0.1 to 0.9) may help to qualitatively track and visualize the level of agreement within a group. A value of 0 selects every proposition, even if it occurs only once in the group (union set). A value of 1 generates the intersection: Only propositions which appear in *every* model of the group are selected. If the set from the ACSMM analysis is empty, the model will be tagged with the status 0 (no model). A corresponding note will be written to the protocol file whenever later commands try to use the model.

If all of the models which one wishes to aggregate are text based, then T-MITOCAR may also be specified. One then creates the new model by using T-MITOCAR on all of the texts by reading out the *text* key from the model hashes. The first model number on this list is the new model number the aggregated model is stored in.

The arguments may be the individual model numbers which will be aggregated, separated by semicolons. The first model number on this list is the new model number the aggregated model is stored in: 210;4;19;1;102 creates a model with the number 210 which is an aggregation of the models in 4, 19, 1, and 102. The *type* of the model will either be *acsmm* or *t-mitocar*. If the new model number already exists, an error will be written to the protocol file. After aggregation, the model number can be used like any other model which is already in the system. It may be used for visualization, analysis, and comparison and even as a model for further aggregation. The model number can only be used if the command appears before any command which refers to the new model number.

6.4.5 Upload, Feedback, and Analysis

After uploading the data in a ZIP file, the data is pre-checked by the upload server. Both the data files and the script files are checked. The user receives an email along with a ticket number of the data analysis process if the format is valid. Otherwise, the user receives a list of errors which occurred and may choose to upload the data again after a revision.

6.4.6 Server Topology

AKOVIA places no explicit limits on the size of data which can be investigated and analyzed. Large concurrent analyses used to slow our servers down to the point

where the browser experienced time outs (both in HIMATT and T-MITOCAR). Therefore, we separated the topology of the small analysis grid into the *upload server*, which takes in the files, and the *analysis servers*. The latter access the upload server and process the tickets offline. Afterward, the results are uploaded to the upload server and the user is notified. Depending on the number and size of concurrent jobs, a response may take hours or sometimes even days. Figure 6.13 shows a simplification of the server topology.

When users upload data, they only receive an initial confirmation email with either a list of errors (if the data is not formatted correctly) or a short note confirming that their data is being processed. If the data is correct, an available analysis server downloads the files as soon as it has finished previous analyses. After completing a script, the analysis server packs and uploads the results and a protocol to the upload server, which sends an email to the user. The email contains abbreviated information on the progress of the analysis and a protected download link with which the user can access the package for a limited time. Afterward, the package is deleted from the server as is the download link.

6.4.7 Data Warehousing

The data remains in the AKOVIA system from the upload until the analysis. All data is completely deleted after the analysis is sent to the user. We do not have plans for a long-term data warehousing. We only store the analysis protocol along with the ticket number in case there are questions about the process later on. Should the results be lost, users will of course be able to upload and order an analysis again.

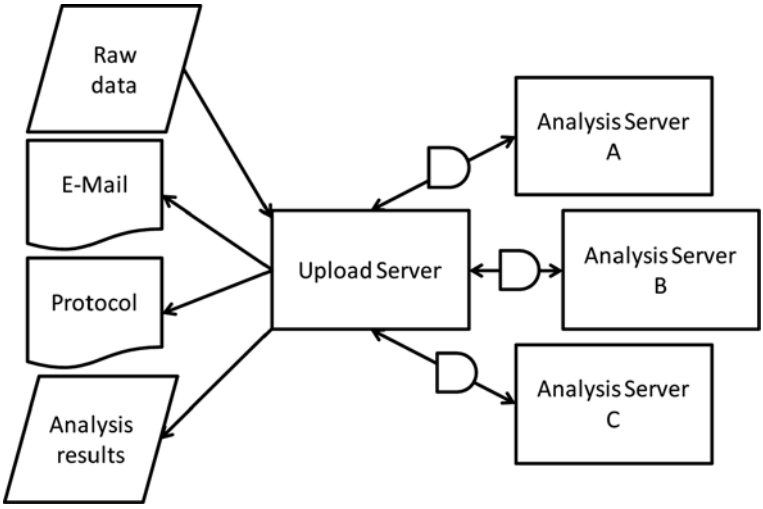


Fig. 6.13 AKOVIA server topology

6.5 Applications And Future Perspectives

The design and development of useful diagnostic systems has always been a goal for researchers and engineers in an effort to serve professional communities in the field of learning and instruction. Future work on automated computational diagnostics will provide more and more powerful systems for the comprehensive analysis of large amounts of data in a short space of time. The main applications will be the enhancement or complementation of existing assessment strategies as well as technologies to support self-assessment for learners.

6.5.1 Applications

The technologies discussed in this chapter aim at the assessment, re-representation, analysis, and comparison of knowledge. The tools were developed independently and integrated step by step afterward. Two comprehensive toolsets emerged. HIMATT focuses on the assessment side and may be used mainly for laboratory experiments or online studies with the usual limitations (e.g., lack of control over who is performing the task and where the material comes from). It also shows limits in the interface. Integrating the toolset into any kind of main software is not very easy. Thus, the main applications of AKOVIA are clearly in analysis and comparison, whereas the assessment step itself is left to the tools and experimental setups of the researchers. AKOVIA is designed to complement any kind of technology as it uses interfaces which allow all kinds of data to be analyzed. The visualizations of the models have shown to have an especially positive effect on learning within tasks which involve writing. Thus, the possible applications reach beyond the structural and semantic analysis and comparison of knowledge. In addition, AKOVIA allows the development of self-assessment technologies. In this case, specifically formatted and interpreted outputs of the analysis and comparison may be embedded into feedback, e.g., on ongoing writing.

6.5.2 Future Perspectives

There are three main areas of interest for further development: the *interfaces* for different commonly used formats to AKOVIA, the design and validation of *more analysis and comparison measures*, and the development of *tools for self-assessment*. The last of these issues will involve a major transformation of the program surfaces, their usability, and also the output. Using graph and its numeric features to generate feedback in natural language for learners will be one of the next challenges. Technologies like this could provide individual coaching for situations in which usually no individual mentoring is possible (e.g., due to large classes).

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *Journal of Technology, Learning, and Assessment, 1*(5), 3–63.

Bonato, M. (1990). *Wissenstrukturierung mittels Struktur-Lege-Techniken. Eine grapentheoretische Analyse von Wissensnetzen*. Frankfurt am Main: Lang.

Brill, E. (1995). *Unsupervised learning of dismabiguation rules for part of speech tagging.* Paper presented at the Second Workshop on Very Large Corpora, WVLC-95, Boston.

Cañas, A. J., Hill, R., Carff, R., Suri, N., Lott, J., Eskridge, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak, & F. M. González (Eds.), *Concept maps: Theory, methodology, technology, proceedings of the first international conference on concept mapping* (pp. 125–133). Pamplona: Universidad Pública de Navarra.

Chartrand, G. (1977). *Introductory graph theory*. New York: Dover.

Chung, K. L. (1968). *A course in probability theory*. New York: Harcourt, Brace & World.

Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2003). *GraphViz and Dynagraph. Static and dynamic graph drawing tools*. Florham Park, NJ: AT&T Labs.

Hannafin, M. J. (1992). Emerging technologies, ISD, and learning environments: Critical perspectives. *Educational Technology, Research & Development, 40*(1), 49–63.

Harary, F. (1974). *Graphentheorie*. München: Oldenbourg.

Ifenthaler, D. (2006). *Diagnose lernabhängiger Veränderung mentaler Modelle. Entwicklung der SMD-Technologie als methodologisches Verfahren zur relationalen, strukturellen und semantischen Analyse individueller Modellkonstruktionen*. Freiburg: FreiDok.

Ifenthaler, D. (2008a). Practical solutions for the diagnosis of progressing mental models. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 43–61). New York: Springer.

Ifenthaler, D. (2008b). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*. doi: 10.1007/s11423-008-9087-4

Ifenthaler, D. (in press-a). Learning and instruction in the digital age. In J. M. Spector, D. Ifenthaler, P. Isaías, Kinshuk, & D. G. Sampson (Eds.), *Learning and instruction in the digital age: Making a difference through cognitive approaches, technology-facilitated collaboration and assessment, and personalized communications*. New York: Springer.

Ifenthaler, D. (in press-b). Model-based feedback for improving expertise and expert performance. *Technology, Instruction, Cognition and Learning*.

Ifenthaler, D., Masduki, I., & Seel, N. M. (2009). The mystery of cognitive structure and how we can detect it. Tracking the development of cognitive structures over time. *Instructional Science*. doi: 10.1007/s11251-009-9097-6

Janetzko, D. (1996). *Knowledge tracking. A method to analyze cognitive structures*. Freiburg: Albert-Ludwigs-Universität.

Janetzko, D. (1999). *Statistische Anwendungen im Internet. Daten in Netzumgebungen erheben, auswerten und präsentieren*. München: Addison-Wesley.

Johnson-Laird, P. N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness*. Cambridge, UK: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. (1991). *Deduction*. Hove: Lawrence Erlbaum.

Johnson, T. E., Ifenthaler, D., Pirnay-Dummer, P., & Spector, J. M. (2009). Using concept maps to assess individuals and team in collaborative learning environments. In P. L. Torres & R. C. V. Marriott (Eds.), *Handbook of research on collaborative learning using concept mapping* (pp. 358–381). Hershey, PA: Information Science Publishing.

Jonassen, D. H., & Cho, Y. H. (2008). Externalizing mental models with mindtools. In D. Ifenthaler, P. Pirnay-Dummer & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 145–160). New York: Springer.

Kirschner, P. A. (2004). Introduction to part II of the special issue: Design, development and implementation of electronic learning environments for collaborative learning. *Educational Technology, Research & Development, 52*(4), 37.

Koper, R., & Tattersall, C. (2004). New directions for lifelong learning using network technologies. *British Journal of Educational Technology, 35*(6), 689–700.

MySQL AB. (2008). MySQL. The world's most popular open source database. Retrieved 11.02.2008, from http://www.mysql.com/

Pirnay-Dummer, P. (2006). *Expertise und Modellbildung: MITOCAR*. Freiburg: FreiDok.

Pirnay-Dummer, P. (2008). Rendevous with a quantum of learning. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 105–144). New York: Springer.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2009). Highly integrated model assessment technology and tools. *Educational Technology Research and Development*. doi: 10.1007/s11423-009-9119-8

Seel, N. M. (1991). *Weltwissen und mentale Modelle*. Göttingen: Hogrefe.

Seel, N. M. (1999). Educational diagnosis of mental models: Assessment problems and technology-based solutions. *Journal of Structural Learning and Intelligent Systems, 14*(2), 153–185.

Seel, N. M. (2003). Model-centered learning and instruction. *Technology, Instruction, Cognition and Learning, 1*(1), 59–85.

Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning in complex domains (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building)*. Syracuse, NY: Syracuse University.

The Apache Software Foundation. (2008). Apache HTTP Server Project. Retrieved 11.02.2008, from http://httpd.apache.org/

The Perl Foundation. (2008). Perl 6. Retrieved 11.02.2008, from http://www.perlfoundation.org/

The PHP Group. (2008). PHP 4.4.8. Retrieved 11.02.2008, from http://www.php.net/

Tittmann, P. (2003). *Graphentheorie. Eine anwendungsorientierte Einführung*. München: Carl Hanser Verlag.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Chapter 7
Deriving Individual and Group Knowledge Structure from Network Diagrams and from Essays

Roy B. Clariana

7.1 Introduction

This chapter focuses on two somewhat fundamentally related ways to elicit knowledge structure, network diagrams (usually as concept maps), and essays (see Chapter 4, this volume). Because of the utility of the proven Pathfinder Network approach and its well-established research base, we developed software to convert concept maps and essays into data representations that can be analyzed with Pathfinder software. Our initial research focused on computer-based methods for scoring concept maps (e.g., Clariana, 2002) and since concepts maps are frequently used in classrooms to replace outlining as an organizational aid for writing essays, we became interested in measuring the relationship between concept maps and the essays derived from these maps (Clariana & Koul, 2008; Koul, Clariana, & Salehi, 2005), and so it was a natural progression to develop software based on our concept map scoring approach to score essays (*ALA-Reader*, 2004). This chapter begins by describing Pathfinder network analysis and then describes the *ALA-Mapper* and *ALA-Reader* scoring approach. Next the investigations with these two tools are reviewed, and finally suggestions for future research are provided.

7.2 Pathfinder Network Analysis

Pathfinder network analysis is a well-established system for deriving and representing the organization of knowledge (Jonassen, Beissner, & Yacci, 1993; Schvaneveldt, 1990). The Pathfinder algorithm converts estimates of relatedness of pairs of terms into a network representation of those terms called Pathfinder Networks (*PFNETs*) that are usually a two-dimensional representation of a matrix of relationship data in which concept terms are represented as nodes (also called

R.B. Clariana (✉)
The Pennsylvania State University, University Park, PA, USA
e-mail: rclariana@psu.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_7,

vertices) and relationships are shown as weighted links (also called edges) connecting the nodes. *PFNETs* resemble concept maps, but without link labels.

There are three steps in the Pathfinder approach. In Step 1, raw proximity data is collected typically using a word-relatedness judgment task. Participants are shown a set of terms two at a time, and judge the relatedness of each pair of terms, for example, on a scale from one (low) to nine (high). The number of pairwise comparisons that participants must make is $(n^2 - n)/2$, with n equal to the total number of terms in the list.

In Step 2, the Knowledge Network and Orientation Tool for the Personal Computer software (KNOT, 1998) is used to reduce the raw proximity data into a *PFNET* representation that is a least-weighted path that links all of the terms. The set of links derived from Pathfinder analysis is determined from the patterns in the raw proximity data, and these are influenced by two parameters that can be manipulated by the researcher, q and Minkowski's r. These parameters for calculating the least-weighted path can be adjusted to reduce or prune the number of links in the resulting *PFNET* (refer to Dearholt & Schvaneveldt, 1990). The resulting *PFNET* is purported to represent the most salient relationships in the raw proximity data.

In Step 3, the comparison of the participant's *PFNET* to an expert or other referent *PFNET* is calculated also using *KNOT* software (Goldsmith & Davenport, 1990). The two most commonly reported similarity measures are Common and Configural Similarity. Common is the number of the links shared by two *PFNETs* (the intersection of two *PFNETs*). Similarity, which is also called neighborhood similarity, is the intersection divided by the union of two *PFNETs*.

Note that *KNOT* has a group average feature that can average multiple proximity files to obtain a group average *PFNET* representation. This feature is especially useful for comparing one group to another or for comparing different groups to some referent. Our experience is that group average *PFNETs* are more robust than individual *PFNETs*. Averaging seems to remove idiosyncratic and error responses contained in individual *PFNETs*.

7.3 Network Diagrams and Knowledge Structure

What information components of concept maps can be collected automatically by a computer and how can the Pathfinder approach be used to score concept maps? Concept maps consist of terms, links, and link labels (i.e., propositions); also there is an overall visual layout of the map that consists of patterns of links and also closeness of terms. *ALA-Mapper* uses either links between terms or distances between terms as an alternative to word-relatedness judgment tasks in Step 1 for obtaining raw proximity data, while Steps 2 and 3 are conducted in the conventional way. Thus, the main contribution of *ALA-Mapper* is its capability to convert components of a concept map into raw proximity data in Step 1 of the Pathfinder analysis approach (Taricani & Clariana, 2006).

So what information in concept maps can be measured? There are at least three or four different cognitive tasks involved when creating a concept map that can leave a

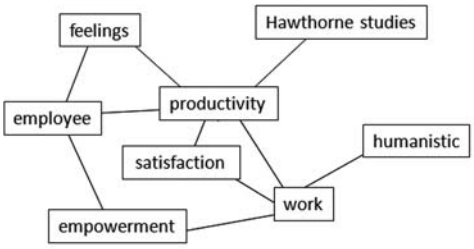
"cognitive residue" in the map. First, if the map is open-ended, where students may use any terms in their map, then a critical task is recalling (or possibly recognizing from a list) the most important terms/concepts to include in the map. Alternately, if a list of terms is provided and the students are told to use all of the terms (fixed or closed mapping), then recall of terms is not a factor. Note that it is easier for both instructors and computers to score closed maps compared to open maps. Next, students must group related terms together, often in an intuitive way, and this most likely relates to their internal network structure of associations. Then students identify propositions by linking pairs of terms with a line and adding a linking phrase to show the meaning of the proposition in that context. While students work on the later stages of their map, they continually revise small components of their map making it easier to grasp, and this also seems to be an intuitive activity of making it "feel" right that likely reflects both the structure of their knowledge and an internalized graphic grammar or norm of what things like this should look like (Clariana & Taricani, 2010).

Probably the most fundamental meaningful psychological components of concept maps and essays are propositions (Einstein, McDaniel, Bowers, & Stevens, 1984; Kintsch, 1974; Kintsch & van Dijk, 1978). In essence, a proposition consists of a subject, a verb, and an object, which in a concept map consists of a term-(linking phrase)-term. *ALA-Mapper* converts node–node information in concept maps and other types of network diagrams into two separate kinds of raw data, links between terms and distances between terms measured in screen pixels (see Fig. 7.1).

Note that linking phrases may not be critical for concept map analysis. Harper, Hoeft, Evans, and Jentsch (2004) reported that the correlation between just counting link lines (i.e., node–node) compared to counting valid propositions (i.e., node–label–node) in the same set of maps was $r = 0.97$, suggesting that link labels add little additional information over just counting links. Also, link labels are more computationally difficult to collect, handle, compare, and analyze than just the presence or absence of a link between terms. Thus *ALA-Mapper* pragmatically uses links only rather than matching link labels.

7.3.1 ALA-Mapper Investigations

Clariana, Koul, and Salehi (2006) used *ALA-Mapper* for scoring open-ended concept maps. Practicing teachers enrolled in graduate courses constructed concept maps on paper while researching the topic, "the structure and function of the heart and circulatory system" online. Participants were given the online addresses of five articles that ranged in length from 1,000 to 2,400 words but were encouraged to view additional resources. After completing their research, participants then used their concept map as an outline to write a 250-word text summary of this topic. *ALA-Mapper* was used to measure the distances between terms in the concept maps and to represent the links that connected terms (see Fig. 7.1). Using Pathfinder *KNOT* software, the raw distance and link data were converted into *PFNETs* and



	empowerment	employee	feelings	Hawthorne studies	humanistic	productivity	satisfaction	work
empowerment	--							
employee	1	--						
feelings	0	1	--					
Hawthorne studies	0	0	0	--				
humanistic	0	0	0	0	--			
productivity	0	1	1	1	0	--		
satisfaction	0	0	0	0	0	1	--	
work	1	0	0	0	1	1	1	--

Link array

	empowerment	employee	feelings	Hawthorne studies	humanistic	productivity	satisfaction	work
empowerment	--							
employee	66	--						
feelings	97	50	--					
Hawthorne studies	166	160	120	--				
humanistic	143	161	135	57	--			
productivity	91	87	60	77	77	--		
satisfaction	52	68	66	114	98	39	--	
work	114	143	127	83	34	68	76	--

Distance array

Fig. 7.1 A sample network diagram and its link array and distance array

then were compared to an expert's *PFNET* to obtain network similarity scores. Five pairs of raters using rubrics also scored all of the concept maps and text summaries. The Pearson correlation values for the concept maps scored by raters compared to: (a) *ALA-Mapper* link-based scores were 0.36, (b) *ALA-Mapper* distance-based scores were 0.54, and (c) text summaries scored by raters was 0.49; thus the *ALA-Mapper* distance scores were a bit more like the raters' concept map scores than were the link scores. The correlation values for the text summaries scored by raters compared to (a) *ALA-Mapper* link-based scores were 0.76 and (b) *ALA-Mapper* distance-based scores were 0.71; thus the *ALA-Mapper* link and distance scores were both quite like the raters' text summary scores.

In a follow-up study, Poindexter and Clariana (2004) used the same Pathfinder scoring technique applied to posttest network diagrams (e.g., no linking phrases) rather than concept maps. The mapping directions specifically directed the participants to use spatial closeness to show relationships and intentionally deemphasized the use of links. Participants completed one of three print-based text lesson treatments on the heart and circulatory system. The three lesson treatments included adjunct constructed response questions (an item-specific approach that emphasizes propositions), scrambled-sentences (a relational approach that emphasizes concept associations), and a reading only control. Participants then completed three multiple-choice posttests that assessed identification, terminology, and comprehension and were finally asked to draw a network diagram given a list of 25 pre-selected terms. The adjunct question lesson treatment was significantly more effective than

the other lesson treatments for the comprehension outcome, and no other treatment comparisons were significant. *ALA-Mapper* network diagram scores based on link data were more related to terminology ($r = 0.77$) than to comprehension ($r = 0.53$), while *ALA-Mapper* network diagram scores based on distance data were slightly more related to comprehension ($r = 0.71$) than to terminology ($r = 0.69$). It was suggested that the links drawn to connect terms related to verbatim knowledge from the lesson text covering facts, terminology, and definitions; while the distances between terms in the network diagram related to comprehension of the processes and functions of the heart and circulatory system.

In a follow-up study, Taricani and Clariana (2006) asked 60 undergraduate students to read a print-based instructional text on the heart and circulatory system and then create concept maps of that content. Half of the participants were given feedback in the form of a prepared hierarchical concept map and the other half did not receive this feedback map. Then all completed a multiple-choice posttest with 20 terminology and 20 comprehension questions. The concept maps were scored using *ALA-Mapper* and these concept map scores were compared to the terminology and comprehension posttest scores. Similar to Poindexter and Clariana (2004) above, concept map scores derived from link data were more related to terminology ($r = 0.78$) than to comprehension ($r = 0.54$) whereas concept map scores derived from distance data were more related to comprehension ($r = 0.61$) than to terminology ($r = 0.48$).

This supports the idea that there is worthwhile relational information in the distances between terms in a network diagram. However, our view is that this distance information is fragile, and that pre-map training or strong directions that emphasize proposition-specific elements in the map damages the distance information captured in the map. For example, concept map training that demands that all map elements be propositions that are term-(linking phrase)-term direct the participants' focus to those elements and away from distance-related relational aspects of their knowledge. Ironically, if the rubrics used to score these concept maps are also strongly proposition oriented, then those maps that do have a focus on propositions will score relatively higher, thus confirming that propositions are key elements in concept maps. In contrast, the Poindexter and Clariana (2004) investigation described above provided mapping directions that intentionally deemphasized propositions (links and linking phrases were optional) and emphasized distances between terms, with the result that the distance scores obtained larger correlations with the comprehension measures. Thus, investigators must be sensitive to the relational or proposition-specific effects of their pre-map training, their mapping directions, and the rubrics used to score the maps.

7.3.2 Rubrics and Network Diagram Scores

So far, the investigations above have avoided the issue of what precisely are raters scoring when they score a concept map or other type of network diagram. Those studies above used holistic scoring that only considered the content accuracy reflected in the concept maps, a quantitative approach where more correct

ideas obtains a higher score form the raters. However, Koul, Clariana, and Salehi (2005) reported that *ALA-Mapper* data correlated better with raters' scores using a qualitative than a quantitative rubric. In their investigation, teachers enrolled in a graduate course worked in pairs to research a science topic online and then created a concept map of the topic. Later, participants individually wrote a short essay from their concept map. The concept maps and essays were scored by *ALA-Mapper* and *ALA-Reader* and by human raters using qualitative and quantitative rubrics. The quantitative rubric was adapted from the Lomask, Baron, Greig, and Harrison (1992) rubric. This rubric considered *size* (the count of terms in a student map expressed as a proportion of the terms in an expert concept map) and *strength* (the count of links in a student map as a proportion of necessary, accurate connections with respect to those in an expert map). The qualitative rubric for scoring concept maps was based on research by Kinchin and Hay (2000). This rubric deals with three common map structures which may be interpreted as indicators of progressive levels of understanding: (1) *Spoke*, a structure in which all of the related aspects of the topic are linked directly to the core concept, but are not directly linked to each other; (2) *Chain*, a linear sequence of understanding in which each concept is only linked to those immediately nearby; and (3) *Net*, a network both highly integrated and hierarchical, demonstrating a deep understanding of the topic. *ALA-Mapper* concept map scores were a good measure of the qualitative aspects of the concept maps (link $r = 0.84$ and distance $r = 0.53$) and were an adequate measure of the quantitative aspects (link $r = 0.65$ and distance $r = 0.50$).

These various results were evidence to convince us that *ALA-Mapper* scores were not really concept map "content" scores, but rather that *ALA-Mapper* scores are a measure of structural knowledge that correlates somewhat with some forms of concept map content scores as well as with different kinds of traditional posttests. We hold that this measure of knowledge structure is tapping a fundamental level of knowledge, the association network that can be drawn from to create meaningful propositions on the fly. Also, distance data and link data in network diagrams can both contain interesting and useful information. However, internal and external context factors can enhance or suppress the information content in this raw data and so must be well controlled. So our focus turned to measuring knowledge structure and on refining the writing prompts to elicit better knowledge representations.

7.4 Essays and Knowledge Structure

The *ALA-Reader* essay analysis approach was adopted directly from the *ALA-Mapper* network diagram analysis approach. *ALA-Reader* searches for key terms in text that are then represented as links in an aggregate array either (a) between all terms that occur in the same *sentence* or else (b) between consecutive terms in a *linear* pass through the text. The aggregate array raw data for a text processed by *ALA-Reader* is similar in form to the network diagram link raw data (see the bottom left portion of Fig. 7.1) and is analyzed by Pathfinder *KNOT* software in the same way.

Compare–contrast type essay questions have been used to assess relational understanding that is part of knowledge structure (Gonzalvo, Canas, & Bajo, 1994) although any text genre is likely influenced by the writer's knowledge structure. Goldsmith, Johnson, and Acton (1991) state "Essay questions, which ask students to discuss the relationships between concepts, are perhaps the most conventional way of assessing the configural aspect of knowledge" (p. 88). It is rather critical to keep in mind that an essay contains different kinds of information, and that the scoring approach determines what is actually measured and most if not all essay-scoring approaches, human- or computer-based, do not intentionally measure knowledge structure. But whether it is intentionally measured or not, essays contain at least a reflection of an individual's knowledge structure.

7.4.1 Sentence Aggregate Approach

The *ALA-Reader sentence aggregate approach* was developed to analyze text at the sentence level because sentences are an important unit of text organization. Sentences contain one or more propositions and the sentence aggregate approach seeks to capture the important node–node associations represented by propositions in sentences. To analyze sentences in text, first *ALA-Reader* disregards all of the words in the text except for pre-selected key terms (and their synonyms and metonyms). Then the key terms that co-occur in the same sentence are represented in a proximity array, the lower triangle of an n-by-n array containing $(n^2 - n)/2$ elements. Each cell in the array corresponds to a pair of key terms (see the left panel of Fig. 7.2). A "1" entered in the appropriate cell of the array indicates that two key terms co-occurred in the same sentence and a "0" indicates that those two key terms did not occur in the same sentence. The software continues to aggregate sentences into the array until all of the text is processed. For example, given the

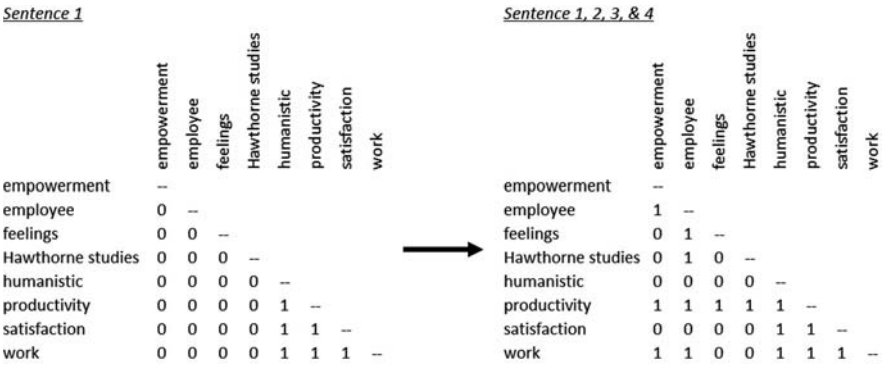


Fig. 7.2 The sentence aggregate proximity file created by *ALA-Reader* for sentence 1 (*left panel*) and for all four sentences (*right panel*)

following four sentences of a participant's essay regarding the humanistic management approach from Clariana, Wallace, and Godshalk (2008) shown with key terms or their synonyms underlined:

> *Humanists* believed that *job satisfaction* was related to *productivity*. The *Hawthorne studies* tried to determine if lighting caused people to be more *productive employees*. However, it was found that *employees* valued being selected to participate in the study and were more *productive* when they *felt* "special." They found that if *employees* were given more *freedom and power* in their *jobs*, then they *produced* more.

These four sentences would be translated by *ALA-Reader* into this sentence aggregate array of selected key terms shown in Fig. 7.2. The force directed network diagram of the four sentences is shown in Fig. 7.3.

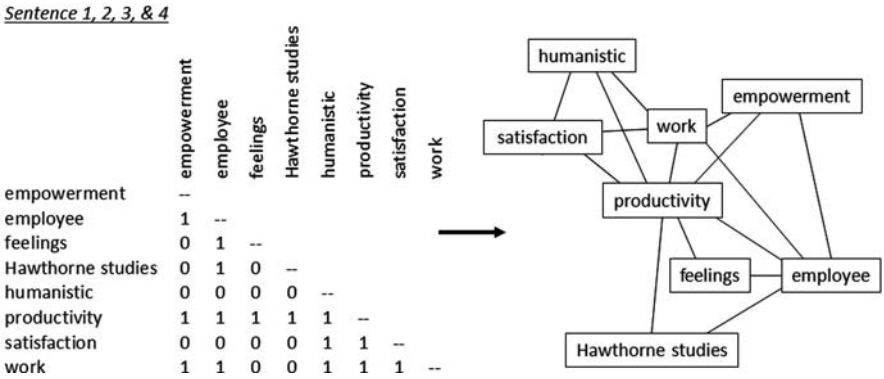


Fig. 7.3 The force-directed graph of the four-sentence aggregate

7.4.2 Linear Aggregate Approach

In contrast to the sentence aggregate approach, the *ALA-Reader linear aggregation approach* enters a "1" (1s) in the appropriate cell of the array to represent adjacent key terms during a linear pass through the text, and so will always obtain a connected graph. However, the result is almost certainly not just a linear chain of words, as important words are used multiple times in the essay passage, those terms will have more links coming in and going out, and the structure when represented as a force-directed network diagram begins to fold bringing related terms closer together in the two-dimensional space. The same text used above in Fig. 7.2 would be reduced by *ALA-Reader* to this linear sequence of selected key terms (with link numerical order shown here for clarity): "humanistic -1- work -2- satisfaction -3- productivity -4- Hawthorne studies -5- productivity -6- employee -7- employee -8- productivity -9- feelings -10- employee -11- empowerment -12- work -13- productivity" (see the linear visual representation of this sequence in Fig. 7.4 and its *PFNET*). The linear sequence of terms can also be represented as a force-directed graph (a *PFNET*)

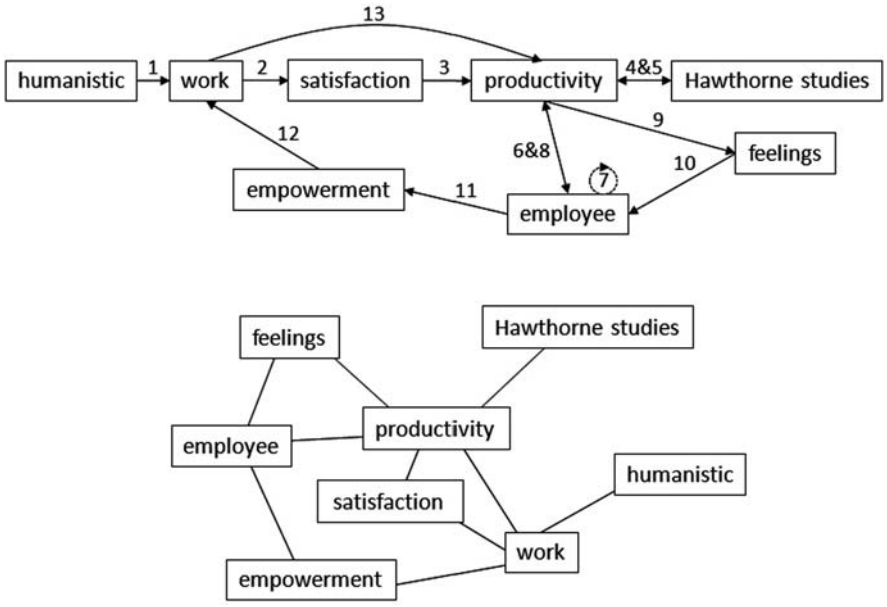


Fig. 7.4 A visual representation of the linear sequence of key terms from a participant's essay passage about the Humanistic management approach (*top panel*) and its equivalent force-directed graph (*PFNET, bottom panel*)

that highlights the more and less salient relationships in the passage based on the degree of the nodes but also provides some idea of possible indirect relationships based on spatial closeness (see the bottom panel of Fig. 7.4). For example, the key term "productivity" with five links is a high-degree node (i.e., with three or more links) and so is a central node in this *PFNET*. This indicates that the student's essay passage describes the humanistic management approach in terms of its relationship to productivity. The terms "work" and "employee" are also high-degree nodes and so are also important terms. Also, compare the number and pattern of links for the sentence aggregate *PFNET* in Fig. 7.3 to the linear aggregate *PFNET* of the same text shown in Fig. 7.4 to note the difference in *ALA-Reader* representation output between the sentence and linear approaches.

Clariana and Koul (2004) used *ALA-Reader* software to score 12 students' essays on the structure and function of the heart and circulatory system relative to an expert's essay. At that time the software could only analyze using the sentence aggregate approach. For benchmark comparison, the essays were also scored by 11 pairs of human raters and these 11 scores were combined together into one composite essay score, then all scores were correlated with the human raters' composite score. Compared to the composite score, the *ALA-Reader* scores were 5th out of 12, with an $r = 0.69$ (the 12 scores correlations ranged from $r = 0.11$ to 0.86), which indicates that four raters were better than the *ALA-Reader* sentence aggregate

scores, but eight raters were worse. Also, the *ALA-Reader* scores were not included when creating the composite score, and so this is a conservative comparison that strongly favors the raters' scores.

Koul et al. (2005) used the *ALA-Reader* sentence aggregate approach to score students' essays on the structure and function of the heart and circulatory system. Working in pairs, participants researched this topic online and created concept maps using *Inspiration* software. Later, using their concept map, participants individually wrote a short essay. The concept maps and essays were scored by *ALA-Mapper* and by *ALA-Reader* relative to an expert's map and essay, by another software tool called Latent Semantic Analysis (*LSA*), and by 11 pairs of human raters using two different rubrics. As in the previous study, the 11 rater scores were averaged together into one composite essay score (a conservative value that favors the raters). Compared to the composite essay score, the *ALA-Reader* essay scores were 5th out of 13, with an $r = 0.71$ (the 13 scores ranged from $r = 0.08$ to 0.88) and *LSA* scores were 9th out of 13, with an $r = 0.62$. As before, relative to the rater composite score, *ALA-Reader* performed better than eight of the raters and also was better than *LSA* on this specific biology content essay.

Clariana and Wallace (2007) used *ALA-Reader* to score essays on management theories relative to an expert referent and also to establish and compare group average knowledge representations derived from those essays. As part of their final course examination, undergraduate business majors ($N = 29$) were asked to write a 300-word compare-and-contrast essay on four management theories from the course, a relevant and high stakes essay. The essays were scored by *ALA-Reader* using both a sentence and a linear aggregate approach. To serve as benchmarks, the essays were also separately scored by two human raters who obtained a Spearman rho inter-rater reliability of $\rho = 0.71$. The linear aggregate approach obtained larger correlations with the two human raters (ρ rater 1 = 0.60 and ρ rater 2 = 0.45) than did the sentence aggregate approach (ρ rater 1 = 0.47 and ρ rater 2 = 0.29). In addition, the group average network representations of low- and high-performing students were reasonable and straightforward to interpret, the high group was more like the expert, and the low and high groups were more similar to each other than to the expert.

In a follow-up investigation, Clariana, Wallace, and Godshalk (2008) considered the effects of anaphoric referents on *ALA-Reader* text processing. Participants in an undergraduate business course ($N = 45$) again completed an essay as part of the course final examination. The investigators edited these essays to replace the most common pronouns "their", "it", and "they" with the appropriate referent. The original unedited and the edited essays for the top- and bottom-performing groups were processed with *ALA-Reader* using both approaches, sentence and linear aggregate. These data were then analyzed using Pathfinder analysis. The network average group representation similarity values comparing the original to the edited essays were large (i.e., about 90% overlap), but the linear aggregate approach obtained larger values than the sentence aggregate approach. The linear approach also provided a better measure of individual essay scores, with a Pearson correlation $r = 0.74$ with the raters' composite score.

These studies show a moderate correlation between human rater essay scores and *ALA-Reader* scores. Note that the essays in the first two studies used mostly technical biology vocabulary while essays in the second two used fairly general vocabulary that included a number of synonyms for key terms, such as manager, supervisor, and boss for the key term "management". *ALA-Reader* may be more appropriate for some types of essays and may be inappropriate for many types of essays. In these few studies, the more technical or specific the vocabulary in the essays, the better *ALA-Reader* performed. In addition, the first two studies used the sentence aggregate approach only and obtained an adequate measure of essay performance, while in the third and fourth study, the linear aggregate approach provided a satisfactory measure of essay performance but the sentence aggregate approach did not. The linear approach appears to be better than the sentence approach, and this may relate to both the nature of structural knowledge and the forced linearity of expository text. As with *ALA-Mapper*, the evidence is persuading us that *ALA-Reader* is not really an essay-scoring tool, but rather it is a tool to measure knowledge structure and this measure of knowledge structure happens to correlate with various kinds of essay scores.

7.5 Next Steps

This chapter described two related software programs that were designed to complement Pathfinder analysis, *ALA-Mapper* for processing graphs and *ALA-Reader* for processing text. The findings from several investigations were presented that indicate that these software tools may be measuring participants' knowledge structure. These two tools show potential but there are several critical issues yet to be resolved regarding these approaches.

A critical area for further investigation is which key terms to use and how many should be used during *ALA-Mapper* and *ALA-Reader* analysis because some key terms appear to be far more important than others. Typically, the course instructor or another content expert selects the key terms for the analysis phase. But further research must establish the best approach for determining these key terms. Contrary to expectations that using more terms means improved concurrent validity (see Goldsmith et al., 1991), Clariana and Taricani (2010) used *ALA-Mapper* to score distance data from a set of 24 open-ended concept maps using either 16, 26 (those 16 + 10 more), or 36 (those 26 + 10 more) most important terms (as selected and prioritized by a content expert). The greatest correlations with the multiple choice terminology and comprehension posttests were observed for 16 terms, then 26, then 36. Increasing the number of terms used to score the concept maps did not increase the predictive ability of the scores, probably due to students not selecting enough of the most important words to include in their concept maps.

Another area for further research involves the effects of providing participants with the key terms during concept mapping or when writing their essay. In open-ended concept mapping, students are typically given a blank page and a prompt,

while closed mapping often also includes a list of terms, sometimes a list of linking phrases, and even in some cases a partially completed map with blank boxes for missing terms. The different approaches involve different cognitive activities (e.g., levels of generativity; Lim, Lee, & Grabowski, 2008). A students' ability to recall the important terms is a critical task in open-ended concept mapping. Probably this generation task should be separated from the actual map formation task by asking students to first list all terms that they would like to include in their map, and then in a second activity, provide a list of researcher-selected terms for the students to use during actual mapping. This two stage approach would maintain some of the power of open-ended mapping (the gold standard) related to understanding the important concepts in a domain question, while also requiring a full range of interaction with the concepts during the second stage.

Another critical area for further research is the setup and prompt used for eliciting a concept map or an essay. Internal and external context factors strongly influence the kind of information elicited during concept mapping. For example, training participants to create hierarchical concept maps, whether the domain organization is hierarchical or not, must alter the obtained knowledge structure improperly toward hierarchical relationships. In a series of experiments, Derbentseva, Safayeni, and Canas (2007) showed that simply requiring participants to draw cyclic concept maps where clusters of four terms were connected in a circle with each leading to the next compared to tree (hierarchical) concept mapping resulted even in fundamentally different propositions. On average, 45% of the linking phrases between terms in the cyclic maps were dynamic phrases compared to only 14% of the linking phrases in the tree maps. Network diagrams contain both associations (distances) and propositions (links), but a strong focus on either one by pre-training, the drawing prompt, or other context factors increases the information content of that aspect but at the expense of the other aspect. Many concept map investigations demand a strong emphasis on propositional correctness and the focus is so great that the distances between terms no longer have psychological meaning. In any case, strong context factors likely devastate the relationship between the artifact obtained and the participant's actual knowledge structure.

Similarly, context variables that influence essays should be more closely examined to determine if context factors, such as the essay writing prompt, providing a list of terms, or the essay genre, can be manipulated to obtain essays that better capture the students' structural knowledge. For example, compare–contrast type essay questions or other writing prompts which ask students to discuss the relationships between concepts may be most appropriate for eliciting knowledge structure (Goldsmith et al., 1991; Gonzalvo et al., 1994). Further research should consider what conditions best elicit essays that reflect student's knowledge structure.

The referent used for comparison analysis also requires considerable thought and further research. During the analysis phases, the referent data set and *PFNET* that is used as the baseline or standard to compare to the participants' *PFNETs* is critical because error, idiosyncrasies, or spurious links in the referent *PFNET* produce error in every comparison. Referents should be carefully crafted. When expert's concept maps or essays are used as the referent, probably several should

be used since different experts may have different but correct representations of the domain question. At any rate, the approach used by investigators to obtain or create a referent must be carefully described, and if possible, the *PFNET* representation of that referent should then be inspected for under specification and for errors.

Also, there are more than two approaches (i.e., linear and sentence) for translating essays into arrays. Lambiotte et al. (1989, p. 342) proposed a taxonomy of "map devices" based on the signaling device used to represent relationships among ideas: Spatially based, node-based, link label based, and hybrid. The distance between terms in concept maps appears to be important information related to inference and comprehension (Cernusca, 2007, pp. 138–139; Clariana & Poindexter, 2003; Poindexter & Clariana, 2006; Taricani & Clariana, 2006), and so not only in network diagrams but also the distances between key terms in a text passage may also be important information. A feature will be added to *ALA-Reader* to capture these linear distances between key terms in text as a proximity array in order to consider this notion.

In summary, the history of science has shown that new observation tools lead to different ways to conceptualize phenomenon, and this leads to new and more powerful theories. The software tools described in this chapter and in this volume show considerable promise for the systematic analysis of knowledge.

References

Cernusca, D. (2007). *A design-based research approach to the implementation and examination of a cognitive flexibility hypertext*. Unpublished doctoral dissertation, University of Missouri, Columbia. Downloaded November 11, 2008, from http://edt.missouri.edu/Summer2007/Dissertation/CernuscaD-071907-D8036/research.pdf

Clariana, R. B. (2002). *ALA-Mapper software*, version 1.01. Retrieved September 28, 2008, from http://www.personal.psu.edu/rbc4/ala.htm

Clariana, R. B. (2004). *ALA-Reader software*, version 1.01. Retrieved December 24, 2008, from http://www.personal.psu.edu/rbc4/score.htm

Clariana, R. B., & Koul, R. (2004). A computer-based approach for translating text into concept map-like representations. In A. J.Canas, J. D.Novak, & F. M.Gonzales (Eds.), *Concept maps: Theory, methodology, technology, vol. 2*, in the Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain, Sep. 14–17, pp. 131–134. See http://cmc.ihmc.us/papers/cmc2004-045.pdf.

Clariana, R. B., & Koul, R. (2008). The effects of learner prior knowledge when creating concept maps from a text passage. *International Journal of Instructional Media, 35*, 229–236

Clariana, R. B., Koul, R., & Salehi, R. (2006). The criterion-related validity of a computer-based approach for scoring concept maps. *International Journal of Instructional Media, 33*, 317–325.

Clariana, R. B., & Poindexter, M. T. (2003). *The influence of relational and proposition-specific processing on structural knowledge.* A paper presented at the Annual Meeting of American Educational Research Association (AERA), San Diego, CA, April 2003.

Clariana, R. B., & Taricani, E. M. (2010). The consequences of increasing the number of terms used to score open-ended concept maps. *International Journal of Instructional Media, 37*(2), 218–226.

Clariana, R. B., & Wallace, P. E. (2007). A computer-based approach for deriving and measuring individual and team knowledge structure from essay questions. *Journal of Educational Computing Research, 37*, 209–225.

Clariana, R. B., Wallace, P. E., & Godshalk, V. M. (2008). Deriving and measuring group knowledge structure via computer-based analysis of essay questions: The effects of controlling anaphoric reference. In D. G. Kinshuk, J. M. Sampson, P. Spector, D. Isaías, & D. Ifenthaler (Eds.), *Proceedings of the IADIS international conference on cognition and exploratory learning in the digital age* (pp. 88–95). Freiburg, Germany: International Association for Development of the Information Society.

Dearholt, D. W., & Schvaneveldt, R. W. (1990). Properties of pathfinder networks. In Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 1–30). Norwood, NJ: Ablex Publishing Corporation.

Derbentseva, N., Safayeni, F., & Canas, A. J. (2007). Concept maps: Experiments on dynamic thinking. *Journal of Research in Science Teaching, 44*, 448–465.

Einstein, G. O., McDaniel, M. A., Bowers, C. A., & Stevens, D. T. (1984). Memory for prose: The influence of relational and proposition-specific processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 133–143.

Goldsmith, T. E., & Davenport, D. M. (1990). Assessing structural similarity of graphs. In Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 75–87). Norwood, NJ: Ablex Publishing.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing knowledge structure. *Journal of Educational Psychology, 83*, 88–96.

Gonzalvo, P., Canas, J. J., & Bajo, M. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology, 86*, 601–616.

Harper, M. E., Hoeft, R. M., Evans, A. W. III, & Jentsch, F. G. (2004). Scoring concepts maps: Can a practical method of scoring concept maps be used to assess trainee's knowledge structures? *Factors and Ergonomics Society Annual Meeting Proceedings, 48*, 2599–2603.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kinchin, I. M., & Hay, D. B. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research, 42*, 43–57.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillside, NJ: Erlbaum.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychology Review, 85*, 363–394.

KNOT (1998). *Knowledge Network and Orientation Tool for the Personal Computer*, version 4.3. Retrieved October 3, 2003 from http://interlinkinc.net/

Koul, R., Clariana, R. B., & Salehi, R. (2005). Comparing several human and computer-based methods for scoring concept maps and essays. *Journal of Educational Computing Research, 32*, 261–273.

Lambiotte, J. G., Dansereau, D. F., Cross, D. R., & Reynolds, S. B. (1989). Multirelational semantic maps. *Educational Psychology Review, 1,* 331–367.

Lim, K. Y., Lee, H. W., & Grabowski, B. (2008). Does concept-mapping strategy work for everyone? The levels of generativity and learners' self-regulated learning skills. *British Journal of Educational Technology, 40*, 606–618.

Lomask, M., Baron, J. B., Greig, J., & Harrison, C. (1992, March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science*. Paper presented at the annual meeting of the National Association of Research in Science Teaching, in Cambridge, MA.

Poindexter, M. T., & Clariana, R. B. (2006). The influence of relational and proposition-specific processing on structural knowledge and traditional learning outcomes. *International Journal of Instructional Media, 33*, 177–184.

Schvaneveldt, R. (1990). *Pathfinder associative networks: studies in knowledge organization*. Norwood, NJ: Ablex Publishing.

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development, 54*, 61–78.

Chapter 8
A Self-Organising Systems Approach to History-Enriched Digital Objects

Andrew F. Chiarella and Susanne P. Lajoie

Authors augment their texts using devices such as bold and italic typeface to signal important information to the reader. These typographical text signals are an example of a signal designed to have some effect on others. However, some signals emerge through the unplanned, indirect, and collective efforts of a group of individuals. Walking paths emerge in parks without having been designed by anyone. Objects, such as books, accumulate wear patterns that signal how others have interacted with the object. Books open to important, well-studied pages because the spine has worn in that location (Hill, Hollan, Wroblewski, & McCandless, 1992). Digital text and the large-scale collaboration made possible through the Internet provide an opportunity to examine how unplanned, collaborative signals could emerge in a text. CoREAD, a social software application, was designed using a self-organising systems perspective to enable indirect, collaborative text signalling. A brief description of self-organising systems, social software, and text signalling follows. The current work will also be situated within the research on history-enriched digital objects (Hill & Hollan, 1993; Hollan, Hutchins, & Kirsh, 2000). Finally, a study of undergraduate students using CoREAD will be presented.

8.1 Self-Organising Systems

A classic example of a self-organising system is a social insect colony, ant colonies for example (Bonabeau, Dorigo, & Theraulaz, 1999). In such collectives, rather simple individuals are able to solve difficult problems collectively (such as food foraging) in a decentralised fashion, that is, without a leader or a plan. Trails to nearby, rich food sources emerge from initially random individual searches of the environment. These random searches become organised as the ants respond to pheromones they leave in the environment, which creates a kind of positive feedback loop. A dominant pheromone trail emerges as more and more ants find the food source and

A.F. Chiarella (✉)
Athabasca University, Athabasca, AB, Canada
e-mail: andrewc@athabascau.ca

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_8,

return to the nest. This chemical trail is similar to trails made by people through parks since in both cases future activity becomes organised by past activity and the resulting trail was not designed, but emerged through decentralised activity.

The interaction of the ants in the colony is a form of stigmergy, or stigmergic interaction. Stigmergy originally referred to products of work left in the environment that influenced future work (Grasse, 1959). An example is the building behaviour of termites whereby termites are influenced by current structures to add to those structures. The result is the creation and maintenance of the termite colony's nest.

> La coordination des tâches, la régulation des constructions ne dépenendent pas directement des ouvriers, mais des constructions elles-mêmes. *L'ouvrier ne dirige pas son travail, il est guidé par lui*. C'est à cette stimulation d'un type particulier que nous donnons le nom de STIGMERGIE (*stigma*, piqûre; *ergon*, travail, oeuvre = oeuvre stimulante). (Grasse, 1959, p. 65, emphasis in the original)

When translated this reads,

> The coordination of tasks and the regulation of constructions does not depend directly on the workers, but on the constructions themselves. The worker does not direct his work, but is guided by it. It is to this special form of stimulation that we give the name STIGMERGY (stigma, goad; ergon, work, product of labour = stimulating product of labour). (Holland & Melhuish, 1999, p. 1)

The term stigmergy[1] has also been extended to include signs left in the environment that are not considered products of work. The pheromones that ants deposit in the environment are a classic example. These chemicals help organise the ants' behaviour but are not themselves part of the task, such as foraging for food.

The key to any stigmergic sign is that it is present in the *local* environment (physical or symbolic) in which the agents "work". Agents only have access to this local information and do not have access to the entirety of the group's efforts. Also, the signs in the environment are readily perceived and interpreted by the agents. A short summary of the key features of self-organising systems – with respect to this study – is presented below:

1. There are many individual agents acting, often simultaneously,[2] using rather simple rules.
2. The agents act on the local information in their environment.
3. The environment plays a role by storing information.
4. The collection of agents is able to explore multiple solution paths, often through "random" exploration by individual agents.
5. Positive feedback amplifies actions such that more and more agents engage in the same action, negative feedback counterbalances this tendency.

---

[1] Marsh and Onof (2008) have also suggested the term stigmergic cognition to refer to indirect communication mediated by modifications to the environment – a form of extended mind (Clark, 2001; Clark & Chalmers, 1998).

[2] Simultaneous, or synchronous, interaction is the norm in natural systems like ant colonies. However, sequenced, or asynchronous, interaction is also possible.

Ultimately, global-level phenomena emerge from these local interactions – the agents' collective behaviour becomes organised although there was no guiding plan or leader coordinating their behaviour. This is nicely summarised by Sulis in the following quote.

> A collective intelligence consists of a large number of quasi-independent, stochastic agents, interacting locally both among themselves as well as with an active environment, in the absence of hierarchical organization, and yet which is capable of adaptive behavior. (Sulis, 1997, p. 35)

Self-organising systems typically require fairly large groups of diverse individuals. Bringing larger groups of people together efficiently has often been difficult given communication and organisation costs. In educational settings, for example, groups are often limited to three to five students. As another example, online discussion forums do not scale well as the number of participants and postings increases. Eventually too many posts and too many threads accumulate. Extracting anything useful from the forum becomes a challenge in itself.[3] However, new forms of internet-based software permit vast numbers of people to share their knowledge, opinions, preferences and skills *far more efficiently*. These social software tools open up a whole new set of possibilities for collaborative work modelled on self-organising systems.

8.2 Social Software

> A few perfumes and pheromones aside, we humans seem noticeably lacking in native trail laying skills. Here the contemporary cyborg has a distinct edge, for she is already an electronically tagged agent. . . .we can automatically lay electronic trials, which can be tracked, analyzed, and agglomerated with those laid by others. (Clark, 2003, p. 145)

The electronic trails to which Andy Clark refers in the above quote is a large part of what makes social software possible. What others have done becomes known to current users in the form of stigmergic signals left in the environment. Each user benefits from the experience of the others.

In general, social software supports direct and indirect communication among a group of users. The focus of this discussion will be on indirect forms of communication, since they are typically asynchronous and stigmergic in nature. Much of the social software available is online, browser-based, and freely available to anyone. As such, the groups that result can often number in the thousands. This creates a form of collaboration that is quantitatively and qualitatively different from traditional software designed for computer supported collaborative work/learning (CSCW/CSCL). Whereas these traditional forms of collaborative software support small, defined groups, social software typically supports what Dron and Anderson (2007) refer to as networks and collectives.

[3] Though this can be minimised with good search tools.

One common type of social software is collaborative bookmarking and tagging systems.[4] These features often go together, typically a website is bookmarked (and added to a database) and the user may then tag it.[5] Tagging involves categorising the website using any number of keywords entered by the user. These tags may also be selected from a list of tags that the user has already used, or a set of tags created by the community of users. Since the set of community tags is often enormous only the more popular tags are typically displayed.

The community tags evolve in a decentralised fashion; there are no rules or screeners vetting the tags. These tags can also be used for browsing websites since the tags function as hyperlinks to a list of websites so tagged. Browsing by tags is like using keywords in a search, with the added benefit that the user knows that there are at least some pages associated with the tag. The tag also represents some user's (or many users) interpretation of the website, not just a simple keyword text match. Tag clouds provide additional information to the user about the popularity of the tag by showing popular tags in larger font sizes.

Some social software has been specifically designed for educational purposes, with the goal of generating educationally useful metadata. This developing field of research is described next.

8.2.1 Social Software for Education

> This paper explores an alternative approach to the use of computers in education, where machines are not in control nor are they the tools of teachers, but instead amplify and embody the combined intelligence of the learners who use them. In such systems the machines knit together with their users to form a landscape, allowing emergent behaviours based on the values and knowledge of the communities that inhabit them to shape them. This is possible because computers occupy a unique position as the tools and the medium as well as the environment in which interactions between people occur. (Dron, 2007b, p. 201)

Several scholars have begun developing social software for education (Bateman, Brooks, McCalla, & Brusilovsky, 2007; Brooks, Hansen, & Greer, 2006; Dron, 2007a, 2007b; Dron, Boyne, & Mitchell, 2001; Dron, Boyne, Mitchell, & Siviter, 2000; Farzan & Brusilovsky, 2005, 2006; Koper, 2004; Recker, Walker, & Lawless, 2003; Tattersall et al., 2005; Vassileva et al., 1999). In some cases, scholars have very explicitly designed their software using a complex or self-organising systems framework (Dron, 2007b; Dron et al., 2001, 2000; Koper, 2004; Tattersall et al., 2005).

These approaches share a few things in common. First, the systems are designed to allow many students to access and amend, add to, or augment learning resources. The information students provide can be collected passively (e.g. time students

[4] Since most social software supports multiple functions, bookmarking and tagging might be better considered key functions of many social software systems.

[5] Any digital object, such as photographs and videos, may also be bookmarked and tagged.

spend using a resource) or actively (e.g. asking the students to provide a rating of the resource). Second, the learning resources themselves or information about the resources – often called metadata – are in some way modified over time. This is based on the learners' use of the resources and any information they added or attached to the resources. The metadata may indicate the target audience, difficulty, uses, or type of content – review, study, theoretical piece, etc. – of a learning resource. Metadata may be indicated using text describing the learning resource or by signs and symbols in the software environment (e.g. icons adjacent to name of the learning resource). Finally, as a direct result of the two features above, the resulting learning resources and metadata are not completely pre-designed by instructors or designers. Based on how many learners use a learning resource and what they do with it – tagging, annotations and note taking, etc. – stigmergic signs are attached to the resource. Learners affect the environment through their actions but the environment, in turn, influences these actions by presenting learners with the community's current opinion. The form that the learning resources take emerges over time as a result.

Most of the social software for education has been designed primarily to assist learners *locate* useful learning resources given their current goals. The resources generally remain the same (e.g. the texts themselves are not modified) but their associated metadata change over time. Traditionally, metadata have been provided by the authors of documents. This is very similar to the use of text signals by writers. In both cases, the producers of the text or object provide information to consumers that is designed to guide or assist them. However, text signals are provided within the text itself, whereas metadata are external to the document or object. As such, text signals assist with processing the text. A brief overview of text signalling is presented next.

8.3 Text Signalling

There is a fairly extensive body of research on text signalling, with much of the more recent work by Lorch and his colleagues (Lemarie, Lorch, Eyrolle, & Virbel, 2008; Lorch, 1989; Lorch & Lorch, 1996; Lorch, Lorch, & Klusewitz, 1995; Lorch, Lorch, Ritchey, McGovern, & Coleman, 2001; Mautone & Mayer, 2001; Meyer & Poon, 2001; Naumann, Richter, Flender, Christmann, & Groeben, 2007). Text signals are writing devices that emphasise parts of the text or indicate the structure of the text (Lorch, 1989, p. 209). They may include a variety of devices: previews, overviews, summaries, titles, subheadings, typographical cues indicating importance (e.g. **bold**, *italics*), and even phrases that emphasise content or explicitly describe structure (e.g. "in summary", "first of four parts"). More recently, Lemarie et al. (2008) have proposed a new model, and definition, of text signalling whereby signalling is a text act designed by the author to have a desired effect on the reader.

> For example, typographical contrast expresses the author's intention to emphasize particular text content. As another example, a system of headings communicates the author's

> organization of a text. Thus, as illocutory acts, signals may be viewed as the realization in a printed text of an author's instructions regarding how the text is structured and how emphasis is to be distributed across the text content. Finally, an instruction/signal by the author may be heeded by the reader with the result that the reader's processing may be influenced. (Lemarie et al., 2008, p. 32)

Most studies have shown that text signals produce better performance on the recall of the signalled content. A comprehensive review by Lorch (1989) found memory was improved for signalled content for a variety of text-signalling devices (p. 229).

Whereas text signalling is an act by the author designed to affect the reader, readers themselves annotate texts in ways "designed" to affect their current or future processing of the text. Additionally, if such annotations and marginalia (i.e. notes, highlighting or underlining, emphasis marks like stars, etc.) could be shared with others, a new form of text signalling, or shared annotations (Marshall, 1998; Marshall & Brush, 2004), might be possible. One where the illucutory acts are from one reader, or set of readers, to yet other readers: processing suggestions "by readers for readers". Since digital texts permit annotations to be easily shared, they afford social interaction that is not normally possible with hardcopy texts (Marshall, 2005; Marshall & Brush, 2004).

As previously described, many types of social software derive their benefits from a self-organised process, that is, a large group of people creates useful "knowledge" through stigmergic signals left in a digital environment. The "history-enriched digital objects" research of the early 1990s (Hill & Hollan, 1993; Hill et al., 1992) can now be seen as a form of social software that supports the creation of text signals, and is discussed in more detail in the next section.

8.4 History-Enriched Digital Objects

> The physics of the world is such that at times the histories of use are perceptually available to us in ways that support the tasks we are doing. *While we can mimic these mechanisms in interface objects, of potentially greater value is exploiting computation to develop new history of interaction mechanisms that dynamically change to reflect the requirements of different tasks*. . .. Digital objects can encode information about their history of use. By recording the interaction events associated with use of digital objects . . . it becomes possible to display graphical abstractions of the accrued histories as parts of the objects themselves. (Hollan et al., 2000, p. 187, emphasis added)

The work by Hill and Hollan on history-enriched digital objects was based on the metaphor of physical wear (Hill & Hollan, 1993; Hill et al., 1992; Hollan et al., 2000). For example, as a book is used, it tends to accumulate wear on important pages. These pages become dog-eared, and the book will tend to open at such pages because of the wear to the book's spine. The now familiar example of a footpath through a park is another example of wear.

Hill and Hollan designed software that would signal to the reader of a document the extent to which other people had read or edited particular lines (Hill et al., 1992).

This was indicated using bars in a margin next to the text, for example. Thick bars indicated heavy reading and were based on the amount of time users spent reading that line (estimated from time spent on the overall "page" in a scrolling text). Other forms of digital wear were based on the edits made to the text (e.g. edits to software code). Similar visual indicators were used to signal to the current user which text section (lines) had been most heavily edited.

The "history-enriched digital objects" work (Hill & Hollan, 1993; Hill et al., 1992) was not guided by a self-organising systems perspective, nor did it explicitly refer to the text-signalling literature. However, it can be seen that the software enabled implicit, and indirect, collaboration that resulted in signs (i.e. text signals) being added to a text. The "use history" and the resulting visual signs were generated as by-products of the users' individual activities. Interaction with objects that are augmented with a visual representation of their history-of-use is, therefore, *stigmergic interaction*. The artifacts that result are, therefore, *stigmergic artifacts*.

8.5 Summary

The literature examined has led to the idea that social software could be used to enable readers to generate self-organised collaborative text signals. The readers themselves create the signals through their interaction with the text, augmented with stigmergic signs. Unlike authored text signals, these collaborative text signals are processing "suggestions", rather than "instructions", from the community of readers to individual readers. CoREAD, a software application for collaborative reading, was designed for this purpose and is described below.

8.6 The Design of CoREAD

The software designed was based on a self-organising system, and was modelled on the indirect interactions of ants that is supported by pheromone trails left in the environment (Bonabeau, 2002; Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000). Rather than adding typographical text signals in a planned, static fashion we ask if text signals could be added, *and modified dynamically*, by aggregating the decisions of many readers – here the signals emerge as a by-product of individual work. Using positive feedback to drive the system, learners affect the text through their actions but the text, in turn, influences these actions by presenting learners with the community's current opinion.

CoREAD is a software application (see Fig. 8.1) that was designed and programmed by the author for his doctoral research study (Chiarella & Lajoie, 2006, 2009). CoREAD presents a text in a page-by-page format (no scrolling) and provides the reader with a highlighting function. Font colour, a typographical text-signalling device, is used to signal sections of the text depending on the history of highlighting of the group thus far. As such, each reader potentially reads the text

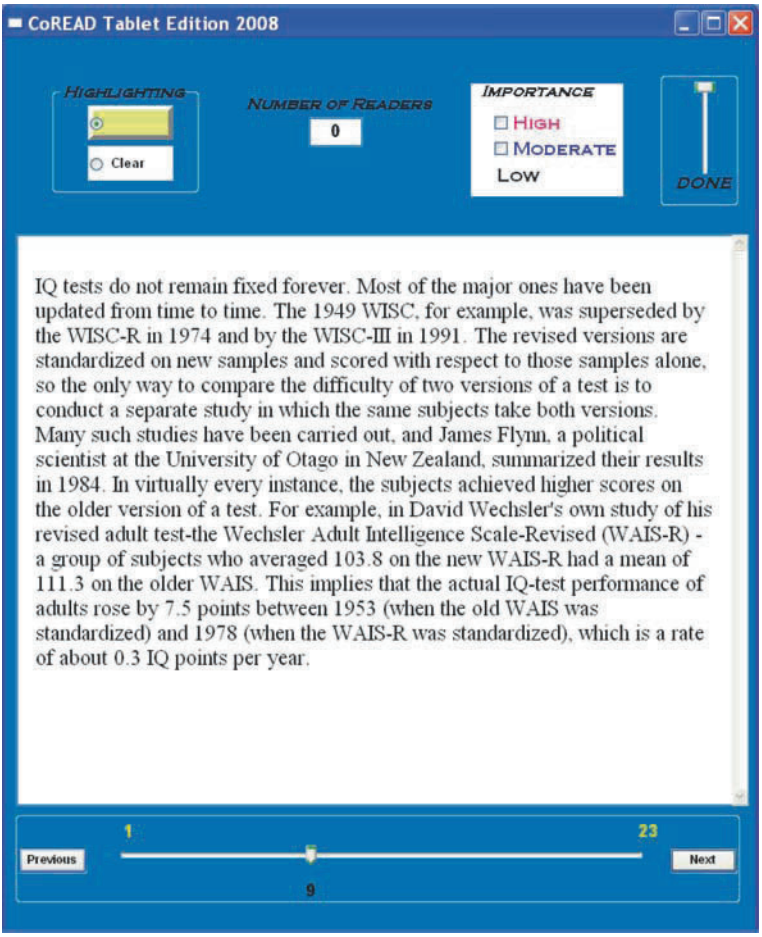


Fig. 8.1 Screenshot of CoREAD

with different parts of the text signalled. In general, red font indicates sections that have been highlighted by *many* readers, blue font for sections highlighted by *some* of the readers, and black font for when *few* to no readers highlighted the text. CoREAD was designed to foster a self-organising system of readers where:

1. There are many readers reading a text, making decisions about what sections are important and highlighting them.
2. Readers base their decisions on the text section they are currently reading and do not have any direct interactions with other readers.
3. The text is modified based on the readers' collective actions. The font colour of the text is altered to reflect the current collective opinion about the importance of each word (i.e. red for high importance, blue for moderate importance and black

for low importance). Technically, for each word an importance score is calculated based on a weighted history (to implement a form of negative feedback, see 6 below) of readers' actions (i.e. highlighted or not). This score, ranging from 0 to 1, is then converted into one of the text signal states above with the moderate category ranging from 0.30 to <0.70.

4. Because the readers have different prior knowledge and reading skills, they will each respond to the text differently. This diversity permits varied decisions about the importance of the text sections. In relation to self-organising systems, this heterogeneity is a source of randomness, and promotes the modification of the text signals through competing opinions (exploration).
5. By viewing the text signals, readers' decisions about the text may be influenced. If so, this is a form of positive feedback; readers are more likely to attend to and highlight sections currently favoured by the group. This is similar to the way that social insects follow pheromone trails of others that precede them (Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000). This positive feedback loop encourages the use of the text signals by other readers (exploitation).
6. To balance the positive feedback, CoREAD places more weight on the actions of recent readers – a type of negative feedback. This allows any earlier contributions that were not subsequently reinforced by other readers to fade away more quickly (i.e. reset to black font). This negative feedback returns words to a low-signal state unless they are consistently highlighted over time, and, in general, prevents the moderate and high-text signals from continually growing in number.

Using CoREAD, each learner reads a text, makes decisions about what sections are important and highlights them. Learners base their decisions on the text section they are currently reading and do not have any direct interactions with other learners. The typographical text signals are modified after each learner reads the text to reflect the current collective opinion about the importance of each word.

8.7 The Study

To examine the collaborative text signals that would emerge when using CoREAD, a study with 40 undergraduate students was conducted. However, the analysis presented here will examine the quality of the individual students' highlighting and written summaries using Latent Semantic Analysis rather than the collaborative text signals that emerged. The method used for the study is described next with the results presented thereafter.

8.8 Method

Forty undergraduate students participated in the study of which 73% had taken at least one psychology or cognitive science course. Ethics approval for this research was granted by the university's Research Ethics Board.

Although two texts from the field of psychology were read by the participants, only the results from the first text will be discussed in this chapter. This text is described in more detail below.

An article from American Scientist (Neisser, 1997) on the Flynn Effect (i.e. the rise in IQ scores over time) was adapted for the study.[6] The three introductory paragraphs, headings and any typographical signals were removed from the text. The title was replaced with "FE". Only two sections describing the possible causes of the Flynn Effect (length of schooling and exposure to visual media) were retained in the version used in the study. The text as used was 2,894 words, 126 sentences in length. The text was presented on 23 "pages" (screens) in CoREAD, each one a paragraph in length.

The participants read each text using CoREAD with the intention of writing a summary when they were done. As they read, they highlighted sections of text as they wished. Participants had complete freedom to move forward or backward and to go directly to any particular page of the text. Immediately after reading the text, they wrote a one-page summary (approximately 410 words maximum). The text was available to the participants while they wrote their summaries. Although the participant's highlighting was displayed, the collaborative text signals were removed from the text during the writing phase. Additionally, the summary was written on a separate computer and, therefore, they could not copy and paste from the text to their summaries. There was no time limit for either the reading or writing tasks. After writing the summary, participants completed a one-page demographic questionnaire that included questions about their prior knowledge of the content presented in the text.

8.9 Results

In order to compare the students' highlights and written summaries to the text and a summary written by the author himself (also modified for the study), Latent Semantic Analysis (LSA) was used (Landauer, Foltz, & Laham, 1998; Landauer, McNamara, Dennis, & Kintsch, 2007). This analysis can compute the semantic similarity between two pieces of text. It is sometimes described as a "bag of words" approach as it does not use word order, syntax, or even sentence boundaries in the analysis. Similarity scores, based on a cosine metric, may range between –1 and +1 (similar to a correlation coefficient, but computed differently), where +1 is maximum semantic similarity. An analysis was performed using the LSA software provided by the SALSA group at the University of Colorado at Boulder (http://lsa.colorado.edu). Hyphenated words often caused problems with the LSA analysis and so hyphens were removed for these analyses. As such, the number of words in the text is necessarily slightly larger for the LSA analyses. These analyses

[6]For a copy of the text please contact the first author.

were performed using the psychology semantic space (Myers, 1995) using all 400 factors and document to document comparison.

One participant was excluded from the analyses that follow because the number of words highlighted was an outlier which would have affected the correlations and regression analyses performed. As such the descriptive statistics, graphs and other analyses were performed with this participant removed ($N = 39$). The participant excluded was the 35th and had only highlighted 92 words or 3% of the text, whereas the mean was 840 words or 29% of the text. (Note: these statistics were calculated with this participant included in the sample.)

To examine the quality of the students' highlighting the words that a student highlighted were assembled into a "text". For the summaries, minor changes were made to correct spelling errors. As well, some forms of punctuation (e.g. hyphens) that cause LSA problems were removed. These two "texts" were then compared to the Flynn Effect text and the summary written by the author himself using LSA. Additionally, each student's own highlighting was compared to her own summary using LSA in order to measure the extent that the summary was a match to what the student had highlighted. Finally, the number of words highlighted and the length of the summary (in words) were used in some of the correlational analyses conducted. These were included since the length of a text is positively associated with the LSA score generated because more words increase the semantic coverage of the text in question. This allows the text to be more semantically similar to another text it is being compared to. As such, there were seven base variables in total (see Table 8.1).

Table 8.1 Variables

Variable	Label
Number of words highlighted	*Number of words highlighted*
Number of words in the summary	*Summary length*
LSA scores	
Participant's highlighting to participant's summary comparison	*Participant's highlighting to own summary*
Participant's highlighting to text comparison	*Highlighting to text*
Participant's summary to text comparison	*Summary to text*
Participant's highlighting to author's summary comparison	*Highlighting to author*
Participant's summary to author's summary comparison	*Summary to author*

8.9.1 Descriptive Statistics

The statistics for the seven variables and an eighth-derived variable are presented in Table 8.2 below. The derived variable is the difference between the LSA scores for the *Summary to text* and *Summary to author* comparisons. Positive values, therefore, indicate that the participant's summary was more semantically similar to the original text than the author's summary.

Table 8.2 Variables with descriptive statistics ($N = 39$)

Variable	Minimum	Maximum	Mean	Standard deviation
Number of words highlighted	371 (13%)	1,257 (43%)	859 (30%)	273
Summary length	197	460	373	67
Participant's highlighting to own summary	+0.63	+0.90	+0.80	0.066
Highlighting to text	+0.79	+0.92	+0.87	0.038
Summary to text	+0.63	+0.83	+0.75	0.050
Highlighting to author	+0.60	+0.71	+0.68	0.024
Summary to author	+0.57	+0.70	+0.64	0.028
Summary text–author difference	+0.02	+0.20	+0.11	0.046

8.9.2 The Author's Summary

The author's summary was only 225 words long and is reproduced below. It was derived from the first two of three introductory paragraphs that were removed from the original text when preparing the text for the study. The author's summary was a modified version of these two paragraphs.

In order to comprehend the results that follow, it is important to note the differences between the text and the author's summary. The text had four sections, though this was *not* indicated using headings or the like in the text as read by the participants. The first described intelligence tests in general including the mean and standard deviation of 100 and 15, respectively. In the next section, the Flynn Effect was described as a rise in (raw) test scores over time, that is, when samples of participants complete earlier IQ tests, the mean generated is typically above 100 showing that the current generation is more intelligent than the one for which the test was standardised in the past. The third section discussed whether the rise is real or not, and offered some possible causes for the effect. The last section examined two possible causes of the Flynn Effect in detail; years of schooling and exposure to visual media. The author's summary did not include any mention of the general information about intelligence tests and made references to the possible causes of the Flynn Effect without going into details.

8.9.2.1 The Author's Summary Reproduced

> Average scores on intelligence tests are rising substantially and consistently, all over the world. These gains have been going on for the better part of a century essentially ever since tests were invented. The rate of gain on standard broad spectrum IQ tests amounts to three IQ points per decade, and it is even higher on certain specialised measures. In the Netherlands, for example, all male 18-year olds take a test of abstract reasoning ability as part of

a military induction requirement. Because the same test is used every year, it is easy to see the mean score rising, in this case, at about seven points per decade. The cause of these enormous gains remains unknown. At this point, no one even knows whether they reflect genuine increases in intelligence or just the gradual spread of some specialised knack for taking tests. Greater sophistication about tests surely plays some role in the rise, but there are other possible contributing factors: better nutrition, more schooling, altered child-rearing practices and the technology driven changes of culture itself. Right now, none of these factors can be ruled out; all of them may be playing some part in the increasing scores. Whatever the causes may be, the sheer size of the gains forces us to reconsider many long-held assumptions about intelligence tests and what they measure.

8.9.3 Students' Highlights

8.9.3.1 Differences Between the Text and Author Comparisons

A paired *t*-test was performed comparing the *Highlighting to text* and *Highlighting to author* LSA scores. The LSA scores were statistically significantly higher ($t = 47.72$, $df = 38$, $p = 0.0001$) for the *Highlighting to text*. The mean difference was +0.19 with a standard deviation of 0.02. In fact, the LSA scores were higher for the *Highlighting to text* comparison for all participants (see Fig. 8.2 below). This indicates that the words the participants highlighted, when treated as a text in LSA, are more semantically similar to the overall text than the author's summary. Given that the highlighted words were selected from the text itself this outcome is not unexpected.

8.9.3.2 Trend over Time

Notice that there appears to be no trend over time – that is, over the participants – for either set of LSA scores (see Fig. 8.2 below). Therefore, there is no evidence here to support the hypothesis that the collaborative text signals would assist participants later in the sequence. If these collaborative signals were helpful, then the text segments highlighted later in the sequences would tend to be more semantically related to the text itself, the author's summary, or both.

8.9.3.3 Correlational Analyses

All of the correlations (Pearson) among the *Number of words highlighted*, and the LSA scores for the *Highlighting to text* and *Highlighting to author* comparisons were statistically significant ($p < 0.05$) (see Table 8.3 below). The *Number*

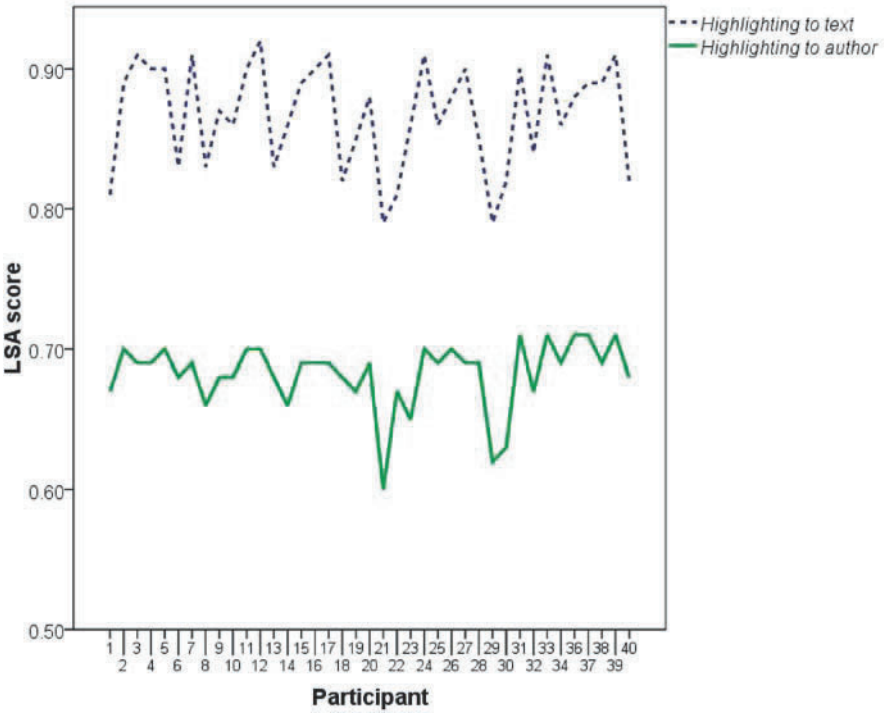


Fig. 8.2 LSA scores for highlighting to text and highlighting to author by participant

Table 8.3 Correlations among the number of words highlighted and the highlighting to text and author LSA comparisons

	Number of words highlighted	Highlighting to text	Highlighting to author
Number of words highlighted	1		
Highlighting to text	0.96*	1	
Highlighting to author	0.75*	0.77*	1

*Indicates statistically significant results, $p < 0.05$.

of words highlighted was strongly and positively related to the semantic similarity between the highlighting "texts" and (1) the Flynn Effect text and (2) the author's summary. When participants highlighted more of the text higher LSA scores for their *Highlighting to text* and *Highlighting to author* comparisons resulted. This is expected given the effect of word count on LSA scores – longer texts typically produce higher LSA scores – and the fact that the participants highlighted a mean of 30% of the overall text, which is quite high.

8.9.4 Students' Written Summaries

8.9.4.1 Differences Between the Text and Author Comparisons

A paired *t*-test was performed comparing the *Summary to text* and *Summary to author* LSA scores. The LSA scores were statistically significantly higher ($t = 14.95$, $df = 38$, $p = 0.0001$) for the *Summary to text* comparison. The mean difference was +0.11 with a standard deviation of 0.05. In fact, the LSA score was higher for the original text comparison for all participants (see Fig. 8.3 below). This is not surprising, as the participants may have attempted to summarise all of the topics in the text, whereas the author's summary focused on the Flynn Effect and its possible causes at a very general level. Such comparisons between the semantic similarity of students' summaries with the text and a model summary (e.g. the author's summary) could prove useful in educational contexts. It can be done efficiently, and potentially indicates whether the students are capturing only the essential ideas, as provided in a model summary, or including other non-essential ideas thereby creating summaries with a higher semantic similarity to the text.

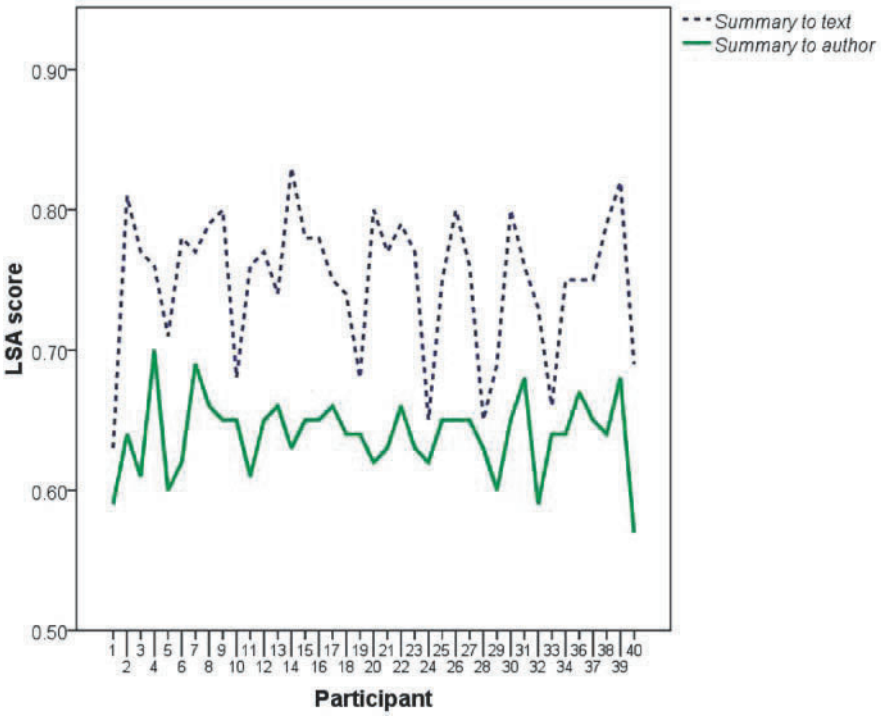


Fig. 8.3 LSA scores for summary to text and summary to author by participant

8.9.4.2 Trend over Time

Notice that there appears to be no trend over time – that is, over the participants – for either set of LSA scores (see Fig. 8.3 above). Therefore, there is no evidence here to support the hypothesis that the collaborative text signals would assist participants later in the sequence. If these collaborative signals were helpful, then the summaries written later in the sequences would tend to be more semantically related to the text itself, the author's summary, or both.

8.9.4.3 Correlational Analyses

Correlations among the relevant variables for the comparison of the participants' summaries are presented below in Table 8.4. As expected, *Summary length* was positively correlated with both the LSA scores for the *Summary to text* and *Summary to author* comparisons. Interestingly, the extent to which the participants' own summaries were semantically related to their highlights (i.e. *Participant's highlighting to own summary*) was also positively correlated to these LSA scores. This suggests that participants who wrote their summaries based on their highlighting produced better summaries than those who did not.

Table 8.4 Correlations among summary length, participant's highlighting to summary comparison, the summary to text and summary to author comparisons and the difference between them

	Summary length	Participant's highlighting to own summary	Summary to text	Summary to author	Summary text–author difference
Summary length	1				
Participant's highlighting to own summary	0.58*	1			
Summary to text	0.57*	0.89*	1		
Summary to author	0.44*	0.38*	0.42*	1	
Summary text–author difference	0.35*	0.74*	0.84*	–0.14	1

* Indicates statistically significant results, $p < 0.05$.

8.9.4.4 Multiple Regression Analyses

To examine whether *Participant's highlighting to own summary* was a unique contribution, multiple regression analyses were conducted. The *Summary length* and *Participant's highlighting to own summary* variables were entered separately as independent variables creating two models predicting *Summary to text* LSA scores and two models predicting the *Summary to author* LSA scores. *Summary length*

Table 8.5 Multiple regression analyses for summary to text and summary to author LSA scores

Dependent variable	Variables entered	R^2	Change in R^2	Change in F ($df1$, $df2$)	p value
Summary to text	1. Summary length	0.32	0.32	17.33 (1,37)	0.0001
	1. Summary length 2. Participant's highlighting to own summary	0.79	0.47	80.46 (1,36)	0.0001
Summary to author	1. Summary length	0.20	0.20	9.03 (1,37)	0.005
	1. Summary length 2. Participant's highlighting to own summary	0.22	0.02	0.97 (1,36)	0.33

was entered first to control for the effects of this variable, then *Participant's highlighting to own summary* was entered to determine if any remaining variance in the dependent variable could be accounted for by this variable. The results are presented in Table 8.5 above. These results indicate that for the *Summary to text* comparison there is a significant independent contribution of the *Participant's highlighting to own summary* after *Summary length* is controlled for. This is not the case for the *Summary to author* comparison. The extent to which the participant writes a summary based on her own highlighting tends to correlate positively with the strength of semantic relationship between the participant's summary and the text but not the author's summary. This might be indicative of a local importance approach whereby the text segments deemed important to the local context upon initial reading are brought together to generate a summary. As such, the resulting summary LSA scores are very high when compared to the text but much lower when compared to the author's summary. This also suggests that some participants must be more selective at the writing stage since they have lower *Participant's highlighting to own summary* scores (also see Table 8.2 for the range of scores obtained for this variable). These participants must include less of the highlighted information, perhaps only that which captures the most global aspects of the text's structure – the gist of the text.

8.9.5 Case Studies

In the following we present the best and worst summaries as determined by LSA analysis when the summaries are compared to both the Flynn Effect text and the author's summary of it.

8.9.5.1 Best Summary When Compared to the Flynn Effect Text – Participant #14

LSA score for *Summary to text* = +0.83
LSA score for *Summary to author* = +0.63
Participant's highlighting to own summary = +0.88
Summary text–author difference = 0.20
Summary Length = 417

> IQ tests typically include a wide variety of items due to the many different forms of mental ability. The degree to which any two tests measure something in common can be indexed by their correlation *r*, which ranges from –1 to +1. For example, a positive *r* means that individuals who score high on one test also tend to score high on the other. When a group of individuals takes a number of different tests, one can compute *r* for each pair of tests considered separately. The result is a correlation matrix that tends to consist of *r*'s that are all positive. Spearman is known for making the first formal factor analyses of such correlation matrices. He concluded that a single common factor that he termed "*g*", for general factor, accounted for the positive correlations among tests. He and his successors regard *g* as the real and perhaps genetically determined essence of intelligence. *G* is best measured by a test of visual reasoning called Raven's Progressive Matrices. An average IQ rests between the range of 85 and 115 and when defined in this way, reflects relative standing in an age group, not absolute achievement. Though a normal child becomes more intelligent with age, their IQ will rarely change much after age 5 or 6. IQ tests are updated over time and studies show that in virtually every instance subjects tested on both an older and newer version of an IQ test scored higher on the older version. On broad tests such as the WISC, Americans have gained about 3 IQ points per decade. This rise of increasing raw scores appearing on every major test is often called the Flynn Effect. The largest Flynn effects appear on highly *g* loaded tests – scores on tests that measure it best are going up at twice the rate of broad spectrum tests. Judging the American children of 1932 by today's standards would have put their IQ at about 80n only. Flynn concludes that the *g* loaded tests do not measure intelligence but only a minor sort of abstract problem-solving ability with little practical significance. Some hypotheses might be the increases in test taking sophistication, the impact of worldwide improvements and others that both support and conflict Flynn's argument. It is largely thought to be the exposure to many types of visual media that has caused the Flynn Effect. Due to the important generational shift, children exposed to these media have a higher developed skills when it comes to visual analysis compared to their elders.

An interesting thing to note about this summary is that it parallels the original text quite closely beginning with information about IQ tests in general rather than beginning with a statement about the Flynn Effect. Much of this information

about mean IQ and correlations among different tests are provided to give a reader additional background information in order to comprehend the point of the article, namely the Flynn Effect and two possible causes of it. This participant seems to have interpreted the summarisation task as one where the key points from all sections of the original text are reproduced. This appears to be an example of the *local importance approach*. Notice that this participant also has a high LSA score for the *Participant's highlighting to own summary* comparison (+0.88 whereas the mean was +0.80 and the maximum was +0.90). It seems clear that this participant used the text segments deemed important upon initial reading to construct the summary without making any further decisions about what to include based on the overall text. As such, the resulting summary scores are very high when compared to the text but much lower when compared to the author's summary. This is because the author's summary focused only on the Flynn Effect and its possible causes. This participant also generated the largest *Summary text–author difference* score for the reasons described above; the difference was 0.20 and approximately two standard deviations above the mean.

8.9.5.2 Worst Summary When Compared to the Flynn Effect Text – Participant #1

LSA score for *Summary to text* = +0.63
LSA score for *Summary to author* = +0.59
Participant's highlighting to own summary = +0.70
Summary text–author difference = 0.04
Summary Length = 266

Various types of tests may be helpful for differentiating different mental abilities. However, the relationship between such tests has had a long history of controversy. There have been many types of statistical scores proposed for describing the relationship that exists between different types of scores, but none really get full consensus. The *g* factor has been proposed to capture a general factor of intelligence but changes in various types of scores over time, such as on the Raven test, may question the *g* factor's credibility. The difference mostly lies in the type of testing; the Raven test focuses more on visual abilities, whereas tests such as the WAIS and the WISC focus more on crystallised abilities. It is this difference in the content of the tests that may explain larger increases in test scores over the last couple of decades. While the schooling system has not changed significantly, and while it is shown that NOT being in the school system affects such test scores of crystallised abilities, the stability of WAIS and WISC score may make sense. In comparison, important increases in visually based test have generated many various hypotheses. Overall, it may be the exposure to a richer visual surrounding

that may best explain these gains over time. All over the world, new generations have experienced a much richer visual world, with media, television, video games and perhaps more use of visual tools in the schooling system. All in all, it may make more sense to regard general intelligence as composed of various subtypes; this way, the fluctuation of tests scores over time would make more sense.

This summary generated low LSA scores for both *Summary to text* and *Summary to author* principally because it was short and did not clearly state the nature of the Flynn Effect. Although this summary was completed by the first participant, the only one that did not benefit from collaborative signals, the role that the absence of such signals had on her performance can not be inferred. It is quite possible that individual differences in relevant prior knowledge and summarisation skills played a large role. However, it will be important to examine the effects of collaborative text signals in a controlled manner in future studies to determine their impact.

8.9.5.3 Best Summary When Compared to the Author's Summary – Particpant #4

LSA score for *Summary to text* = +0.76
LSA score for *Summary to author* = +0.70
Participant's highlighting to own summary = +0.79
Summary text–author difference = 0.06
Summary Length = 358

A debate has arisen over whether or not we are smarter than our grandparents were. Based on the results of various tests geared towards measuring IQ the answer would appear to be yes for with every passing year the score has gone. However, we must examine the various factors which play a role in this rise. The Raven test has played a pivotal role in the analysis of the rise of worldwide test scores. With IQ tests the mean of each age group defines an IQ score of 100 with a general deviation of 15 IQ points. While a child becomes more intelligent with age, his/her score will remain relatively stable at 100. IQ tests have not remained the same and are periodically updated. The rise in raw scores has been named the Flynn Effect. Flynn concludes that the tests do measure intelligence but in fact measure a type of abstract problem solving which has undergone significant improvements. In general, Americans have gained 3 IQ points per decade but contrary to popular belief these gains were not observed in the domain of vocabulary, arithmetic as one would assume since children are now in school longer than their parents and grandparents

were. In fact, tests most closely linked to school content should the smallest gains. On the other hand, for individuals who were kept out of school for a prolonged period of time, their IQ scores decreased dramatically compared to those of their peers. Overall, schooling was deemed to have affected tests of content more so than tests of reasoning. Improvements may also be attributed to the increase in desire to perform well and as such there is no clear link to an overall increase in intelligence. Improvements to technology have also helped create a new form of intelligence thereby exposing today's children to media and stimuli, which were not present to their grandparents. This type of intelligence can be dubbed visual analysis. So while we may be smarter – or more knowledgeable than our grandparents when it comes to visual analysis, there is no clear data to say for certain that we are smarter than them in any other way.

This summary begins with a description of the Flynn Effect and quickly proceeds to referring to it by name and introducing the two possible causes described in the text. Aside from two sentences about IQ tests in general, the entire summary focuses on the Flynn Effect and the two possible causes. It is easy to see why the LSA score for the *Summary to author* comparison was high in this case. Also note that the LSA score for the *Summary to text* was +0.70 which was one standard deviation below the mean and, therefore, lower than average but not overly poor (i.e. the minimum score was +0.63).

8.9.5.4 Worst Summary When Compared to the Author's Summary – Participant #40

LSA score for *Summary to text* = +0.69
LSA score for *Summary to author* = +0.57
Participant's highlighting to own summary = +0.68
Summary text–author difference = 0.12
Summary Length = 197

Testing over the years has changed in many ways. Likewise, so have the results of these tests. Testing has existed in many forms; some verbal and some visual. The IQ test is used to test a variety of items. This is considered a broad spectrum test. The correlation *r* represents the degree to which any two tests measure some thing in common. A positive *r* means that those who score high on one test do the same on another. A negative *r* is the contrary. Charles Spearman concluded that a single common factor accounted for the positive correlation among tests. This is called the general factor or *g*. Raw

test scored may change over the years but the IQ does not change much over the years. Test results have changed throughout the twentieth century. There is much speculation as to why. This rise is known as the Flynn Effect. Flynn concludes that tests do not measure intelligence but the ability to take a test. Possible reasons for the change in test results: Changes in Schooling Cultural reasons Children staying in school longer Test have been performed to study if schooling is in fact the reason test scored are rising.

This summary is quite short thereby limiting the LSA scores possible. However, it is also unclear with respect to the Flynn Effect; a rise is mentioned without any explanation. The first half of the summary provides no mention of the Flynn Effect and only covers general information about IQ testing. This approach is similar to the one taken by participant #14 (best summary when compared to the text) who also produced a summary that provided a lot of tangential information that was meant only as background information but was not about the Flynn Effect per se.

8.10 Discussion

LSA seemed to provide a reasonable means of evaluating the participants' summaries. Some participants scored well when compared to the text but not when compared to the author's summary. This is an interesting outcome as it shows that LSA can be used to differentiate different kinds of performance by using different comparison texts. Using LSA to compare student work with multiple models has been an approach used by other researchers as well (Foltz, Laham, & Landauer, 1999). The outcome also suggests that the participants probably interpreted the task somewhat differently. Since summarisation may not have been explicitly taught to these participants, each may have developed a different understanding of the goals of summarisation and what ought to be included in a summary. Additionally, the task instructions may have inadvertently encouraged some participants to focus on their own highlighting since the highlighted text segments were visible in the text when the writing task took place.

8.10.1 Social Software for Assessment and Feedback

8.10.1.1 Assessment

Traditionally, assessment in educational contexts has been performed by teachers or by adaptive computer environments. Social software systems on their own do not lend themselves to this type of assessment or feedback. Rather, they aggregate and display the actions of many users, which is a form of external assessment if one chooses to compare one's knowledge with that of others. Certainly, more formal

assessment features could be incorporated into social software but these would not be considered social software features in and of themselves.

For this study, LSA was used to compare student work – their highlighting or their summaries – to the text itself or to the author's summary of the text. In the first case, assessment of the students was based on the actions they took within the social software system: what they judged to be important and then highlighted. The results from this study indicated that these actions seemed to capture the meaning of the overall text better than the author's summary. This finding is interpreted as suggesting that the students were highlighting the important ideas for understanding the text at the local level. These highlighted text segments would probably include details that would not form part of the macrostructure (Kintsch, 1990) of the text. In addition, a summary writing task was used to assess the students. This task was external to CoREAD itself though the students had access to the text with their highlighting intact as they wrote their summaries. The results from this study also indicated that the summaries seemed to capture the meaning of the overall text better than the author's summary. As well, the results indicated that some students wrote summaries that were closely matched to their own highlighted text segments. High *Participant's highlighting to own summary* scores, particularly in cases where a significant portion of the text was highlighted, indicates that these students had perhaps included too much of the lower levels of the text structure, instead of the macrostructure. Participants with this pattern seem to have highlighted a substantial portion of the text while reading and then included many of these text segments in their summaries. In order to avoid including too much of the microstructure or details of the text, the reader must engage in a second round of judgements about the importance of individual text segments contained in the text. When initially reading the text, the reader judges importance based on the local context and her current understanding of the text. Later, when writing the summary, the reader must re-evaluate those text segments that were highlighted to determine if they are important enough, given the global context, to be included in the summary. Some set of rules or strategies needs to be employed (for examples see Brown & Day, 1983) – dependent on the reader's prior knowledge and understanding of the text – to select out only some of the text segments initially thought important in the course of reading the text.

Comparison of student work – completed within CoREAD or during post-reading tasks – to model texts using LSA can reveal or at least suggest which strategies students may have used to complete their work. The data provided by social software like CoREAD can be analysed using LSA to provide general assessments of student work in a timely fashion. Educators may find such assessments useful as they are efficient and can be used to direct their attention to students who may require assistance and thoughtful feedback or guidance.

8.10.1.2 Feedback

Typically, following assessment performed by a teacher, the student is provided with some form of feedback. Generally, this feedback is intended to correct student errors or misconceptions or simply direct students to more promising lines of thinking. In

the case of social software, the focus is on providing each user with social feedback. As well, this feedback does *not* follow a formal assessment of the user's thinking or work products. Rather, information about how the other members of a community have acted in the past with respect to the digital object at hand is provided. These actions provide the current user with indirect evidence of how the community thinks about the digital object. For example, if many users have bookmarked a website, then it can be inferred that it is interesting to those users. Popular tags associated with the website indicate which semantic categories the website's content covers. The current user can compare his own thinking to this social, collective thinking and recognise differences or similarities. It is this comparison, performed by the student, that constitutes a kind of feedback. To the extent that it brings about positive outcomes one could consider social feedback as scaffolding (Collins, Brown, & Newman, 1989).

Whereas students might often place a high level of trust in feedback from their teachers or instructional systems as a matter of course, they may have more or less trust in social feedback. For example, in using CoREAD, the reader might be surprised that so many of her fellow readers had highlighted a text segment she deems common knowledge or perhaps even irrelevant to the major themes in the text. This feedback should prompt some sort of judgement or evaluation of the reader's initial response to the text segment. At this time, little is known about this process, if indeed it occurs. From the text-signalling literature (Lorch, 1989; Lorch et al., 1995; Mautone & Mayer, 2001) it seems plausible that readers would at least focus more attention on these collaboratively signalled text segments. But the more interesting question of how the reader compares and possibly amends her judgement in response to the group's judgement remains to be investigated.

8.10.2 Limitations and Future Work

8.10.2.1 Limitations of Highlighting

The highlighting activity is very amenable to the self-organising systems approach upon which CoREAD was developed. The history of actions is easy to maintain and aggregate in order to generate collaborative text signals. However, because simple highlighting is binary, it offers no possibility for determining how important the text segment was for the participant. This could be overcome by offering participants a range of highlighting options (e.g. different colours or colour saturations) associated with different levels of importance (e.g. important, very important, etc.). We did not implement such a system because it increases the complexity of the software for the user and we wanted software that would minimally affect the user as she read the text.

Additionally, because the highlighting activity took place only during the reading task, it was not possible to determine what each participant thought was most important for the summary as opposed to what was deemed important at the time of first reading. Perhaps the simplest solution would be to ask participants to highlight

the text a second time during the summary-writing task (using a new colour). This of course would have the unfortunate consequence of cueing participants to re-evaluate what they initially highlighted. Another approach would be to ask participants to evaluate the importance of certain text segments separately after completing the summary task.

8.10.2.2 No Trend over Time

The study was designed to investigate the formation of collaborative text signals and so each participant read the text with a new set of text signals that depended on the highlighting history of all of the past participants. As such, there was no means to control for the text signals observed by each participant, or individual differences with respect to prior knowledge, reading comprehension strategies, or summary-writing ability. Although the text was "enriched" with collaborative text signals, the limitations of LSA (e.g. the inability to assess the style and organisation of a summary) and individual differences probably limited our ability to see any benefits of these signals. So although no evidence exists from this study to support the hypothesis that collaborative text signals would be useful to future readers, this will need to be properly investigated with controls in place. We intend to design future studies where the collaborative text signals generated from this study are used with a new sample. The main comparison will be reading a text with collaborative signals presented versus no signals. Additionally, controls for prior knowledge, reading comprehension strategies, and summary-writing abilities will be used (e.g. within-group comparison using two texts). Together, these should permit us to investigate whether the collaborative signals provide any assistance to readers.

8.10.2.3 Effects of Text Length on LSA Scores

Since LSA scores are so highly related to the length of the texts compared – with longer texts tending to have higher LSA scores – some means of controlling the length of the summaries and the number of words highlighted would be beneficial. Limiting the length of the summaries to something closer to the length of the author's summary would be fairly simple. The task instructions should perhaps specify not only a maximum but also a minimum number of words for the summary. Limiting the number of words a participant could highlight would be more difficult since it would require that the software code to be modified.

8.11 Conclusion

This chapter has presented CoREAD, a social software application that supports indirect social interactions among a group of readers and a theoretical rationale of the software based on self-organising systems (Bonabeau et al., 1999; Bonabeau & Theraulaz, 2000; Goldstone & Janssen, 2005; Miller & Page, 2007) and history-enriched digital objects (Hill & Hollan, 1993; Hill et al., 1992; Hollan et al., 2000).

As well, a study was presented where Latent Semantic Analysis (LSA) was used to assess the performance of the participants on a summary-writing task. It was shown that LSA can be used compare the work of the participants to different models. The comparison of the participant's highlighting to her own summary using LSA proved to be a useful metric. This LSA score was positively related to the comparison of the participant's summary to the text itself once the length of the summary was controlled for. However, it was not related to the comparison of the participant's summary and the author's summary once the length of the summary was controlled for. This indicates that participants who wrote their summary in a manner that closely matched their highlighting captured the semantic content of the text overall quite well but not the author's summary. As such, the comparison of the participant's highlighting to her own summary provides an indirect measure of the strategy the participant used to write the summary. With high scores indicating that the participant retained much of what was judged important at the time of initial reading in the final summary.

Acknowledgment This work was conducted as part of the first author's doctoral research. It was funded in part by grants awarded to the second author, namely the Social Sciences and Humanities Research Council of Canada INE Fund and the James McGill Research Fund.

References

Bateman, S., Brooks, C., McCalla, G., & Brusilovsky, P. (2007). *Applying collaborative tagging to e-learning*. In Proceedings of the Workshop on Tagging and Metadata for Social Information Organization (16th International World Wide Web Conference).

Bonabeau, E. (2002). Agent-based modelling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(3), 7280–7287.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems*. New York: Oxford University Press.

Bonabeau, E., & Theraulaz, G. (2000, March). Swarm smarts. *Scientific American, 282*, 72–79.

Brooks, C., Hansen, C., & Greer, J. (2006). *Social awareness in the iHelp courses learning content management system.* Paper presented at the Workshop on the Social Navigation and Community based Adaptation Technologies.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Behavior, 22*, 1–14.

Chiarella, A. F., & Lajoie, S. P. (2006). Enabling the collective to assist the individual: CoREAD, a self-organising reading environment. In T. C. Reeves & S. F. Yamashita (Eds.), *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn)* (pp. 2753–2757). Chesapeake, VA: AACE.

Chiarella, A. F., & Lajoie, S. P. (2009, April). *Dynamically modifying text signals: A self-organising systems approach to collaboration.* Paper presented at the Annual meeting of the American Educational Research Association (AERA), San Diego, CA, USA.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind and Language, 16*(2), 121–145.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. New York: Oxford University Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*, 10–23.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In R. Glaser & L. B. Resnick (Eds.), *Knowing, learning,*

and instruction: Essays in honor of Robert Glaser (pp. 453–494). Hillsdale, NJ: L. Erlbaum Associates.

Dron, J. (2007a). Designing the undesignable: Social software and control. *Educational Technology & Society, 10*(3), 60–71.

Dron, J. (2007b). The teacher, the learner and the collective mind. *AI & Society, 21*(1), 200–216.

Dron, J., & Anderson, T. (2007). Collectives, networks and groups in social software for e-learning. In G. Richards (Ed.), *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*. Chesapeake, VA: AACE.

Dron, J., Boyne, C., & Mitchell, R. (2001). *Footpaths in the stuff swamp.* Paper presented at the WebNet 2001, Orlando, FL, USA.

Dron, J., Boyne, C., Mitchell, R., & Siviter, P. (2000). *CoFIND: Steps towards a self-organising learning environment.* Paper presented at the WebNet 2000. San Antonio, TX, USA.

Farzan, R., & Brusilovsky, P. (2005, July). *Social navigation support through annotation-based group modeling.* Paper presented at the International Conference on User Modeling (UM 2005), Edinburgh, UK.

Farzan, R., & Brusilovsky, P. (2006). AnnotatEd: A social navigation and annotation service for web-based educational resources. In T. C. Reeves & S. F. Yamashita (Eds.), *The World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn)* (pp. 2794–2802). Chesapeake, VA: AACE.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhance Learning, 1*(2), Retrieved December 6, 2008, from http://imej.wfu.edu/articles/1999/2002/2004/index.asp.

Goldstone, R. L., & Janssen, M. A. (2005). Computational models of collective behavior. *Trends in Cognitive Sciences, 9*(9), 424–430.

Grasse, P.-P. (1959). La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. la theorie de la stigmergie: Essai d'interpretation du comportement des termites constructeurs. *Insectes Sociaux, 6*(1), 41–80.

Hill, W., & Hollan, J. (1993). *History-enriched digital objects.* Paper presented at the Third Conference on Computers, Freedom, and Privacy, Burlingame, CA.

Hill, W., Hollan, J., Wroblewski, D., & McCandless, T. (1992). Edit wear and read wear. In *CHI '92: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3–9). Monterey, CA: ACM Press.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction, 7*(2), 174–196.

Holland, O., & Melhuish, C. (1999). Stimergy, self-organization, and sorting in collective robotics. *Artificial Life, 5*(2), 173–202.

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7*(3), 161–195.

Koper, R. (2004). Use of the Semantic Web to solve some basic problems in education. *Journal of Interactive Media in Education,* 1–23 (PDF version available online at www-jime.open.ac.uk/2004/2006).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2&3), 259–284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Lemarie, J., Lorch, R. F., Jr., Eyrolle, H., & Virbel, J. (2008). SARA: A text-based and reader-based theory of signaling. *Educational Psychologist, 43*(1), 27–48.

Lorch, R. F., Jr. (1989). Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review, 1*(3), 209–234.

Lorch, R. F., Jr., & Lorch, E. P. (1996). Effects of organizational signals on free recall of expository text. *Journal of Educational Psychology, 88*(1), 38–48.

Lorch, R. F., Jr., Lorch, E. P., & Klusewitz, M. (1995). Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology, 20*, 51–64.

Lorch, R. F., Jr., Lorch, E. P., Ritchey, K., McGovern, L., & Coleman, D. (2001). Effects of headings on text summarization. *Contemporary Educational Psychology, 26*, 171–191.

Marsh, L., & Onof, C. (2008). Stigmergic epistemology, stigmergic cognition. *Cognitive Systems Research, 9*, 136–149.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *Proceedings of Hypertext* (pp. 40–49). New York: ACM Press.

Marshall, C. C. (2005). Reading and interactivity in the digital library: Creating an experience that transcends paper. In D. Marcum & G. George (Eds.), *Digital library development: The view from Kanazawa* (pp. 124–145). Westport, CT: Libraries Unlimited.

Marshall, C. C., & Brush, B. A. J. (2004). Exploring the relationship between personal and public annotations. In *Proceedings of the Joint Conference on Digital Libraries* (pp. 349–357). New York: ACM Press.

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology, 93*(2), 377–389.

Meyer, B. J. F., & Poon, L. W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology, 93*(1), 141–159.

Miller, J. H., & Page, S. E. (2007). *Complex adaptive systems: An introduction to computational models of social life*. Princeton: Princeton University Press.

Myers, D. G. (1995). *Psychology 5th edition*. New York: Worth Publishers.

Naumann, J., Richter, T., Flender, J., Christmann, U., & Groeben, N. (2007). Signaling in expository hypertexts compensates for deficitis in reading skill. *Journal of Educational Psychology, 99*(4), 791–807.

Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist, 85*(5), 440–447.

Recker, M. M., Walker, A., & Lawless, K. (2003). What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education. *Instructional Science, 31*(4–5), 299–316.

Sulis, W. (1997). Fundamental concepts of collective intelligence. *Nonlinear Dynamics, Psychology, and Life Sciences, 1*(1), 35–53.

Tattersall, C., Manderveld, J., van den Berg, B., van Es, R., Janssen, J., & Koper, R. (2005). Self organising wayfinding support for lifelong learners. *Education and Information Technologies, 10*(1–2), 109–121.

Vassileva, J., Greer, J., McCalla, G., Deters, R., Zapata, D., Mudgal, C., et al. (1999). A multi-agent approach to the design of peer-help environments. *Proceedings of AIED*'99, 38–45.

Chapter 9
Performance Categories: Task-Diagnostic Techniques and Interfaces

Michael Yacci

9.1 Introduction

One meaning of *diagnosis* is the investigation of the cause of a condition. Diagnosis should therefore uncover root causes of issues – social, medical, educational, and physical systems can all be examined to find why they function or fail to function. Typically, diagnosis is used to examine the inner workings of dysfunctional systems, with the intended purpose of ultimately prescribing solutions to repair or improve the functioning of the system. Often, the inner working of a system is not directly observable and must be inferred by the presence or absence of tangible, measurable clues.

In the case of human performance, there are numerous potential causes or contributors to ineffective behavior, as elaborated in texts on needs assessment and performance analysis (Mager & Pipe, 1984; Romiszowski, 1981). While some of these causes remain external to the performer (such as lack of functioning equipment or misplaced incentives), in this chapter we concentrate on internal individual skills as inhibitors and facilitators of desired performance. We first examine the tasks that performers are involved in, followed by general approaches to diagnosing problems with the performance of these tasks. Finally, the chapter presents a brief sketch of possible computer-based interfaces and their uses with the task-diagnostic framework presented.

9.2 Tasks

Because of the variety of tasks that humans engage in, there are likely to be a variety of diagnostic approaches that work better or worse for these tasks; therefore we begin by examining the tasks themselves.

M. Yacci (✉)
Rochester Institute of Technology, Rochester, NY, USA
e-mail: may@it.rit.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_9,

9.2.1 Tasks, Outcomes, and Processing

There are numerous taxonomies of human performance that spell out differences between the outcomes of learned behavior (Bloom, 1956; Gagne & Briggs, 1974; Merrill, 1983; Horn, 1989, 2002). Most of these taxonomies follow a similar classification pattern in the cognitive domain, making a distinction between content that is "remembered or recalled" and skills that are "applied." Within the broad category of applied skills, some of these taxonomies make an assumption about "lower level" skills that suggests that mental performance applied to tasks is somewhat algorithmic, step-by-step, or rule-based while "higher level" skills involve unstructured problem solving (labeled *higher order rule using* from Gagne & Briggs, or the *find* level from Merrill, for example). Some taxonomies assume that skills are hierarchically structured with lower level skills being subsumed by higher level skills. This suggests that learning is cumulative in nature, with simpler prerequisite skills acting as components of more complex skills (Gagne, 1968). Additionally, most taxonomies acknowledge the existence of psychomotor skills.

Distinctions between mental processing in tasks are found in the cognitive psychology literature focusing on declarative and procedural knowledge – knowledge of *what* compared to knowledge of *how*; this distinction has neurological (Ullman, 2001) and psychological (Anderson, 1985; Gagne, 1985) support. While the theories that underlie these distinctions are beyond the scope of this chapter, there is ample support of the differentiation in mental processes as applied to different types of tasks. *Production rules* are often hypothesized as an adequate form of representation for procedural knowledge. In well-structured tasks, production rules can be clearly identified that support the task. However, in ill-structured problems, the exact mechanism for problem solving is often less clear.

Other differences in tasks and processing are derived from the literature surrounding creativity, innovation, and invention. For our purposes in this chapter, we use these terms synonymously, referring to activities involving the production of novel ideas, new artifacts, or novel uses of existing tools and technologies. The underlying mental processes as reported in the literature are not algorithmic or orderly. Instead, these various mental processes describe a flash of insight (Bennett, 1997) sometimes occurring from playing with the subject (Feynman, 1997) or a reformulation of the problem with analogies to other problems and solutions (Minsky, 1986).

9.2.2 Tasks at Work

Work tasks themselves are often studied in applied fields such as operations research or training and instructional design. These fields study the impact of humans within technical systems. Hopp, Iravani, and Yuen (2007) suggest that there has been much research on the logistics of manufacturing systems – systems that are well defined and procedural – but little industrial engineering research has been done on "professional work" that they refer to as *discretionary* work. Indeed, a distinction between

what are colloquially labeled "blue-collar" jobs and "white collar" jobs often seems to be based on the degree of initiative and autonomy that the worker needs to deal with situations that range from routine and replicable to highly unusual and unique. This distinction has similarly been noted for at least 40 years in the training field as researchers have worked to quantify the "difficulty" of work-based tasks (Annett & Duncan, 1967). The relative nature of task complexity is discussed by Dorner (1996) who notes that task complexity is mediated by automaticity: tasks that are well-learned require less cognitive processing, and may therefore be considered less complex to more experienced workers. Connell, Sheridan, and Gardner (2003) use the terms *tasks* and *situations* to make a similar distinction between "targeted assignments" that require specific skills (tasks) and work that requires "an orchestration of capacities" (situations) to deal with more complex, problem-related scenarios.

Accompanying many of these mental activities are corresponding and integrated psychomotor skills that follow a similar pattern, moving from well-practiced physical movements requiring little variation or adaptation to physical skills that require creative adaptation to complex or dynamic scenarios (Romiszowski, 1999; Ackerman & Cianciolo, 2000). Many of the examples in the remainder of this chapter describe tasks that combine cognitive and psychomotor behaviors; the continuum that is described next can be extended to tasks that are purely cognitive, that are purely psychomotor, or that have both behaviors integrated within a task.

9.2.3 A Continuum of Tasks

The pattern across these descriptions of mental operations and cognitive and psychomotor tasks suggest a continuum of applied mental processing and physical activities related to specific work environments that run from pure algorithmic application of known steps and rules within a well-defined problem space to a loose, mental circling and playing with a poorly defined problem. For clarity, we use the terms *prescribed* tasks and *discretionary* tasks to encompass the extremes of such a mental and physical behaviors. These labels are viewed in Fig. 9.1 as the end points of a continuum, as tasks may fall anywhere along the line.

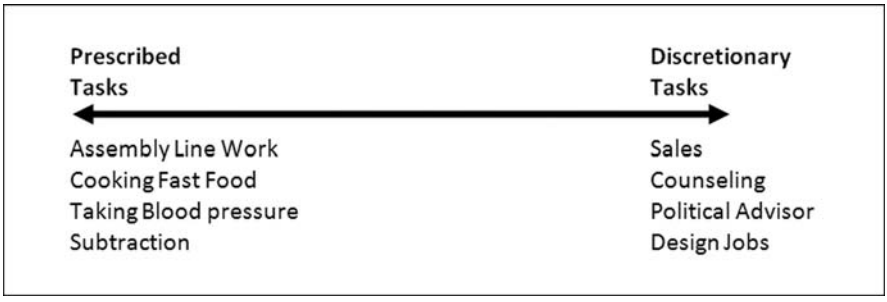


Fig. 9.1 A continuum between prescribed and discretionary tasks

Prescribed tasks, then, are well-defined, highly constrained activities that occur with limited variation in a predictable application domain. The presumed mental and physical activities that occur are well-rehearsed, algorithmic, and with limited generalization required. While there is some degree of rote memorization required, these are not tasks of recall for facts, but rather these tasks require extremely limited degrees of application. Examples in this category might be basic algorithms for subtracting two digit numbers, simple cooking, or taking a blood pressure from an "average" patient. Note that these tasks may vary in the degree of cognitive and psychomotor components. Prescribed tasks are usually well tested, coordinated, and documented, as in factory work or mathematical algorithms for routine operations.

Discretionary tasks involve ill-defined, complex problems with poorly described or open solutions, with almost no constraints on the way that these problems are solved. Note that this category is broader than problem solving because it also entails psychomotor activities as appropriate. Examples of tasks that are closer to the discretionary end of the continuum might be sales and interior design (that have discretionary cognitive aspects). A physical task that falls near the discretionary end of this continuum is extreme skiing (that requires reaction to dynamic, unpredictable physical terrain) as compared to bowling (that occurs in a standard, controlled environment and is more prescribed).

In extreme discretionary tasks there are no readily accepted procedures or processes to follow; a performer cannot be judged as to whether he or she is "doing it right" according to a standard performance algorithm. Instead, the quality of the solution itself (or a created artifact) is the sole determinant of task success. For example, a complex decision-making task for an organization may be judged by its long-term benefits to the system – *how* the decision was reached is not relevant. Or success in football is not judged by "style" but solely on the ability to move past defenders to score (which may require creating physical maneuvers on the fly to react to the competition).

The complete mental operations that underlie discretionary tasks are not fully known. Undoubtedly, these tasks require some degree of rule-using, but are probably not "straight" rule application situations. One might think of the mental operations as being creative in the sense that there are no standard approaches that directly solve the problem – they must be uniquely created or combined for the situation.

9.3 Diagnostic Environments

We are interested in diagnosis as it applies to the achievement of work tasks that fall across the continuum of prescribed tasks to discretionary tasks. Diagnosis in this sense supports successful *task completion at work* not skill improvement per se. This is an important distinction; we are interested in eliminating barriers to the successful completion of tasks and finding bugs in a mental or physical process, not in finding ways to improve an already successful solution; teaching is not the primary goal.

Nonetheless, diagnosis has been most clearly developed in instructional systems of many types.

In intelligent tutoring and computer-based training, diagnostic activities can be embedded into the fabric of instruction. Wenger (1987) provides numerous examples in which many types of diagnostic activities are included in both intelligent tutoring and frame-based computer-assisted instruction. In actual work situations, however, tasks are performed (a) using personal computers, (b) with no computing support, or (c) using specialized and dedicated computers and tools that are not easily modified. Additionally, (d) tasks with psychomotor components may be performed in open and unconstrained spaces. The solutions later described in this chapter are considerate of these varied task environments.

9.3.1 Task Success

To determine whether or not a performer has successfully accomplished a work task, we need to formulate a clear description of the qualities of a successful task performance. This description is fairly straightforward in prescriptive tasks but is more difficult in discretionary tasks.

In prescribed cases, accurately following a standard procedure is often all that is required. A task near the center of the continuum may require moderate judgment in the selection of standard procedures. Describing task success in prescribed tasks often means following a process or a decision-making algorithm; the procedure that workers follow is predesigned to achieve the outcome. Essentially, as in a beginning cookbook, if the steps are followed exactly, then an acceptable outcome is the result.

The outcomes of discretionary tasks are more difficulty to describe. By definition, standard procedures do not exist. In the most extreme discretionary cases, it cannot be unambiguously determined if the intended outcome was achieved, as the intended outcome may never be clearly stated. Dorner (1996) contrasts open-ended *positive* goals that have a clearly stated objective with *negative* goals that merely suggest what is *not* wanted in a solution, amongst other criteria that make goals more or less clear. However, even in the most extreme cases (“make our company more competitive” or “create a unique new product”) a loose goal is provided and one can provide some degree of evidence of having achieved or failed to achieve the goal. The practice of evaluation (Stake, 2004; Fitzpatrick, Sanders, & Worthen, 2004) often deals with these unclear intended outcomes and provides techniques to create defensible criteria. Art and design education often use a panel of trained judges to determine the quality of a product (Jeffries, 2007) in lieu of rigid criteria.

9.4 The Diagnostic Continuum

Given the broad range of skills and task situations that fall across the task continuum, and the different mental and physical activities that are needed to accomplish

these tasks, we conjecture that different diagnostic approaches would be more appropriate at various points on the continuum. We will continue to examine the points closest to the ends of the continuum, as these make for the clearest differentiation between cases.

9.4.1 Prescribed Task Diagnosis

Prescribed performance can be diagnosed with approaches based on hierarchical prerequisite decomposition or procedural decomposition. In hierarchical prerequisite decomposition, a skill is divided into sub-skills that are *absolutely essential* to successful performance (Gagne, 1968). In procedural decomposition, the tasks and decisions that a skilled performer uses to complete a task are enumerated and detailed.

The notion of hierarchical prerequisite decomposition (Gagne & Briggs, 1974) or prerequisite analysis (Smith & Ragan, 2005) is fundamental to most theories and models of instructional design. It begins with a clear description of a task or skill followed by an analysis of supporting prerequisite skills that must be present before the task could successfully be accomplished. Creating a valid prerequisite hierarchy is difficult in common instructional design because it is often based on conjecture rather than empirically verified relationships of skills. In intelligent tutoring, skill sets are assembled into an *expert model* that contains a representation of the knowledge and skills that are required to accomplish a task. These skill sets are verified as the expert model must also be a "runnable" model that is used to produce output that is compared to student output (Wenger, 1987). Work is being done using computers to automatically sequence hierarchical skills based on a Hasse diagram, a graph theory variant that shows prerequisite relationships (Heller, Steiner, Hockemeyer, & Albert, 2006).

A template for a prerequisite hierarchy or prerequisite analysis can be created for some of the more prescribed task categories within content taxonomies. For example, a task category within the Gagne and Briggs (1974) taxonomy is *defined concept*. This is an intentional concept that has a fairly clear boundary condition, defined by a set of attributes common to all instances of the concept. To be able to correctly classify an instance, a performer must be able to identify each of the critical attributes of the concept. Merrill (1983) suggests that recalling the definition of the concept can also assist in learning intentional concepts. Figure 9.2 shows a generic prerequisite hierarchy template for a defined concept.

Each critical attribute may be a defined concept itself (Interrante & Heymann, 1983) in which case the template would reused at different levels of granularity. A diagnosis of an unsuccessful performance in this category would involve identifying the missing prerequisite skills of a performer.

Purely procedural skills can benefit from a procedural analysis in which each step or decision is listed, often in the form of a list or flowchart. Diagnosing an

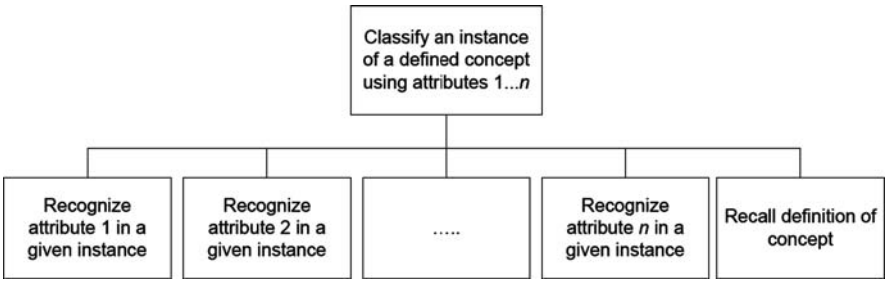


Fig. 9.2 Prerequisite hierarchy template for a defined concept

unsuccessful performance in this task category would entail the identification of missing steps, incorrectly ordered steps, or errors in accuracy of any step.

For example, the steps to prepare American fast food are procedural and fully prescribed: materials and tools are standardized, workers are limited to a handful of repetitive steps (such as using a "condiment gun" to squirt ketchup and mustard on buns) and tasks are carefully timed. If the fast food in incorrectly prepared, one would suspect that a performer skipped a step, or performed steps out of order, or without the proper accuracy. Diagnosis that systematically searches for the presence or absence of required sub-skills or missing or out of order steps is an ideal strategy for creating diagnostic systems for prescribed tasks.

9.4.2 Discretionary Task Diagnosis

Discretionary tasks, as stated, are irregular, innovative, or ill-defined. Because these tasks are irregular in process, there is no effective way to delineate the steps that a performer should take, and consequently there is no effective way diagnose difficulties in *process*; without clear algorithms, it is difficult to find missing steps or to correct poor performance. While these types of tasks are solution-oriented, in which a performer is expected to create solutions rather than follow algorithmic procedures, in the most extreme case, we may not have a clear description of what "success" entails; judging output is a challenge.

It is difficult if not impossible to fully enumerate prerequisite skills in discretionary tasks. Therefore, diagnosis in the form of identifying missing prerequisite skills is not possible with extreme discretionary tasks.

If we cannot determine the underlying knowledge and skills that are lacking, how can we correct behavior? An analogy may help: A physician who cannot determine the underlying cause of a symptom, or who cannot classify a set of symptoms into a "disease" category must still treat the patient. The physician would use more general medical techniques if unsure of the diagnosis. In like manner, we can still offer general heuristic support toward task completion, even if we are not certain of the

underlying processes. Abstractly, we have bypassed discretionary task diagnosis and jumped directly to the stage of performance support.

Diagnosis for extreme discretionary tasks might be better conceived as *process support* rather than *process repair*. Process support would provide general assistance in the process of solution creation, providing techniques for generative innovative thinking and creative problem solving. For example, this support might offer analogies and heuristics to guide a performer toward a solution goal, rather than seek missing skills. Numerous books on creativity (Young, 1960; Adams, 1979; Michalko, 2006; van Oech, 1992; Young, 1960) suggest that generating ideas can be done systematically by techniques such as substitution or nontraditional combinations among others. Many of these techniques use a hill-climbing idea of continual progress toward finding an adequate solution through the generation of many possibilities.

An example of a process support approach that has been used by creative artists, both musical and visual, involves a set of principles referred to as the *Oblique Strategies*. These principles were assembled by visual artist Peter Schmidt and musician/producer Brian Eno. The *Oblique Strategies* began as a set of text notes that each artist assembled to be used under circumstances when time pressure interfered with the creative process. Essentially, the *Oblique Strategies* can be used as reminders of basic creative principles to jog the thinking of the artist (Eno, 1980). A similar set of mental refocusing activities can be found in the form of creative cards called the *Wack Pack* based on van Oech's work (1992).

Yet another process support technique might entail simply reviewing parallel solutions in the same field or in other fields. Metaphors between business and war, business and games, and education and music are all commonly used to jog thinking in a creative way. Solutions in one field may be analogically used as structures for solutions in another area (Gick & Holyoak, 1983).

9.5 Delivery Mechanisms for Diagnosis in Prescribed Tasks

We turn now to a brief review of delivery mechanisms and interfaces that can be used for diagnosis. Although the diagnostic techniques previously described could be delivered by humans, the following discussion is limited to automating the diagnostic effort using computing technologies. These ideas are not meant to be exhaustive, but merely suggestive of the possibilities. First, diagnosis based upon content categories in computer-based training is described, followed by ideas to expand this technique to work tasks. An agent-based technique for prescribed tasks is then described.

9.5.1 Computer-Based Training Diagnosis

Computer-based training (CBT) and computer-assisted instruction (CAI) have a long tradition of using embedded questions within an instructional module to determine branching strategies for each student. Historically, this branching was referred

to as *individualized instruction*; the individualization occurred as some students saw all of the content while others saw a subset of the total content. In the earliest days of CBT, branching was done based on questions that tested for comprehension of content. Students who failed to learn (i.e., were unable to answer comprehension type questions) were branched to remedial sections of the course, that provided alternative explanations. This was a common feature in early CBT and was, in essence, a crudely cut diagnosis of student problems, and a matched effort at remediation.

A more modern approach uses adaptive computer-based instruction that monitors student success and failure within instructional units and builds a student model of student individual differences. The student model is then used to dynamically alter the presentation of content based upon various cognitive styles criteria, such as a preference for structure. Triantafillou, Pomportsis, and Demetriadis (2003) created an instructional system in which two-page variants were created for each instructional page in a CBT lesson. The pages differed in the presentation of instructional elements such as graphics and navigation. The student model determined which pages would be presented differentially to students. Initial formative evaluation results indicated that matching cognitive styles with presentation and navigation features was by and large beneficial. The Triantafillou et al.'s system does not attempt to search for missing task-related skills, but appears to react to more general information processing preferences.

Within a practice session of computer-based training, diagnostics can be created using interaction patterns based on content categories. A somewhat standard diagnostic path can be added to any computer-based instruction that has isolated various content types. For example, a lesson that teaches *defined concept* classification can branch to a diagnostic path under certain instructional conditions, such as a specific number of missed practice questions. A simplified flowchart of such a diagnostic template is seen in Fig. 9.3, and shows that learning any concept is based upon being able to accurately identify the critical features of the concept and is assisted by being able to recall the definition of the concept; failure to correctly classify the concept triggers investigation into probable causes.

9.5.2 Work Task Diagnosis

As discussed previously, some work tasks are done in the absence of computers. How then can inadequate task performance be diagnosed? One solution is through computer-based simulation of the task. Tasks, tools, and environments can be simulated with varying degrees of fidelity, from crude drag and drop interfaces on personal computers to multi-million dollar simulators of complex machinery and human interaction (Yacci, 2004). Higher fidelity interfaces generally provide the opportunity for detecting more subtlety in task performance. In more prescribed tasks, the additional expense of higher fidelity simulators may not always be needed or warranted.

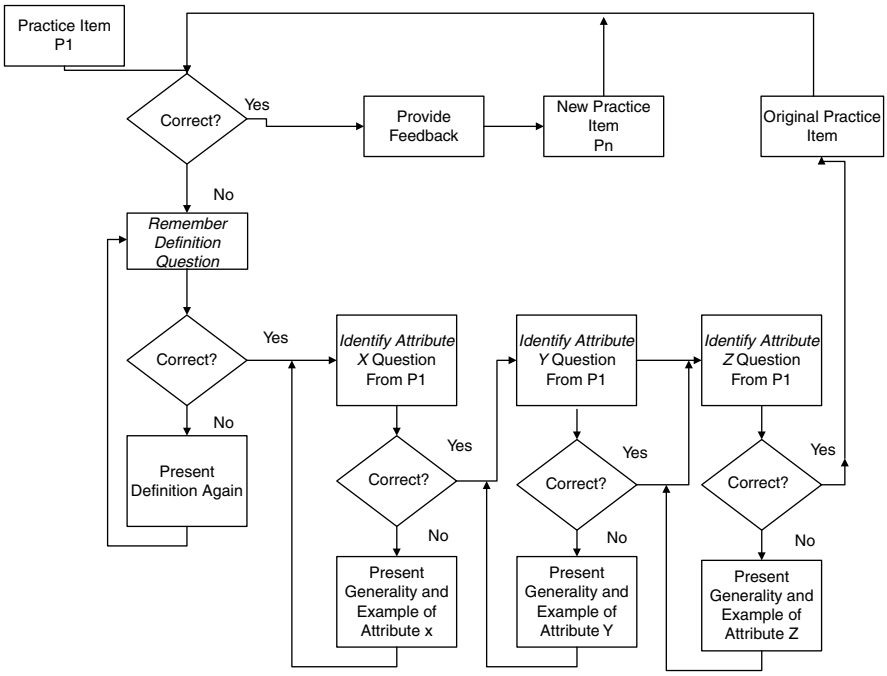


Fig. 9.3 Diagnostic path for defined concept classification

The diagnostic paths previously discussed can be used in conjunction with simulations of work tasks. Although diagnostic paths are generally constructed when frame-based CBT is first designed and built, stand-alone diagnostic applications could be "bolted on" to simulations running within SCORM-compliant Learning Management Systems. These stand-alone diagnostic applications could be used in conjunction with work task performance as a means to determine likely causes of worker error.

9.5.3 Agent-Based Diagnosis

A common agent-based interface is a *text chat bot*, a disembodied intelligence that communicates through text or pictures. Text chat bots use common communication tools such as *Instant Messenger* and can provide a sophisticated level of interactivity. A bot is capable of mixed initiative dialogue, in which it can both answer questions and pose questions to users. In contrast to the embedded diagnostics that are usually built in to instructional systems, a text chat bot would exist as a separate system, apart from an instructional system or a task simulation.

A diagnostic text chat bot can be effective with prescribed tasks using the templates and paths previously displayed in Figs. 9.2 and 9.3. The hierarchical diagnostic system (Yacci & Lutz, 2004) was a prototype using this form of template that operated via *Instant Messenger*. The hierarchical diagnostic system implemented a hierarchical prerequisite model and used variations of the Fig. 9.3 diagnostic path to pose skill-based questions. The student responses to these questions determined what sub-skills existed in the student and what sub-skills were missing. This flowchart was extended into a depth-first search that drilled down into the prerequisite hierarchy to iteratively investigate the lack of sub-skills. One drawback of this system, based on initial evaluation, was that the prototype required many "trivial questions" (Wenger, 1987) be asked of the users.

9.6 Delivery Mechanisms for Diagnosis in Discretionary Tasks

Due to the unconstrained nature of discretionary tasks, it is difficult for the computer to recognize a successful process or solution in these types of tasks. Simply put, the computer cannot easily or accurately determine correctness of responses or process if the task is extremely discretionary. This section examines two techniques for diagnosis in prescribed tasks: simulation and consultant agents.

9.6.1 Discretionary Simulations

Earlier, it was discussed that computer-based simulation could be used as an environment for workers to perform prescribed tasks to allow for diagnosis of worker errors. When computer-based simulation is used for discretionary tasks, how does it differ from that described for prescribed tasks? In simulating the performance of discretionary tasks, the accuracy of an invented solution would depend upon the quality and accuracy of the model of reality that underlies the simulation. To accurately diagnose problems with discretionary tasks, a high fidelity simulation model would be necessary.

Dorner (1996), for example, describes a simulation game in which students are given the opportunity to act as a mayor of a city. The "long-term" success of decisions can be inferred on the basis of the simulation because years can be simulated by minutes and hours. However, the ultimate quality of the decisions that are made in the simulation depend upon the richness and complexity of the simulation model. If the underlying model is not accurate, then the results of the simulation are misleading.

Shank, Berman, and Macpherson (1999) describe goal-based scenarios (GBSs) that are high fidelity simulations of complex environments involving many discretionary tasks. Expert advice and consultation are available at various points of the simulation in the form of *stories* that can help a performer be more successful in the simulation. These stories are usually analogs to the current situation – direct advice

is not provided to learners. While a GBS is complex, it does have a clearly stated goal, and the simulation provides a means to judge successful resolution of the situation. Again, a rich multivariate model underlies the simulation, and task success in the "real world" will only be as accurate as the underlying model.

9.6.2 Consultant Agents

Expert advice can also be provided at an extremely general, high level using *embodied conversational agents*. An embodied conversational agent (ECA) combines conversational ability with an underlying knowledge base and a synchronized visual avatar. In many cases, an ECA is a "talking head" that has audio speech capacity, and is able to provide mixed initiative conversation, helping to direct the user and also responding to questions. An ECA mimics human interaction, featuring synchronized voice and mouth movements, and has many other humanizing features, such as eye blinks and head movements; some even change their wardrobe from day to day. An ECA can exist as a separate system apart from a simulation or an instructional system.

Generally speaking, agents have limited knowledge of a domain, but have fairly good capacity to communicate. Mixed initiative dialogue is possible, as questions can be initiated by either the agent or the student. Most agent-based systems do a reasonable job of replicating the dynamics of conversation, although many agents do not have true natural language capabilities. Open-text input is usually allowed, as ideally, an ECA should have the capacity to converse with users using natural language. However, the complexity of natural language limits the use of this feature in many agents (Allen et al., 2001). A more simple agent can be constructed using a keyword match that gives the appearance of natural language with acceptable results (Plantec, 2004) without the complexity of natural language understanding. In the simplest agents, conversation can be constrained by a menu-based interface in which human input is limited to a multiple choice response.

Perhaps the biggest benefit of agent-based support is the ability to provide general assistance with discretionary tasks. An agent can assist task completion in the form of heuristic idea-generation strategies that *might* work. In essence, the agent provides *consultant* type help for discretionary tasks.

The agent-as-consultant does not need to be able to detect the accuracy of the current procedure or outcome in a discretionary task; it relies on user-initiated requests for assistance. The agent can converse with the user and could customize general creative strategies to the situation. Workers might consult the agent with more general purposes – to get "unstuck" in a creative task, to ask for "sage" wisdom, to maintain motivation, or to simply have a sounding board.

The agent does not closely monitor the task at hand, but instead provides general direction. For example, an agent can carry on a fairly neutral conversation while providing creative techniques such as the SCAMPER approach (Michalko, 2006)

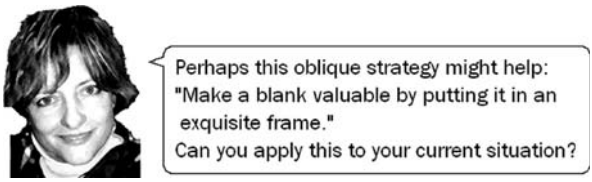


Fig. 9.4 Prototype consultant agent

or the Oblique Strategies (Eno, 1980). Figure 9.4 below shows a prototype under development of a consultant agent that is delivering a randomly selected *Oblique Strategy*. Although this prototype was created using Microsoft Agent, there are many agent development systems that could implement such a system (Prendinger & Ishizuka, 2004).

A recent study showed that students often anthropomorphize ECA and treat them like humans (Doering, Veletsianos, & Yerasimou, 2008). This is consistent with the research of Reeves and Nass (1996) who also found that humans treat *any* communicative computing system with human social rules. The Doering, Veletsianos, and Yerasimou study also showed that many students found ECA to be good study companions; students spent a fair amount of time conversing with them about off-task topics. This suggests that ECA might have the capacity to act as process consultants, due to the fact that students can achieve a relaxed trust with them.

Some designers believe that ECA should have distinct personalities (Plantec, 2004). A personality makes an agent more believable and the believability of the agent is helpful in adding to the motivation of the student and encourages the student to implement a tip or strategy. An agent with an encouraging personality and very little content could easily be perceived as a successful consultant, as evidenced by the historic reactions to Weizenbaum's *Eliza* program. Hayes-Roth (2004) suggests multiple characteristics to add to the believability of agents, but ultimately suggests that an agent does not have to fully mimic a human as much as "suspend disbelief" in users to enable a pleasurable interaction.

9.7 Conclusion

The logic behind the task continuum suggests that there are different forms of diagnosis and repair of tasks to be used across the continuum. Prescribed tasks are amenable to a more systematic diagnosis of missing sub-skills or failed process due to the relative ease of decomposition of the tasks. In discretionary tasks, diagnosis of root causes of task failure may not truly be possible at this time.

Computer-based delivery systems can support some forms of diagnosis better than others. Embedded diagnostic paths can be implemented in almost any interface to diagnose prescribed tasks, even tasks that are not normally performed at a computer. Perhaps the most interesting form of diagnosis surrounds discretionary tasks, in which diagnosis is bypassed for performance consultancy. Creative tasks can be

supported through an agent-based conversational consultant that provides any number of creative thinking approaches useful for generating new ideas and potential solutions. The conversational nature of agents makes them useful as high-level consultants. Current work is being done to improve the capabilities of agents and tests of the agent-as-consultant paradigm will follow.

References

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied, 6*, 259–290.

Adams, J. L. (1979). *Conceptual blockbusting* (2nd ed.). New York: W.W. Norton & Company.

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human–computer interaction. *AI Magazine, 22*(4), 27–37.

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: W.H. Freeman and Company.

Annett, J., & Duncan, K. D. (1967). *Task analysis and training design*. Eric Document ED 019566.

Bennett, J. G. (1997). Living in the medium. In F. Barron, A. Montuori, & A. Barron (Eds.), *Creators on creating*. New York: G.P. Putnam's Sons.

Bloom, B. S. (1956). *Taxonomy of educational objectives*. New York: David McKay Company, Inc.

Connell, M. W., Sheridan, K., & Gardner, H. (2003). On abilities and domains. In R. L. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise*. Cambridge, UK: Cambridge University Press.

Dorner, D. (1996). *The logic of failure.* Reading, MA: Addison-Wesley.

Doering, A., Veletsianos, G., & Yerasimou, T. (2008). Conversational agents and their longitudinal affordances on communication and interaction. *Journal of Interactive Learning Research, 19*(2), 251–270.

Eno, B. (1980, February 2). *Ode to gravity*. [Radio Broadcast]. Host: Charles Amirkhanian. KPFA-FM. Retrieved December 20, 2008, from http://www.archive.org/details/BrianEno

Feynman, R. (1997). The dignified professor. In F. Barron, A. Montuori, & A. Barron (Eds.), *Creators on creating*. New York: G.P. Putnam's Sons.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston, MA: Pearson.

Gagne, E. D. (1985). *The cognitive psychology of school learning.* Boston, MA: Little, Brown & Company.

Gagne, R. M. (1968). Learning hierarchies. *Educational Psychologist, 6*, 1–9.

Gagne, R. M., & Briggs, L. J. (1974). *Principles of instructional design.* New York: Holt, Rinehart and Winston, Inc.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Hayes-Roth, B. (2004). What makes characters seem life-like? In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters*. Berlin: Springer.

Heller, J., Steiner, C., Hockemeyer, C., & Albert, D. (2006). Competence-based knowledge structures for personalized learning. *International Journal on E-Learning, 5*(1), 75–88.

Hopp, W. J., Iravani, S. M. R., & Yuen, G. Y. (2007, January). Operations systems with discretionary task completion. *Management Science, 53*(1), 61–77.

Horn, R. E. (1989). *Mapping hypertext*. Lexington, MA: The Lexington Institute.

Horn, R. E. (2002). *Beginning to conceptualize the human cognome project.* Retrieved October 20, 2008, from http://www.stanford.edu/~rhorn/a/topic/cognom/artclCncptlzHumnCognome.pdf.

Interrante, C. G., & Heymann, F. J. (1983). *Standardization of technical terminology: Principles and practice.* West Conshohocken, PA: ASTM International.
Jeffries, K. K. (2007). Diagnosing the creativity of designers: Individual feedback within mass higher education. *Design Studies, 28*, 485–497.
Mager, R. F., & Pipe, P. (1984). *Analyzing performance problems or 'you really oughta wanna'* (2nd ed.). Belmont, CA: Lake Publications.
Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: An overview of their current status.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Michalko, M. (2006). *Thinkertoys: A handbook of creative-thinking techniques.* Berkeley, CA: Ten Speed Press.
Minsky, M. (1986). *The society of mind.* New York: Simon & Schuster.
Prendinger, H., & Ishizuka, M. (2004). *Life-like characters*. Berlin: Springer.
Plantec, P. (2004). *Virtual humans*. New York: American Management Association.
Reeves, B., & Nass, C. I. (1996). *The media equation*. Cambridge: Cambridge University Press.
Romiszowski, A. J. (1981). *Designing instructional systems*. London: Kogan Page.
Romiszowski, A. J. (1999). The development of physical skills. In C. M. Reigeluth (Ed.), *Instructional-design theories and models. Volume 2: A new paradigm of instructional theory*. Mahwah, NJ: Lawrence Erlbaum Associates.
Shank, R. C., Berman, T. R., & Macpherson, K. A. (1999). Learning by doing. In C. M. Reigeluth (Ed.), *Instructional-design theories and models. Volume 2: A new paradigm of instructional theory*. Mahwah, NJ: Lawrence Erlbaum Associates.
Smith, P. L., & Ragan, T. J. (2005). *Instructional design*. Hoboken, NJ: J. Wiley & Sons.
Stake, R. E. (2004). *Standards-based and responsive evaluation*. Thousand Oaks, CA: Sage Publications, Inc.
Triantafillou, E., Pomportsis, A., & Demetriadis, S. (2003). The design and the formative evaluation of an adaptive educational system based on cognitive styles. *Computers & Education, 41*, 87–103.
Ullman, M. (2001, October). A neurocognitive perspective on language: The declarative/procedural model. *Neuroscience, 2*, 717–726.
Van Oech, R. (1992). *A wack on the side of the head: How you can be more creative*. Berkeley, CA: Atlantic Books.
Wenger, E. (1987). *Artificial intelligence and tutoring systems*. Los Altos, CA: Morgan Kaufmann.
Yacci, M. (2004). *Game based learning: Structures and outcomes*. Proceedings of the Society for Information Technology and Teacher Education (SITE) 15th International Conference, Atlanta, GA.
Yacci, M., & Lutz, P. (2004). *Conversational diagnostic agent*. Proceedings of AACE World Conference on Educational Multimedia, Hypermedia, and Telecommunications, Lugano, Switzerland.
Young, J. W. (1960). *A technique for producing ideas*. Chicago: Advertising Publications.

Part III
Comparison and Empirical Testing Strategies

Intermezzo 3 – The Inner Workings of Knowledge and Its Structure: Reasoning, Comparison, Testing, Evaluation, Decision, and Action

Pablo Pirnay-Dummer and Dirk Ifenthaler

Once the external re-representations have been assessed and aggregated, two competing demands are at hand: First, we need to keep as much information from the external re-representations as possible. Secondly, especially in large datasets the information needs to be condensed in such a way that we are still able to selectively decide on or test our theories and practical goals. Combining both demands is not always easy and the measures need to be chosen carefully with an eye to the research question, evaluation, analysis, or designed plans in order to provide the proper answers. In the field of computer-based diagnostics knowledge artifacts (objects of investigation) are very often graphs. If they are not graphs from the start, they are usually transferred into graphs after assessment. The purpose is aggregation, as we saw in the last part of this book. Purely qualitative methods are the exception. However, their opposition to any kind of aggregation lies in their nature, and they can be aided by computer programs but not carried out automatically. Any aggregation of qualitative research results is at least to be considered a mixed method: Aggregation is quantitative by nature. This does not, on the other hand, mean that all aggregation serves the same purpose or that it can not differ in quality and the amount of information it preserves. As always, the choice of the right measures and comparisons is determined by the research question or practical goal. The main reason for comparison is the further processability of the artifacts, which is especially interesting for computer-based analysis because it can be automated. The indices allow questions about whether one group of experts structures things differently than another or whether a group of learners makes progress over time, e.g., as compared to experts. With computer-based analysis, large data sets are attainable even if resources are limited. When the objects under investigation are graphs, graph theory provides the only logical choice for analysis and a stable basis for several further developments. Surprisingly, the application of graph theory can only rarely be found in research on learning and instruction. Usually, very simple measures are

used as single indicators which do not carry much of the initially rich information and are usually not validated at all. And even in the case that graph theory is applied, the indices used sometimes lack a connection to the theories of learning and instruction, and the scope of the measures is sometimes misinterpreted. The unfavorable application of measures then misleads some colleagues to assume that quantitative measures and comparison are not suitable at all for describing construct-like knowledge. The following third part of the book shows the potential of properly applied graph theory and investigates several measures in great detail. It also shows the standards by which measures should be evaluated methodologically within each chapter. Also, the contingencies of the re-representations for reasoning are discussed to provide a framework for the interpretation of knowledge models, thus closing the circle of the theoretical basis of representation, processes of knowing, and reasoning.

Chapter 10
Graphs and Networks

Peter Tittmann

10.1 Graphs as Representations of Binary Relations

In many fields of science we deal with (technical) *terms*, *concepts*, *notions*, *mental pictures*, or *ideas* that may be *similar*, *dependent*, *correlative*, or in some way *related*. In order to get an overview of the particular field of interest, we may draw small boxes or circles on a sheet of paper that symbolize the terms or ideas. If two ideas are somehow related, then we connect them by a line. In case of a directed relation (like dependency), we use an arrow instead of a line. This simple procedure may work quite well for a dozen of terms and relations. However, we need more sophisticated methods if thousands of concepts are to be analyzed. There is a second, perhaps more important, reason for the introduction of formal methods. Even in case we have got a nice graphical representation of all concepts and relations – what kind of conclusions can we draw from the picture? Here methods from *graph theory* are of great value. They offer possibilities to introduce network indices in order to *measure* the distance between concepts, to *identify* automatically groups of strongly related terms, or to *compare* different ideas. We present here only the first basic concepts of graph theory. Readers who desire to gain deeper insight into this fascinating subject are referred to textbooks like Gross and Yellen (1998) or Bondy and Murty (2008).

The set of concepts (terms, etc.) is in general a finite set $A = \{a_1,\ldots,a_n\}$. A *binary relation* on A is a subset $R \subseteq A \times A$ of ordered pairs of elements from A. Here $A \times A = \{(x,y)|x,y \in A\}$ denotes the set of all ordered pairs of elements from A. Since we deal with binary relations exclusively, we leave out the word "binary" in the following. A relation $R \subseteq A \times A$ is called *reflexive* if for all $a \in A$ the pair (a,a) is contained in R, i.e., each element is related with itself. The relation R is *symmetric* and *antisymmetric*, respectively, if $(a, b) \in R$ for $a \neq b$ implies $(b,a) \in R$ and $(b, a) \notin R$. The relation R is *transitive* if $(a, b) \in R$ and $(b,c) \in R$ imply $(a,c) \in R$. Relations can be vizualized by graphs.

P. Tittmann (✉)
Hochschule Mittweida, Mittweida, Germany
e-mail: peter@htwm.de

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_10,

An *undirected graph* $G = (V,E)$ consists of a set of *vertices* V and a set of *edges* E such that to each edge there are two vertices assigned, namely the *end vertices* of the edge. We write $e = \{v,w\}$ if v and w are the end vertices of the edge e. In this case, the edge e is said to *join* the vertices v and w. The vertices v and w that are joined by an edge e are *neighbors* in G. We also say that v and w are *adjacent* and that the edge e is *incident* to the vertex v (and w). A graph can be *drawn* in the plane. A vertex is pictured as a point (a small circle) whereas an edge corresponds to a line connecting two vertices (points). Figure 10.1 shows an undirected graph with nine vertices and nine edges. Two edges are said to be *parallel* if they share the same pair of end vertices. The edges $a = \{1,2\}$ and $b = \{1,2\}$ of our example graph are parallel. A *loop* is an edge whose end vertices coincide. In Fig. 10.1, we find the loop f attached to vertex 6. A graph that has neither loops nor parallel edges is called *simple*. A *directed graph* (or short *digraph*) has *arcs* rather than edges. An arc $e = (u,v)$ is a directed edge connecting two vertices, namely its *tail* u and its *head* v. In case of a loop, these two vertices coincide. Remark that we use the notation (u,v) for an ordered pair of vertices whereas $\{u,v\}$ denotes an unordered pair. Figure 10.2 shows a digraph with seven vertices.

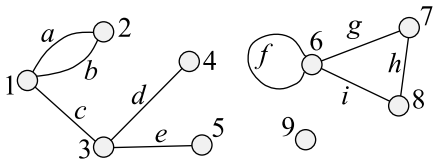


Fig. 10.1 An undirected graph

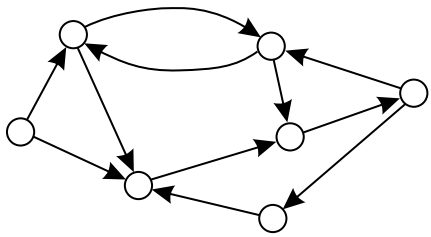


Fig. 10.2 A digraph

The *degree* $\deg v$ of a vertex v is the number of edges that are incident to v. For a vertex v of a digraph G, we distinguish between *indegree* $d^-(v)$ and *outdegree* $d^+(v)$ counting the arcs of G directed to and from v, respectively. The *neighborhood* $N(v)$ of a vertex v in a graph G is the set of all neighbors (adjacent vertices) of v. In a simple graph, we have $|N(v)| = \deg v$.

A *walk* in a graph G is an alternating sequence of vertices and edges (or arcs in a digraph),

$$v_0,e_1,v_1,e_2,v_2,\ldots,v_{k-1},e_k,v_k$$

such that each edge appears between its end vertices within the sequence. If each arc e_i of a walk is directed from v_{i-1} to v_i, then the walk is called *directed*. The *length* of a walk is the number of its edges where repetitions are also counted. A walk is *closed* if the initial vertex of the walk is also the final vertex. A *path* in a graph is a walk such that no vertex is repeated. A *cycle* is a closed walk such that no internal vertex appears twice.

The *distance* $d(u,v)$ between two vertices u and v of a graph $G = (V,E)$ is the length of a shortest path from u to v. In an undirected graph, the distance is symmetric, i.e., $d(u,v) = d(v,u)$. In a digraph, the distance is in general not symmetric. The distance satisfies the *triangle inequality*, i.e., for every three vertices $u,v,w \in V$, we have $d(u,w) \leq d(u,v) + d(v,w)$. There are efficient methods, like Dijkstra's algorithm, in order to find shortest paths in graphs, see, for instance, Korte and Vygen (2008) or Papadimitriou and Steiglitz (1998).

An undirected graph is *connected* if there exists a path between every two vertices of the graph. We call a digraph connected if the underlying undirected graph arising by neglecting the orientation of the arcs is connected. A digraph is *strongly connected* if there exists a directed uv-path for each ordered pair (u,v) of vertices. The digraph drawn in Fig. 10.2 is connected but not strongly connected. The *eccentricity* of a vertex v in a connected graph G is the maximum of all distances from v to other vertices of G. The *diameter* $D(G)$ of a graph G is the maximum over all distances between two vertices of G. Consequently, the diameter is also the maximum eccentricity of a vertex of G.

There are some special graphs that arise in many applications. A *complete graph* K_n with n vertices is a simple graph in which every two vertices are adjacent. Therefore, the complete graph K_n has exactly $\frac{n(n-1)}{2}$ edges. Figure 10.3 shows complete graphs with three, four, and five vertices. A *bipartite graph* $G = (V \cup W,E)$ consists of a vertex set that is composed of two disjoint subsets V and W such that each edge of E has one end vertex in V and the other end vertex in W. Bipartite graphs arise often in connection with assignment problems where the vertex sets correspond to two categories, like objects and classes. An example is presented in Fig. 10.4.

A complete bipartite graph $K_{p,q}$ is a bipartite graph $G = (V \cup W,E)$ with $V = \{v_1, \ldots v_p\}$, $W = \{w_1, \ldots w_q\}$ and $E = \{\{v_i,w_i\} | 1 \leq i \leq p, 1 \leq j \leq q\}$. Hence, each vertex of V is linked by an edge with each vertex of W in $K_{p,q}$. The number of edges in $K_{p,q}$ is pq. A tree is a connected graph without any cycles. A directed tree is a digraph whose underlying undirected graph is a tree. For many applications, rooted

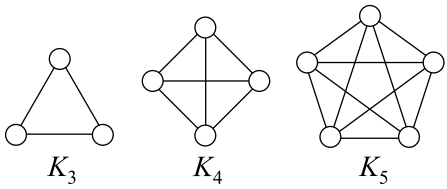


Fig. 10.3 Complete graphs

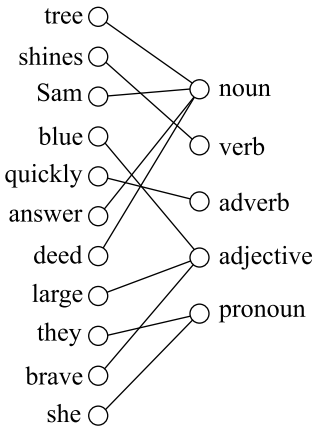


Fig. 10.4 A bipartite graph assigning words to parts of speech

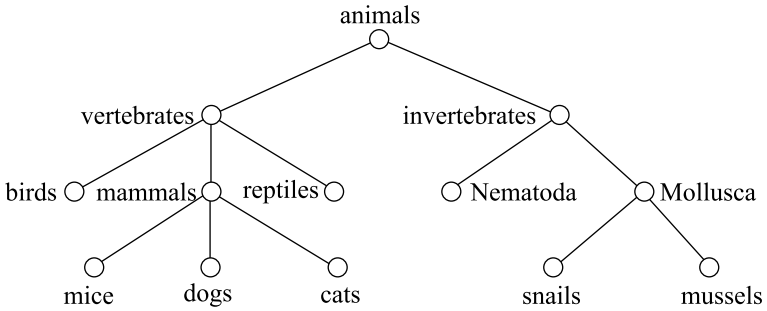


Fig. 10.5 Incomplete classifications of animals

trees are of special interest. A rooted tree is a tree with a distinguished vertex, called the root of the tree. An application of a rooted tree for the classification of animals is shown in Fig. 10.5.

10.2 Graphs and Matrices

In order to process graphs with computers, we need a different representation rather than drawings. A first idea is to store the graph structure in an edge list, that is, a table whose columns are indexed by the edge names. In column e, we find the end vertices of edge e. The following table corresponds to the graph shown in Fig. 10.1. The table may be extended with further rows, in case of given edge weights.

edge	a	b	c	d	e	f	g	h	i
vertex 1	1	1	1	3	3	6	6	7	6
vertex 2	2	2	3	4	5	6	7	8	8

The edge list representation is also suitable for digraphs. Especially in case of simple graphs, the *adjacency list* is a concise graph representation. The adjacency list contains for each vertex the list of neighbors. If the lists are stored with forward and backward pointers then graph operations, like insertion and deletion of vertices or edges, can be performed quickly.

An important data structure for graphs is a matrix representation. Let $G = (V,E)$ be an undirected graph with vertex set $V = \{v_1, \ldots v_n\}$ and edge set $E = \{e_1, \ldots e_m\}$. The *adjacency matrix* $A = (a_{ij})_{n,n}$ of G is a square matrix with the entries

$$a_{ij} = \textit{number of edges between } \upsilon_i \textit{ and } \upsilon_j$$

The adjacency matrix of the graph presented in Fig. 10.6 is

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

The sum of each row or column of the adjacency matrix yields the degree of the corresponding vertex. For a digraph G, we define a_{ij} to be the number of arcs in G that are directed from v_i to v_j. The adjacency matrix of a digraph is in general not symmetric. Representing graphs by matrices requires often more storage than an adjacency list. On the other hand, some graph operations, like test for adjacency, can be performed extremely fast. In addition, matrix representations offer all tools from linear algebra. Let $A^k = \left(a_{ij}^{(k)}\right)$ be the k-th power of the adjacency matrix A of a given graph (or a digraph). Then one can easily show that the entry $a_{ij}^{(k)}$ of this matrix equals the number of walks of length k from v_i to v_j in G. We conclude that if $a_{ij}^{(k)} = 0$ for $k = 0,\ldots,l–1$ and $a_{ij}^{(l)} > 0$ then the distance between v_i and v_j is l. In this case, $a_{ij}^{(l)}$ counts the shortest paths between v_i and v_j. For our example graph presented in Fig. 10.6, we obtain $a_{14}^{(2)} = 2$ which corresponds to the two paths v_1,e_1,v_2,e_4,v_4 and v_1,e_1,v_3,e_5,v_4.

The *incidence matrix* of a graph $G = (V,E)$ with vertex set $V = \{v_1 \ldots v_n\}$ and edge set $E = \{e_1 \ldots e_m\}$ is an $n \times m$ matrix $B = \left(b_{ij}\right)$ with entries

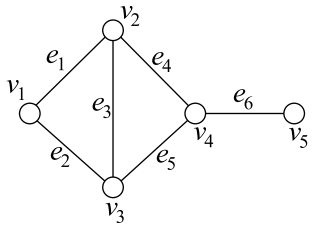


Fig. 10.6 An undirected graph

$$b_{ij} = \begin{cases} 1, & \text{if } v_i \text{ is incident to } e_j, \\ 0, & \text{else.} \end{cases}$$

In a digraph, the entries of the incidence matrix are defined in the following way:

$$b_{ij} = \begin{cases} -1, & \text{if } e_j \text{ is directed to } v_i, \\ 1, & \text{if } e_j \text{ is directed to } v_i, \\ 0, & \text{else.} \end{cases}$$

The incidence matrix of the graph shown in Fig. 10.6 is

$$B = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Let $D = (d_{ij})_{n,n}$ be the *degree matrix* of a graph G, that is, a diagonal matrix with diagonal entries $d_{ii} = \deg v_i$. The definition of the incidence matrix implies $BB^T = A + D$ in a graph and $BB^T = D - A$ in a digraph.

The matrix $L = D–A$ is called the *Laplacian matrix* of G. A graph $H = (W,F)$ is a *subgraph* of a graph $G = (V,E)$ if $W \subseteq V$ (W is a subset of V) and $F \subseteq E$. A subgraph H is *spanning* if it contains all the vertices of the supergraph G. A *spanning tree* of a graph G is a cycle-free connected spanning subgraph of G. Figure 10.7 shows all eight spanning trees of the graph represented in Fig. 10.1. Let $t(G)$ be the number of spanning trees of G. We denote by L_i the matrix obtained from the Laplacian matrix L by canceling an arbitrary row i and column i. The famous Matrix-Tree Theorem by Kirchhoff (1847) states that $t(G) = \det L_i$. The Laplacian matrix of a graph has many interesting applications for measuring the global connectivity of a graph, for finding minimum cut sets (minimum edge sets whose removal from G disconnects the graph), or for graph-drawing procedures. The interested reader is referred to books on algebraic graph theory, for instance, Cvetcović, Doob, and Sachs (1995) or Godsil and Royle (2001).

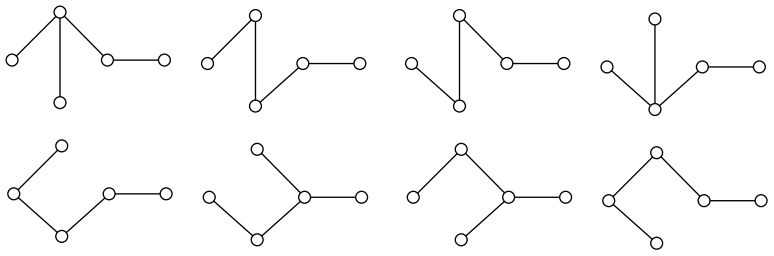


Fig. 10.7 Spanning trees

10.3 Connectivity

So far we can distinguish connected and disconnected graphs. Nevertheless, it is often desirable to measure the "strength" of connectedness of a graph. In some cases, the removal of a single edge of a connected graph G disconnects the graph. We call such an edge a *bridge* of G. In a tree, each edge is a bridge. We denote by $G–e$ the graph obtained from G by the removal of the edge e. More general, let $G–A$ denote the graph obtained from G by removing all edges of the edge set A. The maximal connected subgraphs of a graph G are called the *components* of G. The graph shown in Fig. 10.1 consists of three components. Let $k(G)$ be the number of components of G. A *cutset* of a graph $G = (V,E)$ is an edge subset $A \subseteq E$ such that $k(G - A) > k(G)$. A graph G is *l-edge-connected* if G contains no cutset with less than l edges. The *edge connectivity* $\lambda(G)$ is the greatest integer l such that G is l-edge-connected. For a tree T, we find $\lambda(T) = 1$. The complete graph with n vertices has edge connectivity $\lambda(K_n) = n - 1$. Let $\delta(G)$ be the *minimum degree* of G, i.e., $\delta(G) = \min\{\deg v : v \in V(G)\}$. Since the set of all edges that are incident to a given vertex v forms always a cutset of the graph, we conclude $\lambda(G) \leq \delta(G)$.

The famous Theorem of Menger (1927) states that in a graph with edge-connectivity k, there exists between every pair u, v of vertices at least k edge-disjoint paths, i.e., k paths between u and v that have no edge in common. In a more specific form, this theorem states that if and only if two vertices u and v are connected by k edge-disjoint paths, then each cutset separating u and v contains at least k edges. From a practical point of view, we can say that the connection between two vertices is more robust if there are more edge-disjoint paths connecting them. In another context, e.g., in semantic networks, we may interpret these paths as *independent*.

Let us now consider the effect of removing vertices with respect to connectedness. Let G be a connected graph. A vertex v of G is an *articulation* (or *cut vertex*) if the graph $G–v$, arising from G by the removal of v and all edges that are incident to v, is disconnected. We denote by $G–X$ the graph obtained from G by the removal of all vertices of the vertex subset X. A *separating vertex set* of a graph G is a vertex set X such that $k(G–X)>k(G)$. The *connectivity* $\kappa(G)$ of a graph G is the minimum cardinality of a vertex set X such that $G–X$ is disconnected or a single vertex. Two paths between vertex u and vertex v in G are called *internally disjoint* if they have no vertex in common, except u and v. There exists k internally disjoint paths between every two vertices in a graph with connectivity k (Menger, 1927).

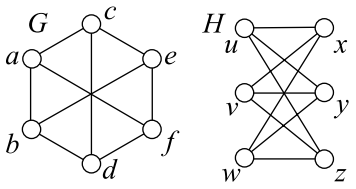
10.4 Graph Isomorphism

To decide whether two given graphs or digraphs are equal seems to be an easy question. First, we check if the corresponding vertex sets coincide and if this is the case we verify whether each pair of vertices is linked in both graphs by the same number of edges. Nonetheless, the question becomes much more intriguing if we ask whether the *structure* of both graphs is the same. With "structure" we mean graph

properties that do not depend on the labeling of the vertices or edges. To be more precise, we call two graphs $G = (V,E)$ and $H = (W,F)$ *isomorphic* if there exists a bijection (a one-to-one mapping) $\phi: V \to W$ such the number of edges between u and v equals the number of edges between $\phi(u)$ and $\phi(v)$ for all $u, v \in V$. In this case, the graph H is obtained from G by relabeling the vertices.

Figure 10.8 shows two isomorphic graphs. A *graph invariant* is a property or a function of graphs that has the same value for any two isomorphic graphs. Examples for graph invariants are the property of being connected, the number of spanning trees, or the diameter.

Fig. 10.8 Isomorphic graphs



10.5 Networks

Networks are graphs with additional weights for vertices and/or edges. There exists a huge variety of weights for edges and vertices depending on the particular application. Vertex weights may represent geographic coordinates, costs, reliabilities, potentials, etc. Edge weights can be capacities, lengths, transition probabilities, availabilities, and others.

One important example of networks is a *Markov graph*. A Markov graph $G = (V,E)$ is a digraph whose vertices correspond to *states* of a Markov chain, whereas the arcs symbolize *transitions* between states. There is a so-called *transition probability* $p_{ij} > 0$ assigned to each arc $e = (i, j)$, $i, j \in V$, such that the relation

$$\sum_{j \in V}^{p_{ij}} = 1$$

is satisfied for each $i \in V$. A state (a vertex) j is *reachable* from a state i if there is a directed path from i to j in G. We write $i \mapsto j$ in case j is reachable from i. The reachability relation is transitive, i.e., if $i \mapsto j$ and $j \mapsto k$ then $i \mapsto k$. A digraph is *strongly connected* if every two vertices of G are mutually reachable. A maximal strongly connected subdigraph of G is a *strong component* of G. A *recurrent state* of a Markov chain is a state that is infinitely often visited with positive probability. For a more precise definition, see Feller (1968). A *transient state* j has a positive probability of no return from j to j. All states within one strong component are of the same type (recurrent or transient). Consequently, we obtain a classification of states by investigating connectedness properties of the Markov graph. If we employ $-\ln p_{ij}$ instead of p_{ij} as arc weights, then we can use a shortest-path algorithm in

order to find a directed path in G that is most likely used for a transition from a state k to another state l. To show this property, consider a directed path P from k to l. Let $E(P)$ be the arc set of the path P. The probability for a transition along this path is

$$R(P) = \prod_{(i,\, j)\in E(P)} p_{ij}.$$

This probability is maximized over the set of all kl-paths in G if the sum

$$-\sum_{(i,\, j)\in E(P)} \ln p_{ij}$$

is minimal.

Another interesting network model arises in case of randomly failing edges in a graph. Assume there exists, independently of other edges, an edge $e = \{u,v\}$ (a relation) with probability p_e between any two adjacent vertices u and v in a graph G. Then graph properties like connectedness or the existence of a path between two given vertices become random events. What is the probability that a graph with stochastic independently failing edges is connected? It turns out that the answer to this question is a computational difficult (**NP**-hard) problem, i.e., a problem for which all existing algorithms require a computation time that is exponentially increasing with the network size. However, this measure is well-known in a reliability theory where it is called the *all-terminal reliability* of the network. David Karger (1995) developed an efficient approximation algorithm for this problem.

A *flow network* $N = (V,E,s,t,c)$ is a digraph $G = (V,E)$ with two distinguished vertices, a *source* s and a *sink* t. A *capacity function* assigns an upper bound (a real number) $c(e)$ to each edge e. Let $E^+(v)$ and $E^-(v)$ be the set of arcs emanating from and pointing at v, respectively. A *flow* on N is a mapping $f:E \to R$ such that for every vertex $v\{s,t\}$ the *conservation constraint* is satisfied:

$$\sum_{e\in E^-(\upsilon)} f(e) = \sum_{e\in E^+(\upsilon)} f(e)$$

A flow is *feasible* if for every arc e the relation $0 \le f(e) \le c(e)$ is satisfied. The *value of a flow* is the amount of flow leaving the source:

$$val(f) = \sum_{e\in E^+(s)} f(e) - \sum_{e\in E^-(s)} f(e)$$

A basic problem within the theory of network flows is to find a feasible flow of maximum value.

There are efficient algorithms for this purpose. The fastest currently available max-flow algorithm is based on push-relabeling techniques, see Goldberg and Rao (1988). Figure 10.9 shows a flow network with a maximum flow of value 18. According to a theorem of Ford and Fulkerson (1956), the value of a maximum

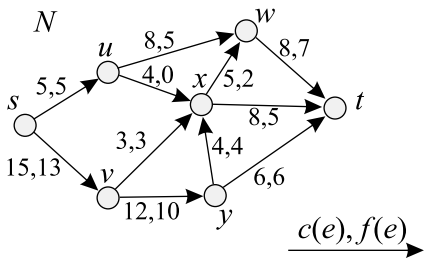


Fig. 10.9 Flow network with capacities and flows assigned to the arcs

flow in N equals the minimum capacity of a st-separating cut. Consequently, the max-flow algorithm can also be applied in order to find minimum st-cuts in N. If we assign a capacity of 1 to all arcs of N, then the value of a maximum st-flow equals the number of directed edge-disjoint st-paths.

In *social network analysis*, graphs and digraphs are used to model different kinds of social relations between people or groups. The vertices are called *actors* in this context. The edges represent relations such as friendship, liking, respect, or kinship. There exist different types (modes) of vertices in some social networks. The edges may be weighted with the "strength" of the relation. The *centrality* of a vertex (actor) is a measure for the position, the importance, or power of the actor. A first simple centrality measure is the degree of the vertex, taking into account that someone with a lot of friends should have a certain influence in a network. More sophisticated measures incorporate also the centrality of the neighbors of a vertex. This idea is realized in Google's PageRank that evaluates the importance of webpages, see, e.g., Brandes and Erlenbach (2005). Distance-based centrality measures employ the eccentricity of a vertex. The centrality of vertices increases with falling eccentricity values.

Another important problem in social networks analysis is the identification of the community structure of a social network. This problem is known in graph theory as *clustering*. We search for subgraphs of a graph that are "dense" or "strongly linked" but have a "week connection" to the rest of the graph. There are different possibilities to define density of a cluster. The requirement that the diameter of one cluster should not exceed two or three can be applied but is often too restrictive for practical applications. We can also measure the edge density in comparison with a local complete graph (a *clique*). A quite different idea is to identify edges that lay between different clusters of the network. These edges are characterized by a high *betweenness*, i.e., they are contained in many shortest paths between pairs of vertices. A fast algorithm using this approach is presented by Newman and Girvan (2004).

10.6 Drawing Graphs

In order to gain some structural insight from a picture of a network, this picture has to be drawn carefully. There are some simple properties that we expect from a "good" drawing of a graph: No two vertices should overlap or be too close to each

other. An edge should not cross a vertex that is not an end vertex of this edge. The angle between two edges leaving one vertex ought not to be too small. The number of edge crossings should be minimized. The last requirement implies that a *planar graph*, i.e., a graph that can be drawn in the plane without edge crossings, should be drawn without edge crossings. There are much more requirements characterizing a good drawing of a graph. The edges (arcs) are generally required to be presented as straight lines. However, if the graph is not simple, then we may also accept bowed lines in order to distinguish parallel edges.

Figure 10.10 shows a circular layout of a random graph with 10 vertices and 18 edges and a drawing of the same graph using a spring embedder. A *spring embedder* uses a physical model to find an embedding (a drawing) of a graph. Each vertex is thought as a metal ring; an edge corresponds to a spring connecting two rings. The springs cause forces acting on the rings. A corresponding system of equations is solved iteratively in order to find a state that minimizes the total system energy.

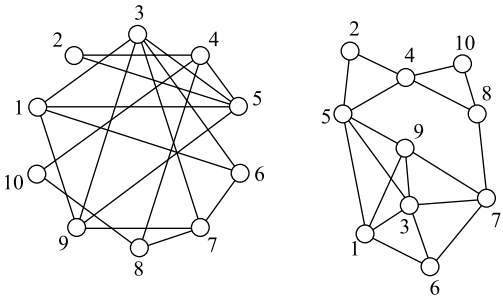


Fig. 10.10 Two drawings of one graph

Many other graph-drawing methods have been proposed, see, e.g., Battista, Eades, Tamassia, and Tollis (1999). Eigenvalues and eigenvectors of the adjacency matrix or of the Laplacian matrix, geometric properties of triangulations, flow methods, and evolutionary algorithms are applied to find suitable graph drawings. Additional requirements lead, for instance, to convex drawings or to orthogonal drawings. The strength of a relation can be visualized by different thickness values for the edges. Vertex color, shape, and size can be varied in order to emphasize important vertices or groups of vertices having special characteristics. A sophisticated free graph-drawing program is Graphviz (www.graphviz.org). The program Pajek (vlado.fmf.uni-lj.si/pub/networks/pajek) for social network analysis contains also a powerful graph visualization routine.

References

Battista, G. D., Eades, P., Tamassia, R., & Tollis, I. G. (1999). *Graph drawing – algorithms for the visualization of graphs*. Englewood Cliffs, NJ: Prentice-Hall.

Bondy, A., & Murty, U. (2008). *Graph theory*. New York: Springer.

Brandes, U., & Erlebach, T., (Eds.). (2005). *Network analysis*. Berlin: Springer.

Cvetcović, D. M., Doob, M., & Sachs, H. (1995). *Spectra of graphs*. Heidelberg: Johann Ambrosius Barth Verlag.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed.). New York: John Wiley & Sons.

Ford, L. R., & Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, *8*, 399–404.

Godsil, C., & Royle, G. (2001). *Algebraic graph theory*. New York: Springer.

Goldberg, A. V., & Rao, S. (1988). Beyond the flow decomposition barrier. *Journal of the ACM*, *45*(5), 783–797.

Gross, J., & Yellen, J. (1998). *Graph theory and its applications*. Boca Raton: CRC Press.

Karger, D. (1995). *A randomized fully polynomial time approximation scheme for the all terminal network reliability problem* (pp. 11–17). Proceedings of the twenty-seventh annual ACM symposium on Theory of Computing, Las Vegas: ACM.

Kirchhoff, G. R. (1847). Über die Auflö der Untersuchung der linearen verteilung galvanischer Ströme geführt wird. *Annalen für der Physik und der Chemie*, *72*, 497–508.

Korte, B., & Vygen, J. (2008). *Combinatorial optimization*. Berlin: Springer-Verlag.

Menger, K. (1927). Zur allgemeinen Kurventheorie. *Fundamenta Mathematicae*, *10*, 96–115.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, 026113-1–026113-15.

Papadimitriou, C. H., & Steiglitz, K. (1998). *Combinatorial optimization: Algorithms and complexity*. Mineola, NY: Dover Publications.

Chapter 11
Abductive Reasoning and Similarity: Some Computational Tools

Roger W. Schvaneveldt and Trevor A. Cohen

11.1 Introduction

This chapter outlines a psychological theory of certain aspects of creative thinking, specifically abductive reasoning, a term coined by the philosopher and logician, C. S. Peirce (1839–1914). Peirce held that the hypothetico-deductive method in science required a logic underlying the generation of hypotheses in addition to the inductive and deductive logic involved in testing hypotheses. Given some observations that are surprising or unexpected, abductive reasoning is concerned with generating hypotheses about the observations or with reasoning to the best explanation. Problem solving, in general, can often be seen to fit the abductive reasoning framework. The problem motivates a search for a solution, and abductive reasoning produces potential solutions to the problem. Peirce suggested that people have an impressive ability to formulate promising hypotheses on the basis of only a few observations.

Issues concerning novelty, evaluation, optimality, consilience, aesthetics, and pragmatics among others arise in the study of abductive reasoning. While these issues will be briefly addressed in the chapter, the primary focus is on the involvement of similarity relations in generating potential abductive inferences. In other words, the focus is on one possible explanation of how new ideas arise. We propose methods for identifying potential new connections among ideas and for displaying connections using Pathfinder networks to assist experts in searching for such promising connections. While reasoning by analogy is a form of abductive reasoning, not all abductive inferences are analogies. We return to this point later.

Similarity-based abduction is proposed as a theory for generating ideas as hypotheses or problem solutions. Abductive reasoning begins by activating a goal state characterized by a problem to be solved with no immediate solution found.

R.W. Schvaneveldt (✉)
Arizona State University, Mesa, AZ, USA
e-mail: schvan@asu.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_11,

Essentially, no available solution means that none are directly associated with the problem. However, a process of spreading activation would lead to the activation of other ideas related to the problem. Over time, continuing to think about the problem or engaging in still other activities would lead to the activation of other ideas together with patterns of connections among the ideas. Interconnections among the activated ideas could lead to an enhancement of the connections of ideas to the elements of the problem in two ways. First, activation among the connections could simply increase the activity in existing weak links between the problem and other ideas. Second, indirect connections of between newly activated ideas and the problem could be detected by means of similar patterns of connections. Such newly activated ideas might be indirectly or implicitly related to the problem. These new promoted weak connections and newly identified indirect connections provide links to potential solutions to the problem. They constitute potential hypotheses.

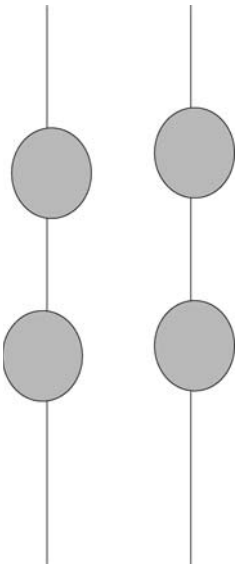
Developing models of similarity-based abduction involves developing methods of generating activation of ideas on the basis of activation of existing connections among ideas. Examples of such methods can be found in GeneRanker (Gonzalez, Uribe, Tari, Brophy, & Baral, 2007), Hyperspace Analog of Language or HAL (Burgess, Livesay, & Lund, 1998), Latent Semantic Analysis or LSA (Landauer & Dumais, 1997), and Random Indexing (Kanerva, Kristofersson, & Holst, 2000). Cohen (2008) has shown how identifying new connections can lead to novel hypotheses concerning potential treatments for medical conditions. Also, developing tools to assist users in identifying fruitful new ideas pertinent to hypothesis discovery and problem solving requires generating possible ideas, ranking the ideas, and providing informative displays of connections for users to examine and evaluate for their potential utility. Examples of models and tools are also presented in the chapter.

11.1.1 Abductive Reasoning

C. S. Peirce wrote extensively about logic and scientific method. Several important pieces were published in 1940 under the editorship of Justus Buchler (Peirce 1940a, 1940b). Peirce proposed that there were three essential types of reasoning including the familiar deductive and inductive reasoning. The testing, confirming, and rejecting of hypotheses is covered by deduction and induction. In contrast with many logicians, Peirce also thought there was a logic underlying the origin of new hypotheses. He called this logic variously “abduction”, “retroduction”, and “hypothesis” in his writings over the years. The kind of reasoning he envisions proceeds something like the following:

I make some observations (O) that are surprising, unusual, or puzzling in some way. It occurs to me that if a particular hypothesis (H) were true, then O would follow as a matter of course. In other words, H implies O so we could say that H explains O. Thus, H is plausible and should be considered further. Abductive reasoning is illustrated by Fig. 11.1.

Fig. 11.1 What is this?



Consider Fig. 11.1 to be a set of observations (O). Now ask, "What is this?" or "How could these observations be explained?" Now we are seeking hypotheses (H) that would explain the diagram (O). We might come up with such conjectures as:

- H1: "It's olives on toothpicks."
- H2: "It's barbeque spits with tomatoes."
- H3: "It's two pair of spectacles."
- Etc.

Notice that each of these conjectures (Hi) has the property that the O would follow if H were true. This is the abductive form of logic. Cast in the form of a syllogism, abductive logic would appear as in Table 11.1. In the example above, the arrangement of the lines and circles constitutes the observations (O). The various suggestions are potential hypotheses.

Table 11.1 Abductive inference

Major premise	O
Minor premise	If H then O
Conclusion	H is plausible

Obviously, this is not a deductive argument which requires that the conclusion necessarily follows from the premises. In abductive inference, H does not follow with certainty so the conclusion about H only reaches plausibility. The observations could have resulted from H so H is a reasonable conjecture about why O is as it is. As such, H deserves further consideration as a possible explanation of

Table 11.2 Inductive inference

Major premise	If H then O
Minor premise	O
Conclusion	H is confirmed

O. Abductive reasoning bears a strong similarity to inductive inference which is illustrated in Table 11.2.

Induction, too, does not carry certainty. In the deductive realm, H does not necessarily follow from the premises given so at best we can say that the observation confirms or supports the hypothesis. Finding confirming evidence for a hypothesis simply allows us to continue entertaining it, perhaps with increased confidence, but confirming evidence does not *prove* a hypothesis. The difference between abduction and induction is due to the temporal relations of the premises. The major premise precedes the minor premise in time so the hypothesis occurs as an explanation in abduction while the observations occur as a test of the hypothesis in induction. Tests of hypotheses do not always lead to confirmation, however, which leads to the third type of inference, deduction (modus tollens) as in Table 11.3.

Table 11.3 Deductive Inference

Major premise	If H then O
Minor premise	O is false (not O)
Conclusion	H is disproved (not H)

Finding that the predictions of a hypothesis fail to hold leads to the certain conclusion that the hypothesis is false. This asymmetry between induction and deduction was the basis of Popper's (1962) philosophy of science. Because disproving hypotheses is more conclusive than confirming them, Popper thought that scientists should make great efforts to disprove their favorite hypotheses rather than seeking more and more confirmatory evidence. This comparison of abduction, induction, and deduction helps to understand the relative roles of these logic forms in certain aspects of forming, confirming, and rejecting hypotheses. Let's return to the abductive case.

Because Peirce sought to characterize abduction as a form of logic, he sought some "rules" of abduction. Harman (1965) characterizes abduction as "inference to the best explanation." We are somewhat more comfortable thinking about abduction in terms of certain kinds of "constraints" rather than rules. For one thing, constraints operate to influence a process without completely determining it. Abduction is concerned with generating plausible ideas, not proving them, so relaxing the requirements of "rules of logic" to constraints seems more appropriate for a theory of abductive reasoning. Peirce proposed that certain conditions associated with testing hypotheses might figure into their value for scientific research. For example, other things being equal, hypotheses that are easy to test might be favored over those that require more time, effort, or money to test. This brings economic criteria to bear in selecting hypotheses.

There are several other criteria or constraints that affect our judgments about the quality of hypotheses. Returning to the example presented in Fig. 11.1, consider the hypothesis that the figure depicts a bear climbing up the other side of a tree. Do you like it? Most people like this suggestion more than the others advanced earlier. Why? One characteristic of the bear hypothesis is that it explains the *entire* figure. It explains not only the existence of the lines and the circles, but it explains the number of lines and circles. It explains why the lines are roughly parallel and why the circles are spaced in just the way they are. In other words, the bear hypothesis makes the most of the observations provided, perhaps even expanding the observations over what they were originally taken to be. Certain features that might have been considered arbitrary or coincidental become necessary and meaningful by virtue of the hypothesis. This is what makes a *good* hypothesis. We might call these constraints *coverage* and *fruitfulness*. Coverage refers to the extent of the coverage of the facts by the hypothesis. Fruitfulness refers to the information added by virtue of the interpretation afforded by the hypothesis.

Syllogisms are usually applied in the realm of deductive reasoning where we say that a syllogism is valid if the conclusion follows necessarily from the premises. When we add qualifications such as "plausible" to the conclusion, we may question the value of presenting the argument as a syllogism. The syllogistic form may tempt one to seek forms of certainty in the realm of abduction, but such an endeavor is fruitless because abduction does not yield certainty. A better quest may be to clarify what it means for a hypothesis to be plausible, and then identify methods that would help to achieve plausibility.

Peirce held that inquiry serves to relieve doubt. If one's beliefs are up to the task of accounting for experience, there is little motivation to examine those beliefs. Thus, surprise leads to a search for explanations to relieve doubt. Such explanations may lead to a change in beliefs either by adding new beliefs or by modifying established beliefs.

11.1.2 The Importance of Novelty

There are some reasons to think that there are different forms of abduction. Eco (1998) discusses the distinction in terms of the prior availability of the hypothesis. Two types of explanation differ in the status of the hypothesis before the abductive step. One form of explanation amounts to providing the general rule under which an observed case falls. This kind of reasoning occurs in medical diagnosis, for example, where a set of presenting symptoms (O) is "explained" by diagnosing the patient as carrying a certain "disease" (H). In this case, the disease was known as a possibility beforehand, and it provides an explanation of the symptoms in a particular case. This form of abduction amounts to determining which of a set of known explanations is to be adduced in a particular case. This might be called *selective* abduction.

A second form of abduction is at play when a new hypothesis is proposed as an explanation. This is where true creativity is at work. Historical examples in science are found in the Copernican heliocentric theory of the solar system, Pasteur's

germ theory of disease, Darwin's theory of evolution, and Einstein's relativity theory. More common examples are to be found in problem solving and other creative activities in which novel ideas are generated to solve problems. This might be called *generative* abduction.

In its various forms, abductive reasoning is actually quite commonplace. Peirce, himself, proposed that perception is fundamentally abductive inference. Sherlock Holmes, notwithstanding, detective work also seems to be better characterized as primarily abductive rather than deductive. Solving a crime involves finding an explanation for the facts of the case (O) by postulating a perpetrator (H). The degree to which detection involves selective as opposed to generative abduction is an open question. It may depend upon the details of a particular case.

By using a variety of constraints in the generation process, the distinction between generation and evaluation may be obscured, but it may be of value to distinguish between cases where abduction leads to new knowledge in a system as opposed to calling up old knowledge. In actuality, novelty may come in degrees as knowledge is modified by abductive inference. Stuart Kauffmann (2000) develops the interesting idea of the "adjacent possible" by which he means that a system may take on a number of novel states that are "adjacent" in some sense to the prior state of the system. Thus, novelty for a system is relative to the state of a system at a given point in time. Still, some state changes may represent larger steps than others. It may be useful in distinguishing different abductive procedures and/or abductive outcomes by the magnitude of the change brought about by the abduction.

Novelty can be introduced at several levels including revising or expanding existing concepts, creating new concepts and categories, forming new propositions in the form of hypotheses or laws, or applying a system of relations to a new situation as in reasoning by analogy. Often abductive reasoning is triggered by a failure of expectation or a conflict between current beliefs and new observations.

11.1.3 Approaches to Understanding the Generation of Hypotheses

Methods for generating new knowledge generally depend in some way on *similarity*. Similarity can take many forms and includes both superficial and relational similarity. New concepts and categories depend on similarity of features or functions. Often some deep similarity is revealed by creative thought as illustrated by Arthur Koestler (1990) in his book, *The Act of Creation,* with the concept of bisociation. Koestler points out how creativity in humor, art, and science often involves bringing two distinct ideas together to reveal a deep similarity. This is illustrated in the following joke:

> A woman observes her friend in apparent deep distress. She asks, "Vats da matta, Millie?" She responds, "Oye ve, it's our son Sammy; the doctor says he has an Oedipus complex." She replies, "Oh Oedipus, Schmedipus, vats da matta as long as he's a good boy and loves his mama."

Here the creative juxtaposition of two ways of loving one's mother (bisociation) produces a humorous result. The similarity of oedipal love and the love for one's mother can be exploited to bring together two quite incompatible ideas.

Similarity is also involved in creating new propositions in Coombs, Pfeiffer, and Hartley's (1992) e-MGR system by combining parts of older propositions located by similarity to the data to be modeled (see also Coombs & Hartley, 1987). Gentner (1983) uses relational similarity as the basis of identifying analogies in her structure mapping system. Case-based reasoning systems (Kolodner, 1993) are related to analogical reasoning systems that attempt to find analogous past cases to use to analyze a current case. Similarity is at the heart of finding cases.

Perhaps an alternative to the use of similarity to guide the formation of new knowledge units is the use of some random process. Genetic algorithms (Holland, 1992) provide a good example of the successful use of randomness in creating new units. Of course, there are other important constraints at work in genetic algorithms besides randomness. Total randomness would hold little value in the search for effective new knowledge. Selective reproduction according to "fitness" helps direct genetic algorithms toward more "fit" units. In his paper, *The Architecture of Complexity*, Simon (1962) suggested that evolution depends on the formation of stable intermediate forms. The following quote makes this point and relates the process of evolution to problem solving:

> A little reflection reveals that cues signaling progress play the same role in the problem-solving process that stable intermediate forms play in the biological evolutionary process. In problem solving, a partial result that represents recognizable progress toward the goal plays the role of a stable subassembly.

In other words, if fruitful steps toward finding a solution to a problem can be recognized, the probability of finding a solution by trial and error can be greatly increased over the probability of generating a complete solution all at once which may be so small as to be nearly impossible. The importance of stable intermediate forms is further analyzed in Simon's 1981 book, *The Sciences of the Artificial.* Several additional constraints at work in abductive reasoning will be discussed in a later section.

11.1.4 Optimizing Versus Satisficing

Several approaches to abduction have been proposed and analyzed by researchers in cognitive science (Aliseda, 2000; Charniak & Shimony, 1990; Fann, 1970; Flach & Kakas, 2000; Josephson & Josephson, 1994; Kakas, Kowalski, & Toni, 1998; Konolige, 1996; Levesque, 1989; Peng & Reggia, 1990; Poole, 2000; Prendinger & Ishizuka, 2005; Senglaub, Harris, & Raybourn, 2001; Shrager & Langley, 1990; Walton, 2004). Many of these researchers have investigated the computational complexity of various algorithms associated with abductive reasoning. Such algorithms often exhaustively search some space of possibilities to optimize

some measure. The algorithms are generally found to have complexity beyond reasonable computability which means they cannot scale up to the demands in most real applications. For example, Thagard and Verbeurgt (1998) showed that deciding about the consistency of a set of propositions is NP hard which is generally believed to be intractable. Bylander, Allemang, Tanner, and Josephson (1991) reached the same conclusion in another analysis of the computational complexity of abduction. Santos and Santos (1996) showed that linear programming leads to good solutions for some abduction problems using relaxation of integer program formulations. Thagard and Verbeurgt also report on a connectionist (neural net) approximation algorithm which gives good results in reasonable time. Reggia and Peng (1993) proposed a connectionist solution to diagnostic problem solving. Adaptive resonance theory (Carpenter & Grossberg, 2003) is still another approach to discovery in the framework of dynamical systems theory. Juarrero (1999) presents a compelling account of the connections between dynamical systems theory and intentional action. Such ideas appear to have a great deal to contribute to the development of theories of abductive reasoning.

Because abduction produces only plausible and not certain conclusions, it seems unnecessary to approach the problem through optimization. There are heuristic methods that arrive at very good solutions in reasonable time. Such methods seem particularly appropriate for abduction. Heuristic solutions amount to what Simon (1947) called satisficing, finding a satisfactory solution rather than an optimal one.

11.2 Generating and Evaluating Hypotheses

Several factors influence the evolution of hypotheses. To varying degrees, the factors affect the generation or the evaluation of hypotheses. Generation and evaluation are not necessarily completely distinct processes. There is likely continually interplay between generating ideas and evaluating them in the search for an acceptable hypothesis. The following section enumerates several of the factors at work in terms of constraints operating in abductive reasoning. The criteria are characterized as constraints, in part, because each criterion is defeasible, that is, useful abductions may result by discounting any of the criteria.

11.2.1 Constraints on Abduction

Although abductive reasoning does not carry the certainty of deduction, there are constraints on what characterizes good hypotheses. A general account of abduction can proceed by identifying the constraints satisfied by the abduction. Abduction systems can be analyzed in terms of the constraints they embody. Different prospective hypotheses can also be compared by the extent to which they meet the constraints. An ordering of the hypotheses by preference follows from such comparisons. An important avenue for research is to determine the proper weighting of the various constraints. *A principled method for varying the weighting of the constraints*

would produce a variety of hypotheses according to different assumptions. Here is a summary of some constraints to be considered as contributing to abductive reasoning.

- *The Observations.* Providing an explanation is a primary constraint on abduction. That the observations follow from the hypothesis is a first condition of plausibility of the hypothesis. At first blush, the observations appear to be primarily involved in evaluation as opposed to generation. However, the observations are also the starting point of the whole process. As discussed later, the observations also enter into similarity relations which are critical in generating potential abductive inferences.
- *Reliability of the Observations.* While observations provide primary constraints, the possibility of error in part or all of the observed data must also be considered. More reliable data should be weighted more heavily. If discounting some aspects of the data lead to a coherent account of the remaining data, the discounted data may be submitted to closer scrutiny.
- *Surprise.* Surprising or unexpected observations point to the need for a new hypothesis. When existing explanations of events fail to cover a newly observed event, abductive inference is called into play. While this is generally true, there may also be value in generating new hypotheses even while the current ones seem to be adequate to the task. Such hypotheses might provide for novel perspectives suggesting new ways to evaluate existing hypotheses.
- *Novelty of Hypotheses.* For observations to be considered surprising, it should be the case that ready explanations for the observations are not available. Thus, by this criterion, the novelty of a hypothesis counts in its favor. Novel hypotheses emphasize generation rather than a search among existing hypotheses.
- *Economics.* Hypotheses that are easier (less expensive) to test should be entertained before those that entail more difficult (more expensive) means of testing. This is one of the criteria suggested by Peirce in his work on abductive reasoning.
- *Parsimony.* Simpler hypotheses are preferred over more complex ones (Occam's razor). Parsimony would appear to be primarily an evaluative criterion, but it is also possible that simpler hypotheses would be easier to generate than more complex ones.
- *Aesthetics.* Beauty, elegance, symmetry, and appeal figure into the value of a hypothesis. Again, this constraint seems to be evaluative, but aesthetic factors could also influence certain characteristics of the hypotheses generated.
- *Plausibility and Internal Consistency.* Hypotheses consistent with each other and with background knowledge are preferred over ones that lead to contradictions. This constraint can also be seen to have both evaluative and generative dimensions. Generation might be expected to be strongly influenced by what is already known, and the acceptability of a generated hypothesis may well affect the likelihood of its survival.
- *Explanatory Power (Consilience).* This criterion is primarily evaluative in the sense that it applies to a hypothesis in hand where its explanatory power can be seen. There are various aspects of consilience such as:

  - *Coverage.* The extent to which a hypothesis explains the details of observations including incidental, in addition to central, details – the greater the coverage, the better the hypothesis.
  - *Fruitfulness.* The information added to the observations by virtue of the interpretation afforded by a hypothesis including providing meaning to features that were previously seen as incidental – the more information added, the better the hypothesis.
  - *Organization of the observations.* Hypotheses that reveal connections among the observations that were not obvious before are of particular value. For example, a hypothesis may suggest related clusters of observations.

- *Pragmatics.* Pragmatics emphasizes the influence of goals and the context in which reasoning occurs. Goals and context are additional constraints on abductions. Pragmatics can operate both to direct generation and evaluation of hypotheses.
- *Analogy Formation.* Analogy formation works by finding sets of relations found in a source domain that can be applied to a target domain. An often cited example is the analogy between the solar system and an atom where parallels can be drawn between the sun and the nucleus of an atom and between planets and electrons of an atom. Once an analogy is drawn on the basis of known relations, characteristics from the source domain can be hypothesized to hold for the target domain.
- *Random Variation.* Hypotheses may be found by some random variation in older hypotheses. A system that constantly seeks for better hypotheses might be expected to occasionally find particularly good hypotheses that had not been considered before. Constraints in addition to randomness are probably necessary. Random variation alone is unlikely to lead to good results. Genetic algorithms employ randomness with other constraints. Genetic algorithms come primarily from the work of John Holland (1992, 1995, 1998). These methods are inspired by genetic reproduction where such processes as crossover and mutation lead to increases in "fitness" of individuals in populations. The methods are used in a variety of optimization problems. Genetic algorithms include a degree of randomness in the selection of mates and in mutation. Mate selection is also controlled by fitness which constrains the influence of random selection.
- *Similarity and Associations.* Similarity at various levels is a weak constraint on abductive reasoning, but similarity, at some level, is often involved in suggesting abductive inferences. Similarity may guide the search for commonalities among features, objects, and rules. In analogical reasoning, similarity of relations is a critical feature. Koestler (1990) proposed bisociation as a prominent feature of creative endeavors. Bisociation is the bringing together of unrelated ideas in a way that draws out a relation between them. In analogical reasoning, patterns of similarities provide constraints on abduction, but with analogies, the similarity is found at the level of relations – similar relations suggest analogies (Gentner, 1983; Gentner, Holyoak, & Kokinov, 2001; Gentner & Markman, 1997; Holyoak & Thagard, 1995). In a study of insight in problem solving, Dayton,

Durso, and Shepard (1990) found that critical associative connections underlying the solution of the problem often appeared in Pathfinder networks before the problem was solved suggesting that arriving at a solution may be mediated by establishing the critical connections. A similar process may be at play in some cases of abductive inference.

11.2.2 Similarity in Abductive Inference

Novel abductive inferences are not strongly associated with the phenomena to be explained. Rather such strong associations would make the inference obvious rather than novel or surprising. However, similarity or association of some kind may well be involved. The similarity may be indirect or the association weak, but the connection is often obvious in hindsight. Bruza, Cole, Song, and Bari (2006) discuss the value of identifying indirect associations in discovering a novel medical treatment involving the use of fish oil to treat Raynaud's Syndrome (intermittent blood flow in the extremities). Swanson (1986, 1987) proposed the treatment by connecting ideas from two unconnected literatures regarding the syndrome and dietary fish oil. Bruza et al. suggest that such connections can be generated from textual sources by identifying concepts (terms) that do not occur together, but they do tend to co-occur with the same other concepts. The HAL system (Burgess et al., 1998; Lund & Burgess, 1996) and the LSA method (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) lead to identifying high degrees of similarity for terms that have similar patterns of co-occurrence with other terms in the database. They use a similarity measure based on the cosine of the angle between the vectors for each of the terms where the vectors represent the co-occurrence patterns of each of the terms. Bruza et al. show the connection of fish oil and Raynaud's Syndrome discovery using such methods.

There is a longstanding interest in the role of geometric models or conceptual spaces in cognition (Gärdenfors, 2000; Kruskal, 1964a, 1964b; Shepard, 1962a, 1962b; Widdows, 2004). Gärdenfors proposes an important role for a geometric level of representation, distinct from both low-level connectionist processes and higher level symbolic representations. An important role of the geometric level is to provide a basis for establishing similarity relations by virtue of the relations among concepts in conceptual space. Much of this work sees concepts as corresponding to regions of low-dimensional conceptual space. In other models, such as HAL and LSA, concepts correspond to vectors in high-dimensional conceptual space. Both views support the idea that similarity can be derived from spatial information.

The use of cosine measures on vectors representing the distribution of terms in text provides a way of assessing similarities between terms. Such similarity reflects both the co-occurrence of terms and similarities in the patterns of co-occurrences across all of the terms in a corpus. By eliminating pairs of terms that occur together in the corpus, one can focus on "indirect" similarity, similarity that derives from

similar patterns of co-occurrence rather than direct co-occurrence. These indirect similarities may suggest possible abductive inferences. Not all indirect similarities can be expected to constitute such inferences. Synonyms rarely occur together in text, but they could be expected to have similar patterns of co-occurrence with other terms. While these would not qualify as novel inferences, they should be relatively easy to identify. Often, we can characterize the *type* of thing that would qualify as a useful inference. For example, if we are looking for possible treatments of a disease or syndrome, only indirect similarity with things that could be treatments would be entertained as potential abductive inferences pertinent to the disease or syndrome. At this stage of our work, we rely on human judgment to determine which, if any, of the terms with indirect similarity to a target of interest constitute interesting potential abductive inferences.

For human evaluation, it is useful to view collections of terms indirectly related to a target term as Pathfinder networks (McDonald, Plate, & Schvaneveldt, 1990; Schvaneveldt, Dearholt, & Durso, 1988; Schvaneveldt, Durso, & Dearholt, 1989; Schvaneveldt, 1990) which depict patterns of relationships among the terms via patterns of links among the terms. Such networks show the strongest similarities among the terms, often revealing interesting paths among the terms as a way of identifying intermediate relationships of interest in addition to showing terms of interest.

11.3 Random Vectors and Pathfinder Networks as Aids for Abduction from Text

In this section, we present some findings from our work on developing computational tools to support abductive inference from textual corpora. Here we provide only a brief look at the work which will appear in more detail in Cohen, Schvaneveldt, and Widdows (2009).

The ability of methods such as LSA and HAL to find meaningful connections between terms (such as “raynaud”, “fish”, and “oil”) that do not co-occur directly in any text passage can be considered as a sort of inference. Landauer and his colleagues describe this as an “indirect inference” and estimate that much of LSA’s human-like performance on tasks such as the TOEFL synonym test relies on inferences of this sort (Landauer & Dumais, 1997). In Fig. 11.2 we illustrate the ability of LSA to identify interesting similarities. These indirect inferences are abductive in nature. They arise from similar patterns of occurrence across the corpus in the absence of co-occurrence.

This figure shows a Pathfinder network (PFNET) of the 20 *nearest indirect neighbors* of the term “beatlemania” in a semantic space derived from the Touchstone Applied Sciences (TASA) corpus using the General Text Parser software package (Giles, Wo, & Berry, 2003), obtained by screening out all terms that occur directly with the term “beatlemania” in any document in this corpus. The links in the PFNET are determined by the cosine similarities between all pairs of the terms, but after the

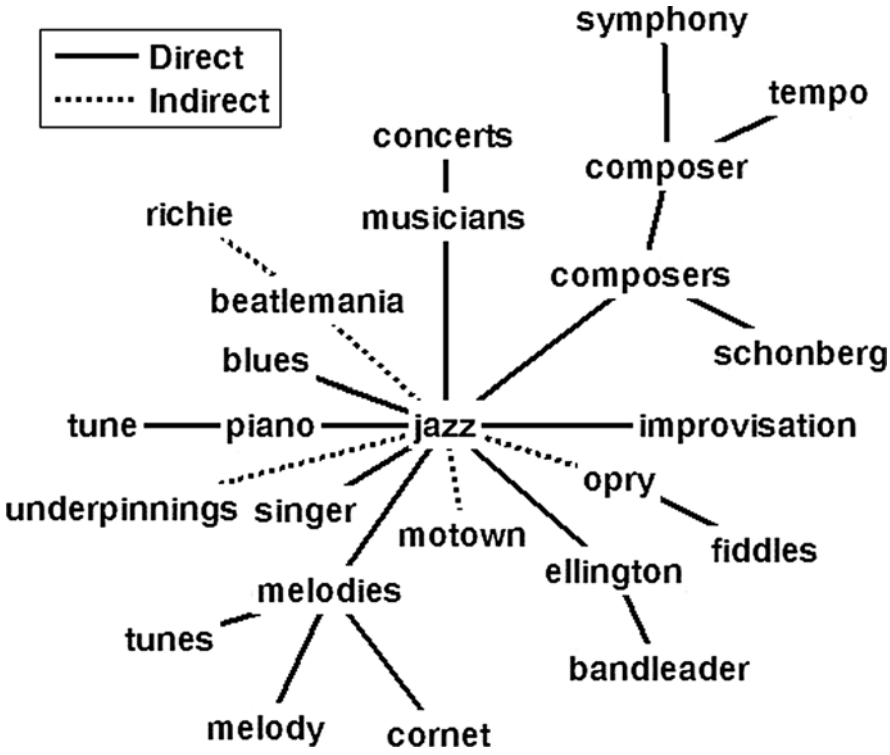


Fig. 11.2 PFNET of nearest indirect LSA neighbors of "beatlemania"

application of Pathfinder network scaling only those links representing the most significant pairwise similarities are preserved. [1]Dashed links illustrate indirect connections between terms that do not co-occur directly in any document. Many of these connections make intuitive sense, as they refer to musical forms and performers more commonly associated with musical genre other than pop. The figure also reveals a number of other interesting indirect neighbors of the term "jazz", such as "motown" and (the grand ′ol) "opry".

PFNETs preserve the most significant links between nodes in a network, and consequently reveal the semantic structure underlying this group of near neighbors, such as the western classical music related connection between "Schonberg", "composers", "composer" and "symphony". It is also possible to use PFNETs to

[1]The PFNETs presented here were all computed with parameters, $r = \infty$ and $q = n - 1$, where n is the number of nodes in the network. The links preserved with these parameters consist of the union of the links in all minimum spanning trees or, in terms of similarities, the union of the links in all maximum spanning trees. The sum of the similarities associated with the links in such trees is the maximum over all possible spanning trees. See "Pathfinder Networks" in Wikipedia for additional information.

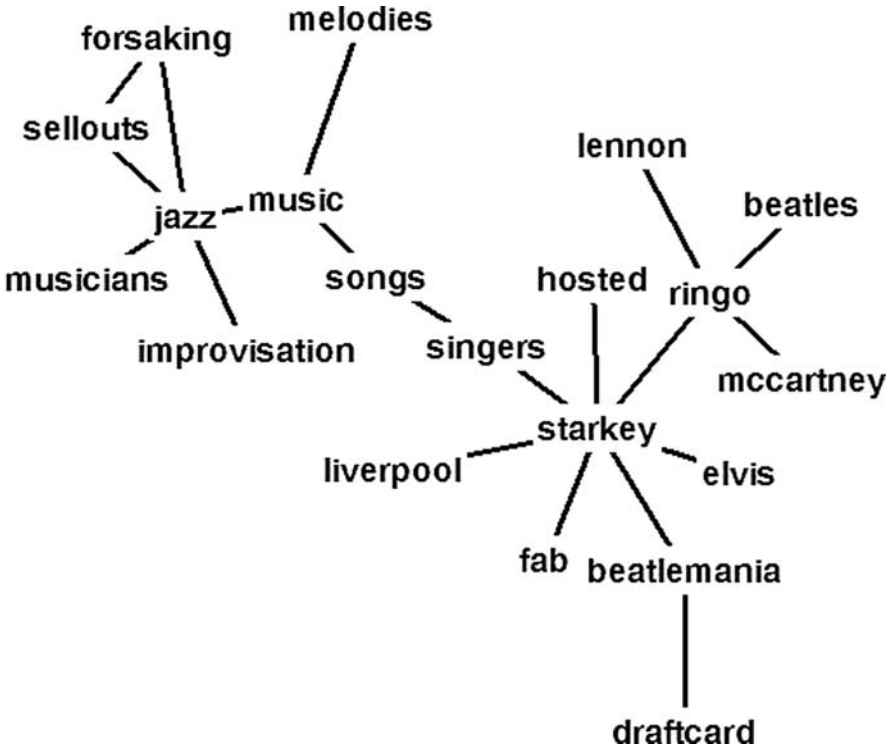


Fig. 11.3 PFNET of the nearest RRI neighbors of "beatlemania + jazz"

attempt to uncover the similarities that lead to an interesting indirect connection. For example, if we combine the representations of "jazz" and "beatlemania" by simply adding their corresponding vectors together and generate the nearest neighbors of this combined representation, we derive the PFNET shown in Fig. 11.3.

Pathfinder has revealed a chain of significant links leading from "jazz" through "music" (jazz is a musical genre), "songs" (a musical form), "singers" (of songs) and "starkey" to beatlemania. This track of relations between "jazz" and "beatlemania" might be seen as a form of bisociation (Koestler, 1990) where the intermediates explain the indirect connection. Starkey here refers to Richard Starkey, the name on the birth certificate of Beatles member Ringo Starr. Although Starr was the group's drummer, he was also a backing vocalist as well as lead vocalist on several well-known tunes such as "Yellow Submarine" and "With a Little Help from My Friends".

Figure 11.3 was generated using as a basis semantic distances estimated by the Random Indexing model (Kanerva et al., 2000; Karlgren & Sahlgren, 2001) using the Semantic Vectors Package (Widdows & Ferraro, 2008). Random Indexing is similar in concept and underlying assumptions to Latent Semantic Analysis in that terms are represented in a vector space according to their distribution across a

large set of documents. However, unlike LSA, Random Indexing does not depend upon computationally demanding methods of dimension reduction to generate a condensed vector representation for each term. Rather, it achieves this end by projecting terms directly into a vector space of predetermined dimensionality (usually >1,000) by assigning to each document a randomly generated index vector in this subspace that is close-to-orthogonal to every other assigned index vector. While more investigation is needed to determine which aspects of LSA's performance this method as able to reproduce, initial investigations show it is possible to use this method of dimension reduction without degrading the model's performance on synonym tests (Kanerva et al., 2000). Unlike LSA, the model scales comfortably to large corpora, as we illustrate below with networks derived from the MEDLINE corpus of abstracts. In our studies, we have found Random Indexing using a term-document approach to be somewhat limited in its ability to generate meaningful indirect inferences. Consequently, the remaining diagrams, with the exception of the "thrombophilia" example, were produced using Reflective Random Indexing (RRI), a method that creates term vectors by an iterative construction (Cohen et al., under review).

The PFNET in Fig. 11.4 was created with a similar approach, however, in this case the semantic distance between terms was generated from the abstracts of

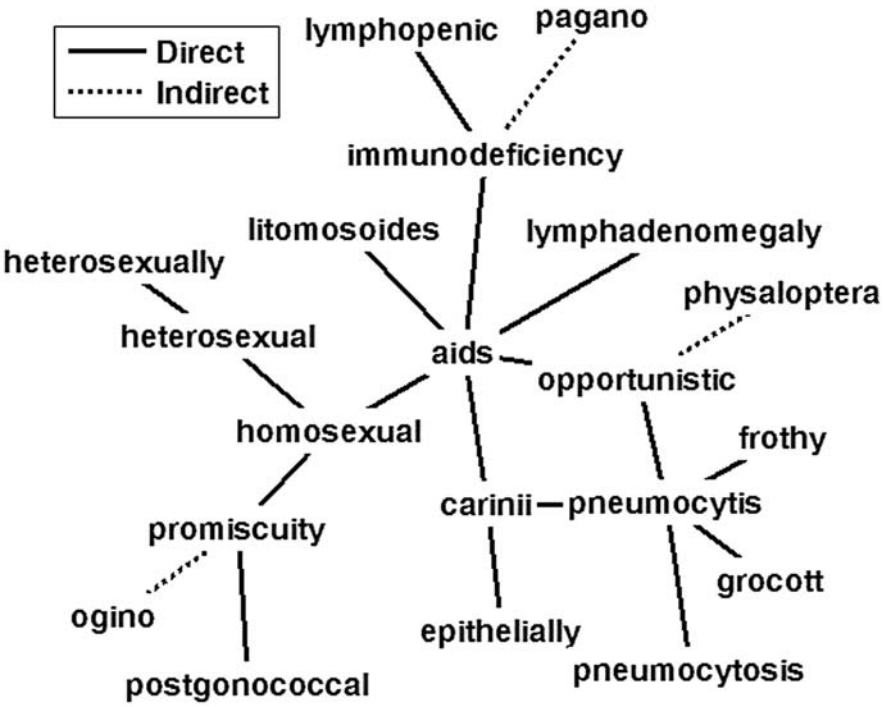


Fig. 11.4 PFNET of nearest neighbors of "pneumocystis + promiscuity"

articles in the MEDLINE database of biomedical literature occurring from 1980 to 1986. An indirect connection between the terms "pneumocystis" (*pneumocystis carinii pneumonia* occurs in immunocompromised patients) and "promiscuity" (promiscuity amongst the homosexual community was implicated in the transmission of the recently discovered Acquired Immune Deficiency Syndrome) was retrieved among the 20 nearest indirect neighbors of the term "pneumocystis". Figure 11.4 shows the 20 nearest neighbors of the combined representation of "pneumocystis" and "promiscuity" using the RRI method of creating the index.

Again, the PFNET reveals a plausible line of reasoning connecting these two terms. Pneumocystis is connected through "carinii", "aids", and "homosexual" to promiscuity. This PFNET illustrates an inferred relationship between pneumocystis carinii pneumonia and promiscuity, which was not explicitly stated in any MEDLINE abstract used to build this model. Interestingly, the "ogino" indirectly connected to "promiscuity" in this diagram refers to Kyusaku Ogino, who measured the fertile period of the female menstrual cycle. While Ogino did not believe this method could be used as a reliable form of contraception, the Rhythm Method of contraception is nonetheless referred to as the "Ogino Method" in the occasional MEDLINE record.

Another interesting indirect connection to emerge from the TASA corpus through Random Indexing is an association between Picasso and impressionism shown in Fig. 11.5. Deriving a PFNET from the combined vector for "picasso" and "impressionism" reveals a pathway from Picasso through "cubism," "cubist," and "carafe," an important cubist work of Picasso, to "manet" to "impressionists" to "impressionism." Manet's work, particularly *Le déjeuner sur l'herbe,* is considered to by many

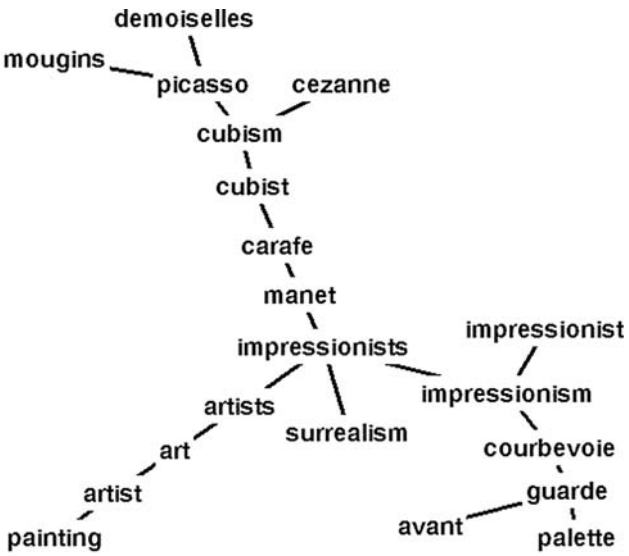


Fig. 11.5 PFNET of nearest RRI neighbors of "picasso + impressionism"

critics to have exerted a seminal influence on the evolution of cubism through the work of Picasso (and George Braque). Picasso also painted variations of several works of Manet, including *Le déjeuner* which at the time of this writing is featured in the exhibition "*Picasso/Manet: Le déjeuner sur l'herbe*" at the Musée d'Orsay in Paris.

We have further investigations underway to produce and evaluate indirect connections obtained from the MEDLINE database. An interesting indirect association was observed between "spongiform" and "cannibalism." This association was noted in a subset of MEDLINE abstracts occurring between 1980 and 1985. The spongiform encephalopathies, such as bovine spongiform encephalopathy in cattle (BSE, aka Mad Cow Disease), scrapie in sheep and Creutzfeld-Jacob Disease (CJD) in humans are degenerative neurological disease that are currently thought to be caused by prions, infectious protein agents that replicate in the brain. While Prusiner's prion hypothesis (Prusiner, 1982) was contested at this time, he was later awarded the Nobel Prize for this work. "Kuru" is an encephalopathy that was transmitted by cannibalistic practice in New Guinea.

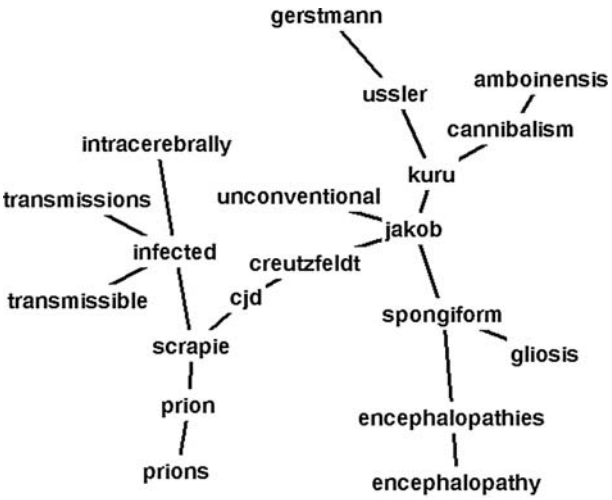


Fig. 11.6 PFNET of nearest RRI neighbors of spongiform + cannibalism

A PFNET for the combined terms "spongiform + cannibalism" is shown in Fig. 11.6. This PFNET reveals a pathway (CJD via "kuru" to "cannibalism"). This pathway reveals a plausible line of reasoning connecting cannibalism through kuru to other spongiform encephalophathies, and the prion hypothesis which was first proposed in the context of scrapie. A similar line of reasoning was explored by Prusiner during the course of his research, in which he developed an experimental model of the transmission of scrapie using the natural cannibalistic activity of hamsters (Prusiner, Cochran, & Alpers, 1985). No direct connection between "prion" and "kuru" exists in the 1980–1985 corpus of abstracts, and while the notion that kuru may also be caused by a prion protein is unlikely to have been novel at the

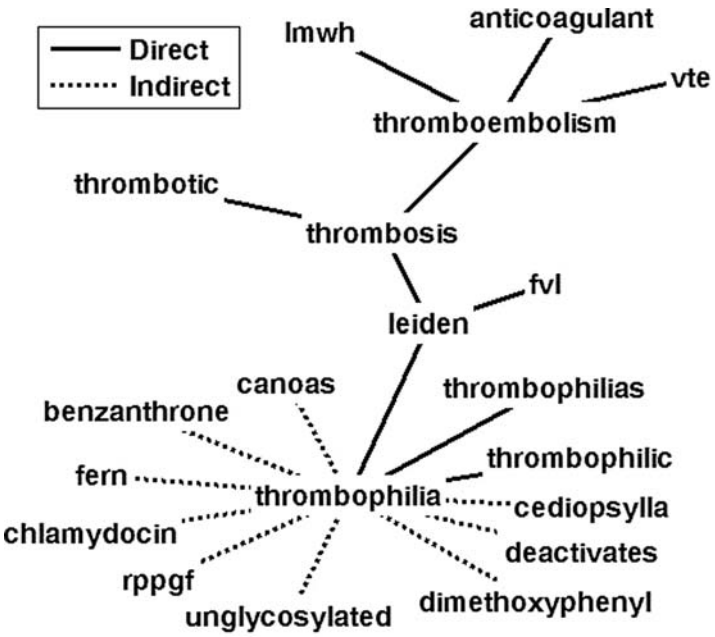


Fig. 11.7 PFNET of nearest RI neighbors of "thrombophilia"

time, this example provides an interesting illustration of how the exploration of one meaningful indirect inference can reveal another.

The discovery of a hypothesis is illustrated in Fig. 11.7, the term "rppgf" was returned as a near neighbor to a cue term "thrombophilia" (Cohen, 2008). RPPGF is the protein sequence Arg-Pro-Pro-Gly-Phe, the sequence of an inhibitor of platelet aggregation that *could* be therapeutically useful in thrombophilia. However, a PubMed search (conducted on June 6, 2008) for "rppgf AND thrombophilia" does not retrieve any results. Further examination of the MEDLINE corpus shows that these terms do not directly co-occur in any of the abstracts in MEDLINE. However, despite this lack of direct co-occurrence, the indirect similarity between these two terms in the RI space derived from these abstracts was sufficient for "rppgf" to be among the nearest neighbors of "thrombophilia." Discoveries of this sort are the focus of our present research including work on using random indexing methods to encode and retrieve the types of relations that exist between concepts (Cohen, Schvaneveldt, & Rindflesch, 2009)

11.4 Predicting "Discoveries"

If the indirect similarity of terms is a harbinger of an undiscovered relationship between the concepts corresponding to the terms, we might expect that indirect

neighbors would tend to become direct neighbors over time. By indirect neighbors, we mean items that are similar to a target item but do not co-occur with the target. Direct neighbors are similar items that do co-occur with the target. Using the MEDLINE database, we assessed the proportion of nearest indirect neighbors between 1980 and 1986 that became direct neighbors after 1986 ("discoveries"). In this experiment we investigated two different methods for creating term vectors, standard random indexing (RI) developed by Kanerva et al. (2000) and a new reflective random indexing (RRI) method adjusted to improve indirect similarity (Cohen et al., under review). The reflective method involves iteratively creating term and document vectors starting with random vectors. The full MEDLINE index of abstracts contains 9,003,811 documents and 1,125,311,210 terms of which 3,948,887 terms are unique. Our index consists of about 300,000 unique terms which excludes terms occurring less than ten times and terms that contain nonalphabetic characters.

Two thousand (2,000) target terms were randomly selected, and the nearest indirect neighbors (NINs) of each of the targets were found in the database between 1980 and 1986. Then each of the indirect neighbors was checked to determine whether it co-occurs with its target after 1986. The ones that did co-occur were dubbed "discoveries." The results are shown in Fig. 11.8.

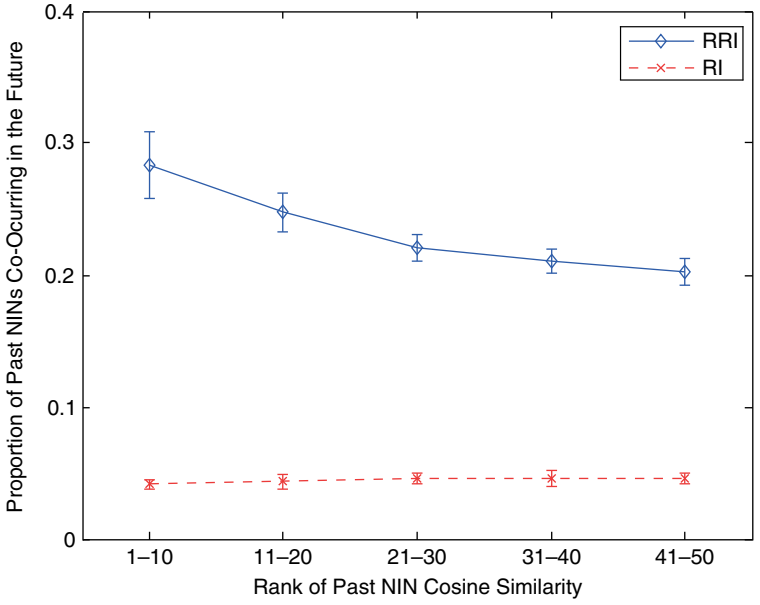


Fig. 11.8 Future "discoveries" from past indirect neighbors (NIN)

The RI index did not produce many "discoveries," a maximum of 4.5% while the rate of discoveries with the RRI index reached 28.4% for the ten nearest indirect neighbors. The difference between the two indexes is statistically significant,

t (1999) = 53.11, $p < 0.0001$. There was also a significant decrease in the rate for the RRI index from 28.4 to 22.0% for nearest indirect neighbors 11–50, t (1999) = 17.81, $p < 0.0001$. This decrease shows that stronger indirect similarity leads to a greater rate of "discoveries" which suggests that the indirect similarity measure does reflect the importance of the relation between the terms. For the less successful RI index, decreasing similarity leads to a slightly increased rate of "discoveries," 4.5% compared to 4.2%.

These findings suggest that indirect similarity may well be a precursor to the future realization of the relations between concepts. Clearly, there is more work to be done to explore and evaluate these findings. At this point, we find some clear support for continuing this line of work.

11.5 Conclusions

New ideas may be sparked by noticing indirect similarities. The spark is essential in leading to novel possibilities in the abductive reasoning found in problem solving and hypothesis generation. We have shown the value of tracing indirect similarities through examples and an analysis of the fate of indirectly similar terms. Our initial efforts at understanding the nature and role of indirect similarity encourage us to continue to pursue the development of this approach and the tools to support it. Although the efforts reported in this chapter have concentrated on finding indirect similarities in textual corpora, it could be argued that analogous processes operate in cognition generally. Exciting work lies ahead to elaborate on such possibilities.

References

Aliseda, A. (2000). Abduction as epistemic change: A Peircean model in artificial intelligence. In P. Flach & A. Kakas (Eds.), *Abduction and induction: Essays on their relation and integration*. Boston, MA: Kluwer Academic Publishers.

Bruza, P., Cole, R., Song, D., & Bari, Z. (2006). Towards operational abduction from a cognitive perspective. In L. Magnani (Ed.), *Abduction and creative inferences in science. Logic journal of the IGPL*. Oxford: Oxford University Press.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*(2–3), 211–257.

Bylander, T., Allemang, D., Tanner, M. C., & Josephson, J. R. (1991). The computational complexity of abduction. *Artificial Intelligence*, *49*, 25–60.

Carpenter, G. A., & Grossberg, S. (2003). Adaptive resonance theory. In M. A. Arbib (Ed.), The *Handbook of brain theory and neural networks* (2nd ed., pp. 87–90). Cambridge, MA: MIT Press.

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 106–111). Boston.

Cohen, T. (2008). Exploring MEDLINE space with random indexing and Pathfinder networks. *American Medical Informatics Association Symposium*, Washington, DC.

Cohen, T., Schvaneveldt, R. W., & Rindflesch, T. C. (2009). Predication-based semantic indexing: Permutations as a means to encode predications in semantic space. *American Medical Informatics Association Symposium*, San Francisco.

Cohen, T., Schvaneveldt, R., & Widdows, D. (2009). Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics,* Sep. 15, [Epub ahead of print]

Coombs, M., & Hartley, R. (1987). The MGR algorithm and its application to the generation of explanations for novel events. *International Journal of Man-Machine Studies*, *20*, 21–44.

Coombs, M. J., Pfeiffer, H. D., & Hartley, R. T. (1992). e-MGR: An architecture for symbolic plasticity. *International Journal of Man-Machine Studies*, *36*, 247–263.

Dayton, T., Durso, F. T., & Shepard, J. D. (1990). A measure of the knowledge organization underlying insight. In R. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.

Eco, U. (1998). Hooves, horns, insteps: Some hypotheses on three types of abduction. In U. Eco & T. A. Sebeok (Eds.), *The sign of three*. Bloomington, IN: Indiana University Press.

Fann, K. T. (1970). *Peirce's theory of abduction*. The Hague: Martinus Nijhoff.

Flach, P. A., & Kakas, A. C. (2000). *Abduction and induction: Essays on their relation and integration*. Boston, MA: Kluwer Academic Publishers.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56.

Giles, J. T, Wo, L., & Berry, M. W. (2003). GTP (General Text Parser) Software for Text Mining. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 455–471). Boca Raton: CRC Press.

Gonzalez, G., Uribe, J. C., Tari, L., Brophy, C., & Baral, C. (2007). Mining gene-disease relationships from biomedical literature: Incorporating interactions, connectivity, confidence, and context Measures. *Proceedings of the Pacific Symposium in Biocomputing*. Maui, Hawaii.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*, 88–95.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Cambridge, MA: MIT Press.

Holland, J. H. (1995). *Hidden order: How adaptation builds complexity*. New York: Addison Wesley.

Holland, J. H. (1998). *Emergence: From chaos to order*. Cambridge, MA: Perseus.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps.* Cambridge, MA: MIT Press.

Josephson, J. R., & Josephson, S. G. (Eds.). (1994). *Abductive inference: Computation, philosophy, technology*. New York: Cambridge University Press.

Juarrero, A. (1999). *Dynamics in action: Intentional behavior as a complex system*. Cambridge, MA: MIT Press.

Kakas, A. C., Kowalski, R. A., & Toni, F. (1998). The role of abduction in logic programming. In D. M. Gabbay, C. J. Hogger, & J. A. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming: Vol. 5, Logic programming*. Oxford: Clarendon Press.

Kanerva, P., Kristofersson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In L. R. Gleitman & A. K. Josh (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (p. 1036). Mahwah, NJ: Erlbaum.

Karlgren, J., & Sahlgren, M. (2001). From words to understanding. In Y. Uesaka, P. Kanerva & H. Asoh (Eds.), *Foundations of real-world intelligence* (pp. 294–308). Stanford, CA: CSLI Publications.

Kauffmann, S. (2000). *Investigations*. Oxford: Oxford University Press.

Koestler, A. (1990). *The act of creation*. New York: Penguin.

Kolodner, J. (1993). *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Konolige, K. (1996). Abductive theories in artificial intelligence. In B. Brewka (Ed.), *Principles of knowledge representation*. Stanford, CA: CSLI Publications.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115–129.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K, Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2–3), 259–284.

Levesque, H. J. (1989). A knowledge-level account of abduction. In *Proceedings of the International Conference on Artificial Intelligence.* Detroit, MI.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, *28*(2), 203–208.

McDonald, J. E., Plate, T. A., & Schvaneveldt, R. W. (1990). Using Pathfinder to extract semantic information from text. In R. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 149–164). Norwood, NJ: Ablex.

Peirce, C. S. (1940a). Abduction and induction. In J. Buchler (Ed.), *Philosophical writings of Peirce*. New York: Routledge.

Peirce, C. S. (1940b). Logic as semiotic: The theory of signs. In J. Buchler (Ed.), *Philosophical writings of Peirce*. New York: Routledge.

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem-solving*. New York: Springer-Verlag.

Poole, D. (2000). Abducing through negation as failure: Stable models within the independent choice logic. *The Journal of Logic Programming*, *44*, 5–35.

Popper, K. (1962). *Conjectures and refutations*. London: Routledge.

Prendinger, H., & Ishizuka, M. (2005). A creative abduction approach to scientific and knowledge discovery. *Knowledge-Based Systems*, *18*(7), 321–326.

Prusiner, S. B. (1982). Novel proteinaceous infectious particles cause scrapie. *Science*, *216*(4542), 136–144.

Prusiner, S. B., Cochran, S. P., & Alpers, M. P. (1985). Transmission of scrapie in hamsters. *The Journal of Infectious Diseases*, *152*(5), 971–978.

Reggia, J. A., & Peng, Y. (1993). A connectionist approach to diagnostic problem solving using causal networks. *Information Sciences*, *70*, 27–48.

Santos, E., Jr., & Santos, E. S. (1996). Polynomial solvability of cost-based abduction. *Artificial Intelligence*, *86*, 157–170.

Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Schvaneveldt, R. W., Dearholt, D. W., & Durso, F. T. (1988). Graph theoretic foundations of Pathfinder networks. *Computers and Mathematics with Applications*, *15*, 337–345.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 24, pp. 249–284). New York: Academic Press.

Senglaub, M., Harris, D., & Raybourn, E. M. (2001). *Foundations for reasoning in cognition-based computational representations of human decision making* (Tech. Rep. SAND2001-3496). Albuquerque: Sandia National Laboratories.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with unknown distance function Part I. *Psychometrika, 27*, 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with unknown distance function Part II. *Psychometrika, 27*, 219–246.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation.* Palo Aalto, CA: Morgan Kaufmann.

Simon, H. A. (1947). *Administrative Behavior*. New York: The Free Press.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*, 467–482.

Simon, H. A. (1981). *The sciences of the artificial* (2nd Ed.). Cambridge, MA: MIT Press.
Swanson, D. R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18.
Swanson, D. R. (1987). Two medical literatures that are logically but not bibliographically related. *Journal of the American Society for Information Science*, *38*, 228–233.
Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science*, *22*, 1–24.
Walton, D. N. (2004). *Abductive reasoning*. Tuscaloosa, Alabama: University of Alabama Press.
Widdows, D. (2004). *Geometry and meaning*. Stanford, CA: CSLI Publications.
Widdows, D., & Ferraro, K. (2008). Semantic vectors: A scalable open source package and online technology management application. *Sixth International Conference on Language Resources and Evaluation (LREC 2008).*

Chapter 12
Scope of Graphical Indices in Educational Diagnostics

Dirk Ifenthaler

12.1 Introduction

Knowledge representation is a key concept in psychological and educational diagnostics. Thus, numerous models for describing the fundamentals of knowledge representation have been applied so far. The distinction which has received the most attention is that between declarative ("knowing that") and procedural ("knowing how") forms of knowledge (see Anderson, 1983; Ryle, 1949). Declarative knowledge is defined as factual knowledge, whereas procedural knowledge is defined as the knowledge of specific functions and procedures for performing a complex process, task, or activity. Closely associated with these concepts is the term cognitive structure, also known as knowledge structure or structural knowledge (Jonassen, Beissner, & Yacci, 1993), which is conceived of as the manner in which an individual organizes the relationships between concepts in memory (Ifenthaler, Masduki, & Seel, 2009; Shavelson, 1972). Hence, an individual's cognitive structure is made up of the interrelationships between concepts or facts and procedural elements.

Further, it is argued that the order in which information is retrieved from long-term memory will reflect in part the individual's cognitive structure within and between concepts or domains. When compared to that of a novice, a domain expert's cognitive structure is considered to be more tightly integrated and to have a greater number of linkages between interrelated concepts. There is thus immense interest on the part of researchers and educators to diagnose a novice's cognitive structure and compare it with that of an expert in order to identify the most appropriate ways to bridge the gap (Ifenthaler, Masduki, et al., 2009; Ifenthaler & Seel, 2005). By diagnosing these structures precisely, even partially, the educator comes closer to influencing them through instructional settings and materials.

However, it is not possible to measure these internal representations of knowledge directly. Additionally, it is argued that different types of knowledge require

D. Ifenthaler (✉)
Albert-Ludwigs-University, Freiburg, Germany
e-mail: ifenthaler@ezw.uni-freiburg.de

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_12,

different types of representations (Minsky, 1981). Therefore, we argue that it is necessary to identify economic, fast, reliable, and valid techniques to elicit and analyze cognitive structures (Ifenthaler, 2008a). In order to identify such techniques, one must be aware of the complex processes and interrelationships between internal and external representations of knowledge. Seel (1991, p. 17) describes the function of internal representation of knowledge by distinguishing three zones – the object zone W as part of the world, the knowledge zone K, and the zone of internal knowledge representation R. As shown in Fig. 12.1, there are two classes of functions: (1) f_{in} as the function for the internal representation of the objects of the world (internalization) and (2) f_{out} as the function for the external re-representation back to the world (externalization). Neither class of functions is directly observable. Hence, a measurement of cognitive structures is always biased as we are not able to more precisely define the above-described functions of internalization and externalization (Ifenthaler, 2008a). Additionally, the possibilities of externalization are limited to a few sets of sign and symbol systems (Seel, 1999b) – characterized as *graphical* and *language-based approaches*.

Lee and Nelson (2004) report various graphical forms of external representations for instructional uses and provide a conceptual framework for external representations of knowledge. Graphical forms of externalization include (1) knowledge maps, (2) diagrams, (3) pictures, (4) graphs, (5) charts, (6) matrices, (7) flowcharts, (8) organizers, and (9) trees. However, not all of these forms of externalization have been utilized for instruction and educational diagnosis (Ifenthaler, 2008a; Scaife & Rogers, 1996; Seel, 1999a). Other forms of graphical approaches are the structure formation technique (Scheele & Groeben, 1984), pathfinder networks (Schvaneveldt, 1990), mind tools (Jonassen, 2009; Jonassen & Cho, 2008), and causal diagrams (Al-Diban & Ifenthaler, in press).

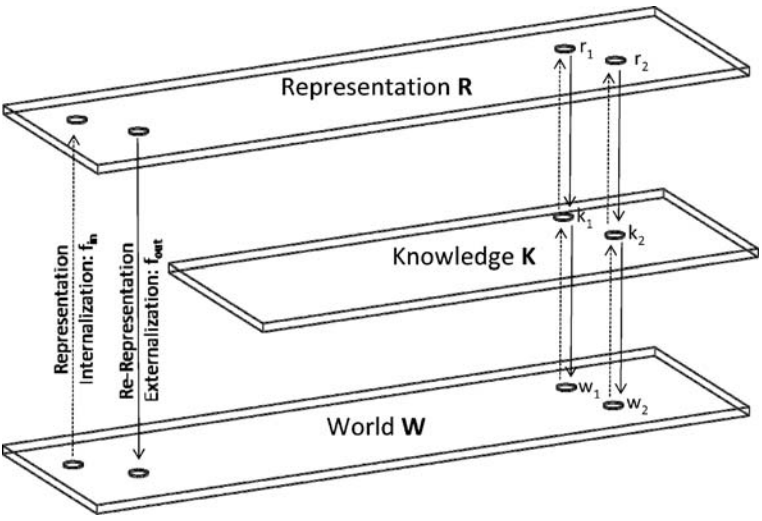


Fig. 12.1 Functions of representation and re-representation

Language-based approaches include thinking-aloud protocols (Ericsson & Simon, 1993), teach-back procedures (Mandl, Gruber, & Renkl, 1995), cognitive task analysis (Kirwan & Ainsworth, 1992), and computer linguistic techniques (Pirnay-Dummer, Ifenthaler, & Spector, 2009; Seel, Ifenthaler, & Pirnay-Dummer, 2009).

As discussed above, there are numerous approaches for eliciting knowledge for various diagnostic purposes. However, most approaches have not been tested for reliability and validity (Ifenthaler, 2008a; Seel, 1999a). Additionally, they are almost only applicable to single or small sets of data (Al-Diban & Ifenthaler, in press; Ifenthaler, 2008b). Hence, new approaches are required which have not only been tested for reliability and validity but also provide a fast and economic way of analyzing larger sets of data. Additionally, approaches for educational diagnostics also need to move beyond the perspective of correct and incorrect solutions. As we move into the twenty-first century, we argue that the application of alternative assessment and analysis strategies is inevitable for current educational diagnostics.

In this chapter, we focus on the scope of graphical indices in educational diagnostics. First, this chapter will provide an introduction to the implementation of graphs as external knowledge representations and present graphical indices and their possible applications in educational diagnostics. We will then highlight recent empirical studies which used graphical indices for educational diagnostics. The chapter will conclude with suggestions for future research for educational diagnostics using graphical indices.

12.2 Graphs as External Knowledge Representation

The underlying assumption is that knowledge can be re-represented (externalized) as a graph (Norman & Rumelhart, 1978). A graph consists of a set of vertices whose relationships are represented by a set of edges. The elements of a graph and their corresponding graphical measures are defined by the methods of graph theory (Diestel, 2000; Harary, 1974; Tittmann, 2003). Graph theory has been applied in various fields of research and applications, e.g., decision making, transactional analysis, network problems, transportation and traffic planning, scheduling problems, topology problems, and project management (see Chartrand, 1977). An overview of applications of graph theory in the social and psychological sciences has been provided by Durso and Coggins (1990) and in educational science by Nenninger (1980).

A widely accepted application of graph theory in social, educational, and psychological science is the use of Pathfinder networks (Schvaneveldt, 1990). Pathfinder provides a representation of knowledge by using pairwise similarity ratings among concepts to create a network. Pathfinder techniques have been combined with other procedures (e.g., multidimensional scaling – MDS) to expand the information for diagnostic purposes (e.g., Acton, Johnson, & Goldsmith, 1994; Goldsmith, Johnson, & Acton, 1991). However, Goldsmith et al. (1991) mention the need for more research regarding the psychological interpretation of graphs as knowledge

representation. Accordingly, we argue that graph theory has potential beyond its application in the Pathfinder approach. The following sections will strengthen our assumptions.

12.2.1 Basics of Graph Theory

A graph is constructed from a set of *vertices* whose relationships are represented by *edges*. Basics of graph theory are necessary to describe externalized knowledge representations as graphs (Bonato, 1990; Ifenthaler, Masduki, et al., 2009).

1. A graph *G(V,E)* is composed of vertices *V* and edges *E*. If the relationship between vertices *V* is directional, a graph is called a directed graph or digraph *D*. A graph which contains no directions is called an undirected graph.
2. The position of vertices *V* and edges *E* on a graph *G* are examined with regard to their proximity to one another. Two vertices *x, y* of *G* are adjacent if they are joined by an edge *e*. Two edges $e \neq f$ are adjacent if they have a common end or vertex *x*.
3. A path *P* is a graph *G* where the vertices x_i are all distinct. The length of a path *P* is calculated by the number of its edges e_j. The vertices x_0 and x_k are called the ends of the path *P*.
4. A graph *G* is indexed when single vertices *V* and edges *E* are distinguished by their names or content.
5. Every connected graph *G* contains a spanning tree. A spanning tree is acyclic and includes all vertices of *G*. Spanning trees can be used for numerous descriptions and calculations concerning the structure of a graph.

Please refer to Chapter 10, this volume, or to Tittmann (2003) for a detailed mathematical introduction to graphs and networks. The following part of this section will provide an overview of measures of graph theory which can be applied for educational diagnostics. However, as available measures of graph theory only account for *structural properties* of knowledge representations, the second to last part of this section will focus on measures beyond graph theory, namely *semantic properties*. The concluding part of this section will briefly describe the HIMATT tool, which integrates graphical indices for educational diagnostics (Pirnay-Dummer et al., 2009).

12.2.2 Measures of Graph Theory

By describing externalized knowledge representations as graphs, including associated vertices and edges, we are able to apply various measures from graph theory to diagnose individual knowledge representations and, in addition, to track the development of knowledge representations over time (Bonato, 1990; Ifenthaler, Masduki,

et al., 2009; White, 1985). Below we briefly describe appropriate structural measures, including information on the (a) operationalization, (b) computation rules, and (c) diagnostic purpose of a knowledge representation. None of the structural measures account for the content of the underlying knowledge representation.

1. *Number of vertices* indicates the number of concepts (vertices) within a graph.

    a. The size of the knowledge representation is indicated by the sum of all embedded concepts (semantically correct or incorrect).
    b. Computed as the sum of all vertices within a cognitive structure. Defined as a value between 0 (no vertices) and *N*.
    c. The diagnostic purpose is to identify additions of vertices (growth of the graph) as compared to previous knowledge representations and track change over time.

2. *Number of edges* indicates the number of links (edges) within a graph.

    a. The size of the knowledge representation is indicated by the sum of all embedded links (semantically correct or incorrect).
    b. Computed as the sum of all edges within a cognitive structure. Defined as a value between 0 (no edges) and *N*.
    c. The diagnostic purpose is to identify additions of links (closeness of associations of the graph) as compared to previous knowledge representations and track change over time.

3. *Connectedness* indicates how closely the concepts and links of the graph are related to each other.

    a. The closeness of the knowledge representation is indicated by all possible paths and their accessibility.
    b. Computed as the possibility to reach every vertex from every other vertex in the knowledge representation. Defined as a value between 0 (not connected) and 1 (connected).
    c. The diagnostic purpose is to identify the strength of closeness of associations of the knowledge representation. A strongly connected knowledge representation could indicate a deeper subjective understanding of the underlying subject matter.

4. *Ruggedness* indicates whether non-linked vertices of a graph exist.

    a. The concepts of a knowledge representation are not accessible from every other concept. Hence, the knowledge representation consists of at least two subgraphs which are not linked.
    b. Computed as the sum of subgraphs which are independent or not linked. Defined as a value between 1 (all vertices are linked) and *N*.
    c. The diagnostic purpose is to identify possible non-linked concepts, subgraphs, or missing links within the knowledge representation. Non-linked

concepts of a knowledge representation point to a lesser subjective understanding of the phenomenon in question.

5. *Diameter* indicates how large a graph is.
   a. The diameter of a knowledge representation is a reliable indicator for its complexity.
   b. Computed as the quantity of edges of the shortest path between the most distant vertices (diameter) of the spanning tree of a knowledge representation. Defined as a value between 0 (no edges) and *N*.
   c. The diagnostic purpose is to identify how broad the subject's understanding of the underlying subject matter is.
6. *Cyclic* indicates the existence of paths within a graph returning back to the start vertex of the starting edge.
   a. A cyclic knowledge representation contains a path returning back to the start concept of the starting link.
   b. Computed as the existence or nonexistence of cycles within the knowledge representation. Defined as 0 (no cycles) and 1 (is cyclic).
   c. The diagnostic purpose is to identify the strength of closeness of associations of the knowledge representation. A cyclic knowledge representation could indicate a deeper subjective understanding of the underlying subject matter.
7. *Number of Cycles* indicates the number of cycles within a graph.
   a. A cyclic knowledge representation contains at least one path returning back to the start concept of the starting link.
   b. Computed as the sum of all cycles within a knowledge representation. Defined as a value between 0 (no cycles) and *N*.
   c. The diagnostic purpose is to identify the strength of closeness of associations of the knowledge representation. Many cycles within a knowledge representation could indicate a deeper subjective understanding of the underlying subject matter.
8. *Average degree of vertices* indicates the average degree of all incoming and outgoing edges of all vertices within a graph.
   a. An increase in the number of incoming and outgoing links adds to the complexity of the knowledge representation.
   b. Computed as the average degree of all incoming and outgoing edges of the knowledge representation. Defined as a value between 0 and *N*.
   c. The diagnostic purpose is to identify a low, medium, or high density within the knowledge representation. Knowledge representations which only connect pairs of concepts can be considered weak; a medium density is expected for most good working knowledge representations.

Hietaniemi (2008) offers a powerful open-source module called Graph-0.84 which includes the graph data structures and algorithms described above.

The module can be implemented into PERL environments (The Perl Foundation, 2008). Features of the module have been implemented into the SMD Technology (Ifenthaler, 2008b), T-MITOCAR (Pirnay-Dummer, Ifenthaler, & Johnson, 2008), and HIMATT (Pirnay-Dummer et al., 2009).

12.2.3 Measures Beyond Graph Theory

Besides the measures of graph theory, which account for *structural properties* of knowledge representations, we argue that an educational diagnostics system should also account for the specific content (*semantic properties*). Therefore, we introduced semantic measures which add to the richness of detail of our proposed educational diagnostics (Johnson, Ifenthaler, Pirnay-Dummer, & Spector, 2009; Pirnay-Dummer et al., 2009). A semantic measure consists of a comparison feature which calculates similarities and contrasts between two or more different knowledge representations. Such measures for comparison can be applied to any knowledge representation which is available as a graph. Some of the measures count specific features of a given graph. For a given pair of frequencies f_1 *and* f_2, the similarity is generally derived by

$$s = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)}$$

which results in a measure of $0 \leq s \leq 1$, where $s = 0$ is complete exclusion and $s = 1$ is identity.

The other measures collect sets of properties from the graph (e.g., the vertices = concepts or the edges = relations). In this case, the Tversky similarity (Tversky, 1977) applies for the given sets A and B:

$$s = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A - B) + \beta \cdot f(A - B)}$$

α and β are weights for the difference quantities which separate A and B. They are usually equal ($\alpha = \beta = 0.5$) when the sources of data are equal. However, they can be used to balance different sources systematically (e.g., comparing a learner model which was constructed within 5 min to an expert model, which may be an illustration of the result of a whole book).

So far, three semantic measures have been developed, implemented, and tested for reliability and validity: (1) concept matching, (2) propositional matching, and (3) balanced propositional matching. Below we briefly describe these three semantic measures, including information on their (a) operationalization, (b) computation rules, and (c) diagnostic purpose.

(1) *Concept matching* compares the sets of concepts (vertices) within a graph to determine the use of terms.

   a. The use of semantically correct concepts (vertices) is a general indicator of an accurate understanding of the given subject domain.
   b. Computed as the sum of vertices of a knowledge representation which are semantically similar to a domain-specific reference representation (e.g., expert solution). Defined as a value between 0 (no semantic similar vertices) and *N*.
   c. The diagnostic purpose is to identify the correct use of specific concepts (e.g., technical terms). The absence of a great number of concepts indicates a less elaborated domain-specific knowledge representation.

(2) *Propositional matching* compares only fully identical propositions between two graphs.

   a. The use of semantically correct propositions (vertex-edge-vertex) indicates a correct and deeper understanding of the given subject domain.
   b. Calculated as the semantic similarity between a cognitive structure and a domain-specific reference cognitive structure. Defined as a value between 0 (no similarity) and 1 (complete similarity).
   c. The diagnostic purpose is to identify the right use of specific propositions (concept-link-concept), i.e., concepts correctly related to each other. Additionally, misconceptions can be identified for a specific subject domain by comparing known misconceptions (as propositions) to individual knowledge representations.

(3) *Balanced propositional matching* should be used instead of the concept and propositional matching to balance the dependency of both measures.

   a. Propositional matching necessarily has its maximum in the value of concept matching. In order to balance this dependency of both indices, the balanced propositional matching index should be used instead of the concept and propositional matching.
   b. Computed as the quotient of propositional matching and concept matching. Defined as a value between 0 (no similarity) and 1 (complete similarity).
   c. The diagnostic purpose is to account for the correct use of single concepts (e.g., technical terms) and their correct connectedness.

12.2.4 Implementation of Graphical Indices for Educational Diagnostics

The demand for an automated and computer-based diagnostic system incorporating a domain independent, fast, reliable, and valid assessment and analysis brought forth the HIMATT system (Highly Integrated Model Assessment Technology and

Tools; see Pirnay-Dummer et al., 2009). Methodologically, the tools integrated into HIMATT touch the boundaries of qualitative and quantitative research methods and provide bridges between them. First of all, text can be analyzed very quickly without loosening the associative strength of natural language. Furthermore, concept maps can be analyzed and compared to those of an expert or other participant.

Figure 12.2 shows the architecture of HIMATT. Within the system, experiments can be laid out and conducted for various educational diagnostic purposes. Additionally, external data in written or graphical formats can be integrated into HIMATT. The data can then be analyzed by the researcher. As a result of the analysis process, HIMATT generates standardized graphical representations and seven quantitative indicators which are based on graph theory.

Reliability measures exist for the individual instruments integrated into HIMATT. They range from $r = 0.79$ to $r = 0.94$ (Ifenthaler, 2008b; Pirnay-Dummer et al., 2009) and are tested for the semantic and structural measures separately and across different knowledge domains. Validity measures are also reported separately for the structural and semantic measures. Convergent validity lies between $r = 0.71$ and $r = 0.91$ for semantic comparison measures and between $r = 0.48$ and 0.79 for structural comparison measures (Pirnay-Dummer et al., 2009).

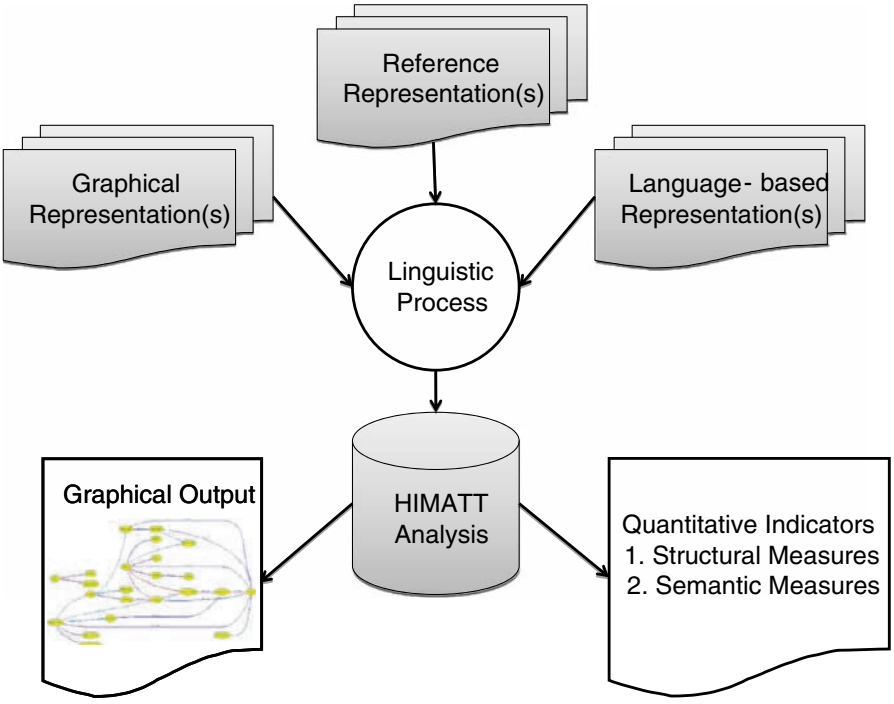


Fig. 12.2 HIMATT architecture

12.3 Empirical Studies

Empirical studies on the application of graph theory are available for almost every field of science. A literature review revealed over 14,000 scientific journal publications. The huge spectrum of research studies includes projects from *management* (e.g., Darvish, Yasaei, & Saeedi, 2009), *geophysics* (e.g., Todd, Toth, & Busa-Fekete, 2009), *medicine* (Chowdhury, Bhandarkar, Robinson, & Yu, 2009), *engineering* (e.g., Huang, Lo, Zhi, & Yuen, 2008; Rao & Padmanabhan, 2007), *neuroscience* (e.g., Bai, Qin, Tian, Dai, & Yang, 2009), *physics* (e.g., Ding & Guan, 2008), *computer science* (e.g., Bronevich & Meyer, 2008; Fiedler, 2007), *biology* (e.g., Ohtsuki, Pacheco, & Nowak, 2007), *chemistry* (e.g., Balaban, 1985), *oceanography* (e.g., Prigent, Fontenelle, Rochet, & Trenkel, 2008), and *anthropology* (e.g., Foster, 1978).

However, the number of empirical studies on the application of graph theory in the field of education is small (e.g., Durso & Coggins, 1990; Goldsmith et al., 1991; Hsia, Shie, & Chen, 2008; Nenninger, 1980; Schvaneveldt, 1990; Xenos & Papadopoulos, 2007). A series of empirical studies focusing on the application of graph theory in educational diagnostics using computer-based assessment and analysis techniques has been conducted recently. The graph theory-based analysis functions have been implemented into the HIMATT system (see above and Chapter 6 in this volume). In the following, we present three of these recent studies which provide insight into the possibilities of applying graph theory in educational diagnostics: (1) development of cognitive structures over time, (2) feedback for improving expert performance, and (3) between-domain distinguishing features of cognitive structures.

12.3.1 Development of Cognitive Structures

The study by Ifenthaler, Masduki et al. (2009) focuses on the issues involved in tracking the progression of cognitive structures, which captures the transition of learners from the initial state to the desired state (Snow, 1989, 1990), and making repetitive measurements of change over an extended period of time for a more accurate diagnosis (Ifenthaler & Seel, 2005; Seel, 1999a). Accordingly, it responds to the claim that research on cognitive structures needs to move beyond the traditional two-wave design in order to capture changes more precisely (Willett, 1988). As individuals reinstate and modify their cognitive structures when interacting with the environment (Jonassen et al., 1993, Piaget, 1976; Seel, 1991), the necessity of conducting multiwave longitudinal experiments is evident. However, the collection and analysis of longitudinal data gives rise to various methodological dilemmas which should not be neglected (see Ifenthaler, 2008a; Seel, 1999a). Besides general concerns about quantitative studies over time (Collins & Sayer, 2001; Moskowitz & Hershberger, 2002), tracking changes in cognitive structures requires valid and reliable assessment techniques, adequate statistical procedures, and specific situations which enable the activation of such cognitive structures (Ifenthaler, 2008a).

Indicators of graph theory have been assumed to be applicable for tracking the development of externalized cognitive structures over time.

Twenty-five students (18 female and 7 male) from the University of Freiburg, Germany, participated in the study. Their average age was 24.7 years ($SD = 1.9$). All students attended an introductory course on *research methods* in winter semester 2007. A total of 125 concept maps were collected at five measurement points during the semester.

Data were collected through concept maps using the software *CmapTools* (Cañas et al., 2004). According to Novak (1998), a concept map is a two-dimensional graphical representation of communicated knowledge and its underlying structure. A concept map consists of *concepts* (graph theory: vertices) and *relations* (graph theory: edges). Research studies on the application of *CmapTools* indicate that our theoretical assumptions on using this software are widely accepted (e.g., Coffey et al., 2003; Derbentseva, Safayeni, & Cañas, 2004). Since the research study focused on the development of cognitive structures, the longitudinal procedure included five measurement points. The main parts of the study were as follows:

(1) In a 60-min introductory lesson, the subjects were introduced to the concept mapping technique and taught how to use the CmapTools software. Additionally, the instructor collected demographic data and delivered documentation on concept maps and the software, including examples.
(2) At five measurement points (MP) during the course on research methods, the subjects were asked to create an open concept map relating to their understanding of research skills. Every subject needed to upload the concept map at a specified date and time during the course.
(3) The course learning outcome was measured by way of (1) five written assignments, (2) a written exam, and (3) a written research proposal. The course learning outcome was rated with a score between 0 and 100 points (Spearman-Brown coefficient, $r = 0.902$).

After uploading the concept maps, the instructor gave the students a brief feedback to notify them that their maps had been successfully uploaded and that they should carry on with their studies in the course. As open concept maps were used in the research study, the subjects were not limited to specific words while annotating the concepts and relations.

An in-depth analysis of $N = 125$ cognitive structures (five re-representations of each of the 25 participants) revealed several patterns that helped us to better understand the construction and development of these constructs over time. Several HLM analyses were computed to test the hypothesis. According to the guidelines suggested by Hox (2002), the sample size of the study is just adequate. However, in order to validate the initial findings further studies with a larger sample size will be necessary.

The results of the *Level-1* HLM analysis (intraindividual change of cognitive structures over time) are described in Tables 12.1 and 12.2. The *Mean Initial Status* π_{0i} indicates that all corresponding measures are significantly higher than 0. Except

Table 12. 1 Level-1 linear growth models of cognitive structures (structural measures)

		Coefficient	SE	*t*	*df*	*p*
Surface structure	Mean initial status π_{0i}	14.95	1.95	7.64	24	<0.001
	Mean growth rate π_{1i}	15.36	2.72	5.65	123	<0.001
Matching structure	Mean initial status π_{0i}	6.02	0.49	12.09	24	<0.001
	Mean growth rate π_{1i}	1.66	0.29	5.62	123	<0.001
Ruggedness	Mean initial status π_{0i}	1.27	0.11	11.48	24	<0.001
	Mean growth rate π_{1i}	0.35	0.11	3.32	123	0.002
Average degree of vertices	Mean initial status π_{0i}	2.01	0.08	24.19	24	<0.001
	Mean growth rate π_{1i}	0.03	0.03	1.32	123	0.189
Number of cycles	Mean initial status π_{0i}	2.85	0.44	6.49	24	<0.001
	Mean growth rate π_{1i}	0.52	0.19	2.69	123	0.008
Number of vertices	Mean initial status π_{0i}	13.68	1.79	7.65	24	<0.001
	Mean growth rate π_{1i}	14.59	2.63	5.56	123	<0.001
Number of edges	Mean initial status π_{0i}	4.14	0.74	5.62	24	<0.001
	Mean growth rate π_{1i}	4.40	0.93	4.71	123	<0.001

Table 12. 2 Level-1 linear growth models of cognitive structures (semantic measures)

		Coefficient	SE	*t*	*df*	*p*
Vertex matching	Mean initial status π_{0i}	8.49	0.85	9.94	24	<0.001
	Mean growth rate π_{1i}	3.67	0.41	8.99	123	<0.001
Propositional structure	Mean initial Status π_{0i}	0.0317	0.0056	5.63	24	<0.001
	Mean Growth rate π_{1i}	–0.0019	0.0016	–1.15	123	0.253

for *average degree of vertices*, all measures reveal a significant positive linear *Mean Growth Rate* π_{1i} per measurement point (e.g., *surface structure* = 15.36).

Therefore, $H_{1.1}$ can be accepted: The *structure* (surface structure, graphical structure, ruggedness, number of cycles, and number of vertices) of the externalized cognitive structures changes during the learning process, except for the measure *average degree of vertices*. The *average degree of vertices* indicates the average number of incoming and outgoing edges. Accordingly, as most of the externalized cognitive structures are very broad and do not center on one vertex, each vertex takes two edges on average. This does not change during the learning process, as the subject domain (research skills) does not change and does not seem to be organized around one central vertex.

The result of the HLM analysis revealed a significant growth in the *structural measures* between measurement points one and five. The overall size of the cognitive structures (*surface structure*) increased many times over. This is an indicator for an accommodation process (see Piaget, 1976; Seel, 1991), i.e., the individuals continuously added new concepts and links between concepts to their cognitive structures while learning. As a consequence, the complexity of the externalized cognitive structures also increased, which is indicated by the growth in the measures

matching structure and *number of cycles*. Therefore, we conclude that in the process of learning and understanding more and more about a given subject matter, individuals succeed in integrating single concepts and links more tightly. However, we also found significant growth in the measure *ruggedness* (i.e., non-linked concepts within the entire cognitive structure). The significant decrease in the measure *connectedness* supports this result. This indicates that newly learned concepts are not immediately integrated into the cognitive structure. The delay involved in integrating concepts into the cognitive structure should be kept in mind when constructing instructional materials and learning environments. We also suggest analyzing this phenomenon more precisely in a future study.

Contrary to the results of the *structural measures*, the HLM analysis revealed significant growth only in the semantic measure *vertex matching*. The individuals use more and more semantically correct concepts (vertices) in the course of the learning process. As individuals become more familiar with the terminology of the subject domain, they use these concepts more frequently. This learning process enables individuals to communicate their cognitive structures more precisely and in a more expert-like manner. The significant positive correlation we found between the final learning outcomes and the number of semantically correct concepts (*vertex matching*) reaffirms these assumptions (see Ifenthaler, Masduki et al., 2009).

Hence, in order to provide effective instruction, it is important for students' prior knowledge to be identified since the subsequent construction and organization of knowledge structures as well as mental models in a particular situation depends on the students' preconceptions and naïve theories (Seel, 1999a). Measures derived from graph theory proved to be reliable and valid indicators.

12.3.2 Feedback for Improving Expert Performance

In this chapter, Ifenthaler (in press-b) investigates different types of model-based feedback generated automatically with the HIMATT (Highly Integrated Model Assessment Technology and Tools) methodology (see Pirnay-Dummer et al., 2009). Since the beginnings of mental model research (e.g., Gentner & Stevens, 1983; Johnson-Laird, 1983; Seel, 1991) many research studies have provided evidence that "mental models guide and regulate all human perceptions of the physical and social world" (Seel & Dinter, 1995, p. 5). Accordingly, mental models are dynamic ad hoc constructions which provide subjectively plausible explanations on the basis of restricted domain-specific information (Ifenthaler, 2008b). Various research studies have shown that it is very difficult but possible to influence such subjectively plausible mental models by providing specific information (see Anzai & Yokoyama, 1984; Ifenthaler & Seel, 2005; Mayer, 1989; Seel, 1995; Seel & Dinter, 1995). Ifenthaler and Seel (2005) argue that it is important to consider how such information is provided to the learner at specific times during the learning process and how it is structured. In accordance with the general definition of feedback (see Wagner & Wagner, 1985), such information for improving individual mental model building processes provided purposely and on the fly is referred to as model-based feedback.

Hence, model-based feedback aims at a restructuring of the underlying representations and a reconceptualization of the related concepts and links (vertices and edges). This is in following with Piaget's epistemology (1950, 1976). New information provided via model-based feedback can be assimilated through the activation of an existing schema, adjustment by accretion, or tuning of an existing schema. Otherwise it is accommodated by means of a reorganization process which involves building new mental models (Seel et al., 2009).

Seventy-four students (66 female and 8 male) from the University of Freiburg, Germany, participated in the study. Their average age was 21.9 years ($SD = 2.3$). The participants were randomly assigned to the three experimental groups: (1) cutaway feedback ($n = 26$), (2) discrepancy feedback ($n = 24$), and (3) expert feedback ($n = 24$).

First, the participants completed a demographic data questionnaire. Second, they completed the concept map and causal diagram experience questionnaire. Next, the participants completed the test on verbal (6 min) and spatial abilities (9 min). Then they answered the 27 multiple choice questions of the domain-specific knowledge test on climate change (pretest). After a short relaxation phase, the participants were given an introduction to concept maps and causal diagrams and were shown how to use the HIMATT software. Then, the participants used the username and password they had been assigned to log in to the HIMATT system, where they constructed a causal diagram on their understanding of climate change (10 min). Immediately afterward, they wrote a text about their understanding of climate change (10 min). A short relaxation phase followed, during which we automatically generated the individual feedback models for each participant. After that, the participants received the text on climate change and the automatically generated feedback model (cutaway, discrepancy, or expert model – depending on which experimental group they had been assigned to). All three types of feedback models were automatically generated with HIMATT. The *cutaway feedback model* included all propositions (vertex-edge-vertex) of the participants' pretest causal diagram. Additionally, the semantically correct vertices (compared to the expert re-representation) were graphically highlighted (circles are semantically correct as compared to the expert; ellipsis are semantically incorrect as compared to the expert re-representation). The *discrepancy feedback model* included only propositions (vertex-edge-vertex) of the participants' pretest causal diagram which had no semantic similarity to the expert re-representation. The *expert feedback model* consisted of a standardized re-representation of an expert's model on climate change. The participants had 15 min to read the text and examine their feedback model. Immediately after working on the text, the participants completed the model feedback quality test. Then they answered the 27 multiple choice questions of the posttest on declarative knowledge. After another short relaxation phase, the participants used their username and password to log in to the HIMATT system for the second time. In the HIMATT posttest, they constructed a second causal diagram on their understanding of climate change (10 min) and wrote a second text regarding their understanding of climate change (10 min). Finally, the participants had to complete a short usability test regarding HIMATT features.

Table 12.3 Average gain of HIMATT measures for the experimental groups ($N = 74$)

	Cutaway feedback ($n = 26$)	*SD*	Discrepancy feedback ($n = 24$)	*SD*	Expert feedback ($n = 24$)	*SD*
Surface matching	1.731	3.779	3.375	2.871	4.826	4.579
Graphical matching	–0.192	1.497	0.875	1.985	1.609	1.438
Structural matching	1.231	3.766	2.583	1.213	3.087	2.353
Gamma matching	0.005	0.099	–0.001	0.142	–0.019	0.155
Concept matching	0.052	0.074	0.020	0.067	–0.010	0.109
Propositional matching	0.007	0.027	0.006	0.026	–0.001	0.002
Balanced propositional matching	–0.008	0.091	0.000	0.044	–0.009	0.079

The graphical re-representations of the participants were analyzed automatically with the HIMATT analysis feature. Hence, the knowledge gain of the seven HIMATT measures was computed by subtracting the pre- from the post-measure. Table 12.3 shows the average gain of the HIMATT measures (surface, graphical, structural, gamma, concept, propositional, and balanced propositional matching) for the three experimental groups (cutaway feedback, discrepancy feedback, and expert feedback).

The results showed a significant effect between participants in the three experimental groups for the HIMATT measure *surface matching*, $F(2, 70) = 4.080$, $p = 0.021$, $\eta^2 = 0.10$, with participants of the *expert feedback group* increasing their number of vertices more than the other experimental groups. The one-way ANOVA also revealed a significant effect for the HIMATT measure *graphical matching*, $F(2, 70) = 7.355$, $p = 0.001$, $\eta^2 = 0.17$. The increase of complexity of participants was higher in the *expert feedback group* than in the others. Between the experimental groups, the increase of the HIMATT measure *structural matching* was also significant, $F(2, 70) = 3.140$, $p = 0.049$, $\eta^2 = 0.08$. Again, the participants in the *expert feedback group* outperformed the other experimental groups. For the semantic HIMATT measure *concept matching* a final significant effect was found, $F(2, 70) = 3.243$, $p = 0.045$, $\eta^2 = 0.08$. Here, participants in the *cutaway feedback group* gained more correct concepts than the participants in the other two groups. However, no further effects for the HIMATT measures were found.

With the help of the seven automatically calculated HIMATT measures, changes in the participants' understanding of the subject domain "climate change" were investigated and re-represented with causal diagrams. Participants who received the expert feedback added significantly more relations to their causal diagrams (*surface matching*) than did those in the other groups. Accordingly, the expert feedback provided them a broad spectrum of concepts and relations, which were then integrated into their own understanding of the phenomenon in question. This also explains the significant differences between the measures *graphical* and *structural matching*. As the number of relations in a causal diagram increases, there is a high probability that its complexity and complete structure will also increase.

However, an increase in these structural measures does not necessarily mean that the solutions of participants in the expert feedback group are better than these of the other participants. As a further analysis of the semantic HIMATT measures revealed, participants in the cutaway feedback group outperformed the other participants with regard to their semantic understanding of the phenomenon in question (*concept matching*). Accordingly, even if the structure increases, the semantic correctness of the learner will not automatically also increase. Hence, learners may integrate a huge amount of concepts into their understanding of the phenomenon which do not necessarily help them to come to a better and more correct solution to the problem.

Thus, measures derived from graph theory also proved to be reliable and valid indicators in the study on model-based feedback. Further studies will focus on the learning trajectories while providing forms of model-based feedback. This will provide more detailed insight into the effects of model-based feedback and how it helps to support and improve expertise and expert performance.

12.3.3 Between-Domain Distinguishing Features of Cognitive Structures

In this study, Ifenthaler and Hetterich (under review) argue that previous empirical studies have focused on within-domain-specific features and the learning-dependent development of cognitive structures (e.g., Clariana & Wallace, 2007; Ifenthaler, Masduki et al., 2009; Koubek, Clarkston, & Calvez, 1994). In contrast to these empirical investigations, this study focuses on between-domain specific similarities and differences. More precisely, it identifies similarities and differences in externalized cognitive structures between three different subject domains: mathematics, biology, and history.

The central research objective was to identify between-domain distinguishing features of externalized cognitive structures. Accordingly, the participants were asked to externalize their understanding of three different subject domains (mathematics, biology, history). Additionally, it is argued that the form of externalization influences the person's communicated output (Ifenthaler, 2008a). Therefore, the participants were asked to externalize their understanding of each subject domain as written text and as a concept map.

Seventy-one students (66 female and 8 male) from the University of Freiburg, Germany, participated in the study. Their average age was 22.2 years ($SD = 2.3$). First, the participants completed a *demographic data questionnaire* and the *experience with concept mapping test*. Second, they completed the test on *verbal*, *mathematical*, and *spatial abilities*. Next, they were given an introduction to concept maps and causal diagrams and were shown how to use the HIMATT software (Pirnay-Dummer et al., 2009). After a short relaxation phase, the participants completed the *domain-specific knowledge test* on history. Then they received the *text on European borders*. The participants had 15 min to read the text. Then, they logged

into the HIMATT system, where they constructed a causal diagram on their understanding of European borders (10 min). Immediately afterward, they wrote a text about their understanding of European borders (10 min). After another short relaxation phase, the procedure was repeated with the domains *mathematics* and *biology* ((1) Domain-specific knowledge test, (2) reading the text, (3) constructing a concept map, and (4) writing a test). In total, the experiment took approximately 2 h.

Overall, we found highly significant differences in the four structural measures of HIMATT between the three subject domains – for both written text (Table 12.4) and concept maps (Table 12.5). The ANOVA revealed a significant effect for written text for the measures *graphical matching* ($F_{(2, 208)} = 3.064$, $p < 0.05$; $\eta^2 = 0.03$) and *gamma matching* ($F_{(2, 208)} = 8.929$, $p < 0.001$; $\eta^2 = 0.08$).

For the concept maps, the ANOVA revealed a different picture. A significant effect was found between the three subject domains for the measures *surface matching* ($F_{(2, 207)} = 25.271$, $p < 0.001$; $\eta^2 = 0.20$), *graphical matching* ($F_{(2, 207)} = 8.186$, $p < 0.001$; $\eta^2 = 0.07$), and *structural matching* ($F_{(2, 207)} = 36.540$, $p < 0.001$; $\eta^2 = 0.26$).

The findings indicate that there are similarities and differences between the structural features of the externalized cognitive structures. Additionally, initial analysis also indicates similarities and differences between the two forms of externalization (written text and concept map). This new research on the application of graph theory measures in educational diagnostics indicates another useful application of these quantitative indices.

Table 12.4 Means, standard deviations of the four structural measures of HIMATT for the written text

	Mathematics		Biology		History	
	M	*SD*	*M*	*SD*	*M*	*SD*
Surface matching	17.47	11.74	22.56	29.08	16.17	19.73
Graphical matching	4.50	2.28	4.23	3.19	3.39	2.77
Structural matching	10.68	4.95	11.62	9.62	9.28	7.81
Gamma matching	0.60	0.55	0.89	0.32	0.86	0.36

Table 12.5 Means, standard deviations of the four structural measures of HIMATT for the concept maps

	Mathematics		Biology		History	
	M	*SD*	*M*	*SD*	*M*	*SD*
Surface matching	10.27	3.52	13.61	4.21	9.29	3.52
Graphical matching	5.07	1.79	5.58	1.79	4.35	1.79
Structural matching	9.97	3.05	13.73	3.79	9.32	2.98
Gamma matching	0.43	0.11	0.47	0.08	0.46	0.13

12.4 Conclusion

There is an immense field of applications for graphical indices in educational diagnostics. Graph theory has proven to be an appropriate diagnostic approach, especially in knowledge representation and analysis. Pathfinder and combined techniques (Durso & Coggins, 1990; Schvaneveldt, 1990) provide a reliable representation of knowledge structures and analysis of learning as they use pairwise similarity ratings among concepts to create networks. These networks are based on proximity data among entities and are determined by calculating the proximities that best fit within the network. Furthermore, newly developed automated applications such as SMD Technology (Ifenthaler, 2008b), T-MITOCAR (Johnson et al., 2009; Pirnay-Dummer et al., 2008), and HIMATT (Pirnay-Dummer et al., 2009) integrate the latest software technology and a great quantity of graph theory-based applications and analysis functions.

Additionally, graph theory can be applied to almost every area of educational diagnostics. Picard (1980) introduced a very promising approach for designing and analyzing questionnaires using graph theory. Furthermore, graph theory has been successfully applied for instructional planning (Hsia et al., 2008) and evaluation purposes (Xenos & Papadopoulos, 2007).

Future applications of graph theory in educational diagnostics include automated self-assessment and forms of automated feedback. A recently implemented application is TASA (Text-Guided Automated Self-Assessment). TASA is a web-based online tool for self-assessment of written essays. It embeds the parts of SMD Technology (Ifenthaler, 2008b) and T-MITOCAR (Johnson et al., 2009) which are necessary to generate a graph from the learners' essay directly after the upload. The uploaded essay provides the learner with a graphical representation of the essay in a format which non-experts have been shown to be able to handle. Additionally, graph theory-based measures make TASA into both a reflection and a preflection tool for the learner: After the upload is finished, the learners receive written feedback on the text. The text provides information on the key concepts, the ways in which they are connected, and concepts and connections which may be circumstantial but still have some added meaning in the text. TASA uses measures from graph theory to generate this feedback. If there is a group of learners which is working on the same task or topic, TASA may also be used as a preflection tool. Preflection will allow the learners to plan their actions based on what is already there and the task (goal) to fulfill. Once all members of the group have uploaded their text, TASA generates a list of the most common terms from all texts throughout the group. The learners are then asked which five terms from the whole list they would like to have in their underlying model (knowledge representation) when they upload their essay the next time. They select from a list of 30 terms. In this way, the individual learner can benefit from the other learners' conceptions.

In our digital age, technology, learning, and educational diagnostics are closely linked (Ifenthaler, in press-a; Ifenthaler, Isaias, Spector, Kinshuk, & Sampson, 2009). Researchers and engineers have always endeavored to design and develop useful diagnostic systems to serve professional communities in the field of learning

and instruction, and they will continue to do so (Ifenthaler, in press-a). Future work on automated computational diagnostics, including approaches such as graph theory, will provide more and more powerful dynamic systems for the comprehensive analysis of large amounts of data in a short space of time.

References

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology*, *86*(2), 303–311.

Al-Diban, S., & Ifenthaler, D. (in press). Comparison of two analysis approaches for measuring externalized mental models: Implications for diagnostics and applications. *Journal of Educational Technology & Society*.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anzai, Y., & Yokoyama, T. (1984). Internal models in physics problem solving. *Cognition and Instruction*, *1*(4), 397–450.

Bai, L., Qin, W., Tian, J., Dai, J., & Yang, W. (2009). Detection of dynamic brain networks modulated by acupuncture using a graph theory model. *Progress in Natural Science*, *19*(7), 827–835.

Balaban, A. T. (1985). Graph theory and theoretical chemistry. *Journal of Molecular Structure: THEOCHEM*, *120*, 117–142.

Bonato, M. (1990). *Wissenstrukturierung mittels Struktur-Lege-Techniken. Eine grapentheoretische Analyse von Wissensnetzen*. Frankfurt am Main: Lang.

Bronevich, A. G., & Meyer, W. (2008). Load balancing algorithms based on gradient methods and their analysis through algebraic graph theory. *Journal of Parallel and Distributed Computing*, *68*(2), 209–220.

Cañas, A. J., Hill, R., Carff, R., Suri, N., Lott, J., Eskridge, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak, & F. M. González (Eds.), *Concept maps: Theory, methodology, technology, Proceedings of the First International Conference on Concept Mapping* (pp. 125–133). Pamplona: Universidad Pública de Navarra.

Chartrand, G. (1977). *Introductory graph theory*. New York: Dover.

Chowdhury, A. S., Bhandarkar, S. M., Robinson, R. W., & Yu, J. C. (2009). Virtual craniofacial reconstruction using computer vision, graph theory and geometric constraints. *Pattern Recognition Letters*, *30*(10), 931–938.

Clariana, R. B., & Wallace, P. E. (2007). A computer-based approach for deriving and measuring individual and team knowledge structure from essay questions. *Journal of Educational Computing Research*, *37*(3), 211–227.

Coffey, J. W., Carnot, M. J., Feltovich, P. J., Feltovich, J., Hoffman, R. R., Cañas, A. J., et al. (2003). *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support*. Pensacola, FL: Chief of Naval Education and Training.

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change*. Washington, DC: American Psychological Association.

Darvish, M., Yasaei, M., & Saeedi, A. (2009). Application of the graph theory and matrix methods to contractor ranking. *International Journal of Project Management*, *27*(6), 610–619.

Derbentseva, N., Safayeni, F., & Cañas, A. J. (2004). Experiments on the effects of map structure and concept quantification during concept map construction. In A. J. Cañas, J. D. Novak, & F. M. González (Eds.), *Concept maps: Theory, methodology, technology, Proceedings of the First International Conference on Concept Mapping* (pp. 125–132). Pamplona: Universidad Pública de Navarra.

Diestel, R. (2000). *Graph theory*. New York: Springer.

Ding, L., & Guan, Z. H. (2008). Modeling wireless sensor networks using random graph theory. *Physica A: Statistical Mechanics and its Applications*, *387*(12), 3008–3016.

Durso, F. T., & Coggins, K. A. (1990). Graphs in social and psychological sciences: Empirical contributions to Pathfinder. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 31–51). Norwood, NJ: Ablex Publishing Corportion.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Fiedler, M. (2007). Reminiscences related to graph theory. *Computer Science Review*, *1*(1), 65–66.

Foster, B. L. (1978). Formal network studies and the anthropological perspective. *Social Networks*, *1*(3), 241–255.

Gentner, D., & Stevens, A. L. (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology*, *83*(1), 88–96.

Harary, F. (1974). *Graphentheorie*. München: Oldenbourg.

Hietaniemi, J. (2008). *Graph-0.84*. Retrieved May 6, 2008, from http://search.cpan.org/~jhi/Graph-0.84/lib/Graph.pod

Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hsia, T. C., Shie, A. J., & Chen, L. C. (2008). Course planning of extension education to meet market demand by using data mining techniques – an example of Chinkuo technology university in Taiwan. *Expert Systems with Applications*, *34*(1), 596–602.

Huang, H. C., Lo, S. M., Zhi, G. S., & Yuen, R. K. K. (2008). Graph theory-based approach for automatic recognition of CAD data. *Engineering Applications of Artificial Intelligence*, *21*(7), 1073–1079.

Ifenthaler, D. (2008a). Practical solutions for the diagnosis of progressing mental models. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 43–61). New York: Springer.

Ifenthaler, D. (2008b). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*. doi: 10.1007/s11423-008-9087-4

Ifenthaler, D. (in press-a). Learning and instruction in the digital age. Introduction. In J. M. Spector, D. Ifenthaler, P. Isaías, Kinshuk & D. G. Sampson (Eds.), *Learning and instruction in the digital age: Making a difference through cognitive approaches, technology-facilitated collaboration and assessment, and personalized communications*. New York: Springer.

Ifenthaler, D. (in press-b). Model-based feedback for improving expertise and expert performance. *Technology, Instruction, Cognition and Learning*.

Ifenthaler, D., & Hetterich, B. (under review). Identifying between-domain distinguishing features of cognitive structures: Instructional implications.

Ifenthaler, D., Isaias, P., Spector, J. M., Kinshuk, & Sampson, D. G. (2009). Editors' introduction to the special issue on cognition & learning technology. *Educational Technology Research and Development*. doi: 10.1007/s11423-009-9127-8

Ifenthaler, D., Masduki, I., & Seel, N. M. (2009). The mystery of cognitive structure and how we can detect it. Tracking the development of cognitive structures over time. *Instructional Science*. doi: 10.1007/s11251-009-9097-6

Ifenthaler, D., & Seel, N. M. (2005). The measurement of change: Learning-dependent progression of mental models. *Technology, Instruction, Cognition and Learning*, *2*(4), 317–336.

Johnson-Laird, P. N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness*. Cambridge, UK: Cambridge University Press.

Johnson, T. E., Ifenthaler, D., Pirnay-Dummer, P., & Spector, J. M. (2009). Using concept maps to assess individuals and team in collaborative learning environments. In P. L. Torres & R. C. V. Marriott (Eds.), *Handbook of research on collaborative learning using concept mapping* (pp. 358–381). Hershey, PA: Information Science Publishing.

Jonassen, D. H. (2009). Externally modeling mental models. In L. Moller, J. B. Huett, & D. Harvey (Eds.), *Learning and instructional technologies for the 21st century. Visions of the future* (pp. 49–74). New York: Springer.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hilsdale, NJ: Lawrence Erlbaum.

Jonassen, D. H., & Cho, Y. H. (2008). Externalizing mental models with mindtools. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 145–160). New York: Springer.

Kirwan, B., & Ainsworth, L. K. (1992). *A Guide to task analysis*. London: Taylor & Francis Group.

Koubek, R. J., Clarkston, T. P., & Calvez, V. (1994). The training of knowledge structures for manufacturing tasks: An empirical study. *Ergonomics*, *37*(4), 765–780.

Lee, Y., & Nelson, D. (2004). *Instructional use of visual representations of knowledge.* Paper presented at the Society for Information Technology and Teacher Education International Conference 2004, Atlanta, GA, USA.

Mandl, H., Gruber, H., & Renkl, A. (1995). Mental models of complex systems: When veridicality decreases functionality. In C. Zucchermaglio, S. Bagnara, & S. U. Stucky (Eds.), *Organizational learning and technological change* (pp. 102–111). Berlin: Springer.

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research*, *59*(1), 43–64.

Minsky, M. (1981). A framework for representing knowledge in mind design. In R. J. Brachmann & H. J. Levesque (Eds.), *Readings in knowledge representation* (pp. 245–262). Los Altos, CA: Morgan Kaufmann.

Moskowitz, D. S., & Hershberger, S. L. (Eds.). (2002). *Modelling intraindividual variability with repeated measures data*. Mahwah, NJ: Lawrence Erlbaum.

Nenninger, P. (1980). Anwendungsmöglichkeiten der Graphentheorie in der Erziehungswissenschaft. *Zeitschrift für Empirische Pädagogik*, *4*, 85–106.

Norman, D. A., & Rumelhart, D. E. (Eds.). (1978). *Strukturen des Wissens. Wege der Kognitionsforschung*. Stuttgart: Klett.

Novak, J. D. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ohtsuki, H., Pacheco, J. M., & Nowak, M. A. (2007). Evolutionary graph theory: Breaking the symmetry between interaction and replacement. *Journal of Theoretical Biology*, *246*(4), 681–694.

Piaget, J. (1950). *La construction du réel chez l'enfant*. Neuchatel: Delachaux et Niestlé S.A.

Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen*. Stuttgart: Klett.

Picard, C. F. (1980). *Graphs and questionnaires*. Amsterdam: North-Holland Publishing Company.

Pirnay-Dummer, P., Ifenthaler, D., & Johnson, T. E. (2008). *Reading with the guide of automated graphical representations. How model based text visualizations facilitate learning in reading comprehension tasks.* Paper presented at the AREA 2008, New York.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2009). Highly integrated model assessment technology and tools. *Educational Technology Research and Development*. doi: 10.1007/s11423-009-9119-8

Prigent, M., Fontenelle, G., Rochet, M. J., & Trenkel, V. M. (2008). Using cognitive maps to investigate fishers' ecosystem objectives and knowledge *Ocean and Coastal Management*, *51*(6), 450–462.

Rao, R. V., & Padmanabhan, K. K. (2007). Rapid prototyping process selection using graph theory and matrix approach. *Journal of Materials Processing Technology*, *194*(1), 81–88.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human – Computer Studies*, *45*(2), 185–213.

Scheele, B., & Groeben, N. (1984). *Die Heidelberger Struktur-Lege-Technik (SLT). Eine Dialog-Konsens-Methode zur Erhebung subjektiver Theorien mittlerer Reichweite*. Weinheim: Beltz.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood: NJ: Ablex Publishing Corporation.

Seel, N. M. (1991). *Weltwissen und mentale Modelle*. Göttingen: Hogrefe.

Seel, N. M. (1995). Mental models, knowledge transfer, and teaching strategies. *Journal of Structural Learning and Intelligent Systems*, *12*(3), 197–213.

Seel, N. M. (1999a). Educational diagnosis of mental models: Assessment problems and technology-based solutions. *Journal of Structural Learning and Intelligent Systems*, *14*(2), 153–185.

Seel, N. M. (1999b). Educational semiotics: School learning reconsidered. *Journal of Structural Learning and Intelligent Systems*, *14*(1), 11–28.

Seel, N. M., & Dinter, F. R. (1995). Instruction and mental model progression: Learner-dependent effects of teaching strategies on knowledge acquisition and analogical transfer. *Educational Research and Evaluation*, *1*(1), 4–35.

Seel, N. M., Ifenthaler, D., & Pirnay-Dummer, P. (2009). Mental models and problem solving: Technological solutions for measurement and assessment of the development of expertise. In P. Blumschein W. Hung, D. H. Jonassen & J. Strobel (Eds.), *Model-based approaches to learning: Using systems models and simulations to improve understanding and problem solving in complex domains* (pp. 17–40). Rotterdam: Sense Publishers.

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in Physics education. *Journal of Educational Psychology*, *63*(3), 225–234.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher*, *18*(9), 8–14.

Snow, R. E. (1990). New approaches to cognitive and conative assessment in education. *International Journal of Educational Research*, *14*(5), 455–473.

The Perl Foundation. (2008). *Perl 6*. Retrieved February 11, 2008, from http://www.perlfoundation.org/

Tittmann, P. (2003). *Graphentheorie. Eine anwendungsorientierte Einführung*. München: Carl Hanser Verlag.

Todd, C. S., Toth, T. M., & Busa-Fekete, R. (2009). GraphClus, a MATLAB program for cluster analysis using graph theory. *Computers & Geosciences*, *35*(6), 1205–1213.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Wagner, W., & Wagner, S. U. (1985). Presenting questions, processing responses, and providing feedback in CAI. *Journal of Instructional Development*, *8*(4), 2–8.

White, R. T. (1985). Interview protocols and dimensions of cognitive structure. In L. H. T. West & A. L. Pines (Eds.), *Cognitive structure and conceptual change*. Orlando, FL: Academic Press.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Xenos, M., & Papadopoulos, T. (2007). Computer aided evaluation of higher education tutors' performance. *Studies in Educational Evaluation*, *33*(2), 175–196.

Chapter 13
Complete Structure Comparison

Pablo Pirnay-Dummer

13.1 Knowledge and Structure

Knowledge is neither observable nor useful as long as it is not actively used, nor is its structure. Once it is processed as the basis of action, a certain well-known sequence is inevitable: Knowledge leads to decision. Decision leads to intention. Intention leads to action. Action is behavior, although not all behavior is action and is therefore not necessarily intentional. We can only measure behavior all the time, and, therefore, the assessment of knowledge itself is also always a heuristic to carefully derive parts of behavior which are most likely to have a connection with the construct of knowledge. In assessment, researchers work with externalized artifacts which are believed to correspond to the actual representation. The chain may be disrupted by certain psychological means (e.g., pathologically), but this discussion is not part of this chapter. Since this work will address the structure of knowledge representation as externalized by humans, an accessible knowledge modeling will be assumed. A structure is what binds conceptual pieces of representation together (Preece, 1976) in such a way that heuristic and strict reasoning of any kind accessible by humans may be performed on its basis (Gentner & Markman, 2006; Seel, 1991). Logic performs on a more or less coherent structure in order to derive one or multiple decisions. Kant illustrates this inspiring distinction between content and logic as follows: "The difference between a confused and a clear representation is merely logical and has nothing to do with content" (Kant, 1787). He elaborates even further on the logical structure within concepts as they may not have phenomenological manifestations:

> No doubt the conception of right, as employed by a sound understanding, contains all that the most subtle investigation could unfold from it, although, in the ordinary practical use of the word, we are not conscious of the manifold representations comprised in the conception. But we cannot for this reason assert that the ordinary conception is a sensuous one, containing a mere phenomenon, for right cannot appear as a phenomenon; but the conception of it

P. Pirnay-Dummer (✉)
Albert-Ludwigs-Unviversity, Freiburg, Germany
e-mail: pablo@pirnay-dummer.de

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_13,

> lies in the understanding, and represents a property [. . .] of actions, which belongs to them in themselves. (Kant, 1787)

Simple or complex, abstract or concrete, and also nested concepts – which may recursively refer to structure as well – are the basic elements, and the structure allows the decision-making process to navigate between them. They can have an inner structure, especially as they are used in higher order and abstract representations. A structure is a model which allows the human heuristics of reasoning and judgment (Gilovich & Griffin, 2002) to maneuver between the conceptual content for deduction (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) and induction (Seel, 1991). The inner structure constitutes at least a hierarchical relationship between different abstraction layers. The abstraction layer of a model is based on the abstraction level of the implicit or explicit task as presented to the system (Pirnay-Dummer, 2006), e.g., a task about money may cover everything between what is in someone's wallet to world economics and will still be structurally and conceptually involved in answering a question like "Do we have enough money for X?" Even in the wallet case, which has a very salient prototype, the question may be semantically complex. Suppose a visit to the movies costs 7 euros and I have 10 euros in my wallet. "No" might still be a valid inference, e.g., if I plan on going out tomorrow night or if I need 4 euros for the train back home. The list of plausible constraints which are implied neither by the question nor by the only key concept the question aims at (money) could easily be expanded almost indefinitely. The same holds true even more so in more complex use of the term (world economics), although the term is still not lexically ambiguous. Wittgenstein provided the same conclusion in a more general way:

> If I have two things, I can of course reconcile them with one another, at least hypothetically, but the characteristic thing about the scope of the concept is its class, and the concept which contains it was only a crutch, a pretense, an excuse. (Wittgenstein, 1994)

Thus, the model layer is composed of the term money itself, the concept it is supposed to represent, and the context of the reasoning task. This also explains the relation of concepts toward their model structure. In the same way the abstraction level may be different depending on the context, the scope may also vary. Later on, I will introduce a simple example of a model which contains a "firefighter" as a concept. Of course, if the task, which is the reason for the model structure to be generated, changes, then "firefighter" can also be the model topic. Then the model structure would, e.g., resemble everything which belongs to a firefighter or clarify the functions a firefighter has. It will play a significant role in the decision-making process, even in very small model structures with a clear scope within the reasoning domain (model layer). Therefore, the symbolic context within which concepts are composed into model structures carries meaning inasmuch as the model structures allow the reasoning process to be completed and lead to a decision (Montague, 1974; Link, 1979) – no matter how incomplete the knowledge may be in terms of expert knowledge.

This is the main reason why the structure of expertise is so interesting, especially to the field of learning and instruction, where expertise is considered to be a good

guide for curriculum, teaching, and learning (Seel, 2003). Learners will always have to measure themselves against expertise because this is the driving force behind any real learning goal (Snow, 1990). Also, the learners' progress will be measured as a convergence toward expert solutions (Ifenthaler, 2006). Expertise rarely has outside criteria but rather is determined in a normative way. Whether somebody is considered an expert is determined by perceived (interpreted) performance. If the person performs better than most other people in the same field, he or she must be an expert (Gruber, 1994). As regards expert knowledge, two aspects are interesting for researchers: the *content* and the *general features* of how content is processed by experts, especially in contrast to nonexperts. The latter raises all kinds of theoretical questions about whether expert knowledge is generally structured (e.g., embedded) differently than non-expert knowledge (e.g., Jonassen, Beissner, & Yacci, 1993). Moreover, different kinds of experts in the same domain appear to differ in their general knowledge – structurally and semantically – when they are confronted with the same reasoning tasks within their field, e.g., *industry salesmen* and *economists* (Pirnay-Dummer, 2006).

The synthesis of the limitations of single concepts and the idea of general features of structure provides a rationale for a development and implementation which allows the elicitation of structure only – without regard to content. Only with these measures at hand can structural similarities of expertise be explored. However, neither a structure mapping perspective (see Forbus, Gentner, Markman, & Ferguson, 1998) nor a graph isomorphism viewpoint (e.g., Kubose, Holyoak, & Hummel, 2002) provides a sufficient solution. The former requires conceptual content and can therefore not account for similar structures between different domains. Although the latter has shown to be highly applicable when analogies are being compared (see Gentner & Markman, 2006), it is limited to the comparison of acyclic graphs, e.g., in short sentence analogies. The structural mapping theory is able to account for priorization and hierarchy: Different types of relations can be given different ranks in order to structure them, which makes it easier to search for coherence (see Gentner & Bowdle, 2008). The method I discuss in this chapter can process cyclic graphs but is not capable of accounting for such a meta-structure within and – more importantly – between domains. It is also built for larger and less organized structures than analogies.

13.2 Retracing Knowledge Structure

Knowledge, especially from text, can be analyzed in many different ways (Helbig, 2006). The more automated ways are of particular interest within the context this book. However, deciding between available analysis methods in general is of course related to research pragmatics. Not all methods are suitable to help answer a specific research question. The decision also depends on the scope of assumptions one would like (or even have) to make (Pirnay-Dummer, 2008). However, two general poles

may be distinguished which are connected to two basic questions: *Who interprets the available artefacts* (data) and *how are they interpreted?*

For the scope of desirable assumptions this translates into the following:

So far, the common aggregate for knowledge re-representation (externalization) is a graph. A graph $G(E,V)$ is defined as a set of edges $E = \{e_1 \ldots e_n\}$ which links a set of vertices $V = \{v_1 \ldots v_n\}$ (usually pairwise). Vertices are in most cases concepts which are used in the context of knowledge but may occasionally also refer to more complex entities (e.g., clusters) (see Chapter 10 for a more detailed description of the graph features). A graph is not necessarily represented in a visual (graphical) way. For computing purposes, the XML standard and also classical list forms (e.g., Ifenthaler, 2008) are available.

There are different ways to construct the graph. Usually, the subjects either draw the graphs, with or without ongoing support, or they express themselves in spoken or written language which is then transferred into a graph afterward. There are manual ways (e.g., Cañas et al., 2004; Nückles, 2004; van Someren, 1994), partly automated ways, like Pathfinder (e.g., Schvaneveldt et al., 1985) or NET (Eckert, 1998), and automated ways, like MITOCAR and T-MITOCAR (Pirnay-Dummer, 2006; Pirnay-Dummer, Ifenthaler, & Spector, 2009), to help with the transition from natural language to an aggregation on a graph.

Semantic content can be retraced by comparing the sets of semantic particles within a structure, e.g., concepts (edges, nodes), propositions (vertices, links) (see Chapter 6 for a more detailed description of commonly available semantic indices). Structure, however, is more difficult to trace because in the investigated knowledge artifacts there is always content. In the past, only superficial indices have been accounted for, e.g., the number of propositions within a graph (Seel, Al-Diban, & Blumschein, 2000; Al-Diban, 2002). An exception is the structure mapping engine (see Falkenhainer, Forbus, & Gentner, 1989; Forbus, Gentner, & Law, 1995), which works on acyclic graphs. While the simple indices still hold some empirical evidence, they also have their evident limits: Today we know that the size of a model alone is not a sufficient indicator for quality or even expertise (see Ifenthaler, 2008). Sometimes experts work on very lean structures that are highly efficient at the same time (Glaser, Abelson, & Garrison, 1983; Glaser, 1992).

Every analysis herein is conducted on graphs. The graphs are the *result* or *way of notation* of the assessment but they do not necessarily have to be the *means* thereof. Every assessment method which data may be transformed to a graph may be analyzed by the algorithm. The structural comparison will obviously be only as good as the assessment helps to map knowledge onto a graph. If the assessment is vague and not valid, reliable, and objective, the analysis will also fail.

13.3 Completeness and Explanatory Power

Very often, common indices from graph theory are chosen to refer to the structure of knowledge, e.g., the number of links (surface, see Chapter 6) or measures of

diameter (graphical matching, see Chapter 6). Such indices have probabilistic co-occurrences with the quality of a structure. They are suitable in most typical cases and reflect single individual properties of the surface of a graph. In terms of graph theory, the measures are complete as they describe details of the graph. In terms of the description of knowledge structure, however, they are incomplete inasmuch as they do not resemble, account for, or reconstruct measures of the inner structure. In exchange they can be computed very fast. In this chapter I will introduce an algorithm which accounts for the inner structure as it is important for the understanding of knowledge. I call this approach "complete" because it retraces every structural component and analyzes the structure on the basis of its parts. Although it is also still a heuristic, it works on a far more detailed level. This does not at all mean that I consider other indices as being "weak" in terms of their explanatory power – which is clearly not the case. I do, however, consider them incomplete because they focus on singular aspects on the surface of graphs. The following example will illustrate this statement:

The three very simple example graphs in Fig. 13.1 have clearly distinguishable structures. However, they all have the same number of links (surface) and spanning trees of the same diameter (graphical matching index). The first and third graphs are their own spanning trees and their diameter is 3. If the second graph lost one of the outside links (of the square) and additionally either the diagonal or the opposite side, the diameter of its spanning tree would again be 3. The number of links (surface) is 5 in all three cases. Thus, on both graph surface-oriented measures the structures seem to be alike. In the following paragraphs, I introduce the structural matching index. To do so, I start with a notion of simple structures and a preliminary structural notation, from which I develop the complete structural traces.

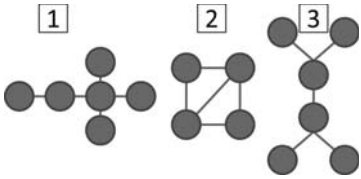


Fig. 13.1 Three different structures with the same surface and graphical matching indices

13.4 Simple Structures and a Preliminary Structural Notation

In this section, I introduce some basic underlying ideas of complete structure comparison. The examples and inductive derivations helped me understand the nature of knowledge structure. Within a graph there are various possible paths between vertices. In some graphs, not all vertices may be interconnected. A path within a graph may be

(I)→ (fire)→ (alert)→ (firefighters)→ (water)→ (no fire)

The path consists of the concepts and the links. There may be more concepts and more paths within the whole graph. Also, there are more paths in the example

above, e.g., (II) (firefighters) → (water) or (III) (alert) → (firefighters), which may play a role in a completely different reasoning model. Especially (III) could be used in a completely different way:

> (III) Peter, please don't play with the button. If you press the *alert* button, then *firefighters* will appear.

Also, several different desired states may be involved in (III): Peter may want firefighters to show up while the speaker does not want to call them. Model structures are reusable for all kinds of situations. If I wanted to look at the structure without regard to the content, I could transform the example in (I) to a structurally tagged form like this:

(Ia) (A) → (B) → (C) → (D) → (E)

The transition (Ia) from the actual concepts to tags for the structural components already allows a better structural comparison. It resembles a simple sequence of things. On this level, I could look for other models which contain sequences like this and find out whether a particular group of experts also construct their knowledge in the same way:

(IV) (crash) → (alert) → (ambulance) → (emergency room) → (healing)

The underlying model of (IV) is structurally identical to (I) and can therefore still be represented by a structure like in (Ia) because the vertices do not resemble the concepts anymore but are more general structural components which can be mapped to similar structures (like in this case: a sequence). Thus, (A) does not refer to any concept (e.g., fire). It states only that there is a concept which is followed by another concept, no matter what the concepts contain in particular. The structural similarity between (I) and (IV) would be 1 because their structure is identical, namely (Ia). Similarities are usually represented as being between $s = 0$, which means exclusion or no similarity, and $s = 1$, which means identity. The same type of example could be used for circular graphs, e.g.:

(V) (ocean) → (sun) → (wind) → (clouds) → (rain) → (river) → (ocean).

The structural components from a cycle like (V) would simply be:

(Va) (A) → (B) → (C) → (D) → (E) → (F) → (A)

The only difference from a sequence is that the last vertex is also the first in both (V) and (Va). Everything else stays the same. We can now compare the structures (Ia) and (Va) and consider the frequencies of the intersection of the components compared to the union set to calculate the similarity:

$$s = \frac{f(I_a \cap V_a)}{f(I_a \cup V_a)} = \frac{5}{6} = 0.833$$

Later on I will show that similarity measure as introduced by Tversky (1977) is a better measure of comparison, especially when the structures being compared have different sizes. So far, this simple notation and visual comparison seems to work fine

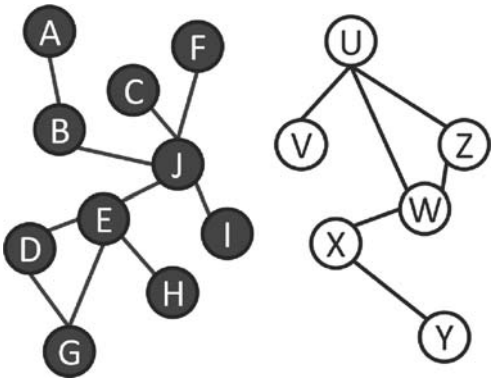


Fig. 13.2 Two different knowledge structures represented on graphs

for sequences and circles. But does it also work for other, more complex structures? Figure 13.2 shows two already abstracted and more complex structures.

I could try to intuitively puzzle together some similarities, e.g., assuming that the geometrical information actually provides meaningful evidence. For instance, I could look for geometrical similarities or congruencies and try to match the triangle (U) – (Z) – (W) – (U) to the other triangle (D) – (E) – (G) – (D). However, the interpretation of the geometric information around the triangles is a trap: Graphical representations have to fulfill visual constraints before content-directed constraints to prevent overlaps or to minimize the overlaps of the edges (see Ellson, Gansner, Koutsofios, North, & Woodhull, 2003). Some visualization algorithms take weights between links into account, but only if there is space.

Two-dimensional scalability (representability) of usually complex structures (about 30 propositions) has a stress (error) of about $S = 0.3$. Nonparametric multidimensional scaling revealed stress values from $S = 0.254$ to 0.316 (mean $= 0.284$, and $SD = 0.019$) throughout eight different validation studies (Pirnay-Dummer, 2006). But this value corresponds only to a distribution of pairwise adjacent concepts when no overlap-constraints and no links are involved. Structural similarity is therefore not a matter of a good eye. The graphical constraints can lead to a different visualization when only a single additional link is embedded, and even trained readers of graphs may be confused when a certain level of model complexity is reached. This can already be the case even if the graph still seems to be small. Moreover, visual judgment cannot be automated and computed either beyond the simple cases of sequences or cycles.

13.4.1 Complete Structural Traces

The goal of achieving structural computability of knowledge structures will require a different approach which can help with the analysis. The model in Fig. 13.2 was constructed in an abstract way to show that the method will work without regard to the semantic content. From this representation to the one that will be needed for

an automated analysis of the paths, it is again only a single simple step. Instead of giving the structural component nodes nominal tags – like in (Ia) – every vertex is tagged with the number of edges that are connected to it.

This leads to a structurally identical graph, as Fig. 13.2 translates directly into Fig. 13.3. Except for the content of the nodes, the two graphs are identical. If interpreted separately, the frequencies show a level of embeddedness in relation to the structure. Once mapped back to the structure again, every possible path between any of the vertices via the edges within Fig. 13.3 is a structural component of the graph, like a piece in a complex puzzle. One path of Fig. 13.2 is:

(VI) (A) – (B) – (J) – (E) – (G)

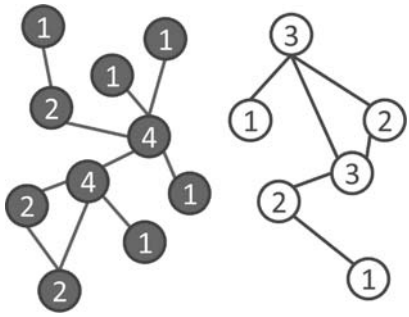


Fig. 13.3 Trace Graph: recoded model structure with adjacency information on the nodes

Within the trace graph (Fig. 13.3) this path translates to:

(VIa) (1) – (2) – (4) – (4) – (2)

The same kind of trace through the structure can be generated for every single possible path within a graph, providing a list of individual structural traces throughout the graph. Even for smaller graphs, it is too labor intensive – and maybe also not very fulfilling – to construct the list manually.

To rebuild the structure by algorithm, the degree of any single vertex is calculated first. The degree is the number of edges which are connected to this specific node:

$$\Gamma(v)$$

For each vertex v_i every possible path to the rest of the graph is generated. Thus, chains K_n of adjacencies are constructed with the set of degrees $K_n = \{k_1 \ldots k_\upsilon\}$. Although not theoretically needed, υ is a necessary stop criterion in most applications – especially in web applications. υ constitutes the maximal search spread (length of traces). Empirical implications of this constraint will be discussed later on. V_v is all vertices, and u refers to the different path lengths. All structural traces of the graph are contained in the set:

$$\Gamma^{V}_{\upsilon,i}(V_v) = \bigcup_{i=1}^{n} \bigcup_{u=2}^{\upsilon} \Gamma_{i,u}(v)$$

If a trace leads back to a previously contained vertex (cycle), it will tag the last vertex in the list as signed, e.g., (2) – (4) – (2) – (–2) for the cycle (D) – (E) – (G) – (D) from Fig. 13.2. In this way, a graph may be cyclic and still be analyzed. Cycles (loops) also play an important role in the understanding of systems in general (Seel, 2003). Traces with a length of 1 are ignored ($u = i + 2$), because every graph which did not consist solely of isolated vertices (no edges) would then be structural similar to another. Such a similarity would yield no theoretical plausibility because it would refer to properties which by definition every model has; and the main goal is to distinguish models structurally. Traces can be built on any kind of graph and for any kind of structural annotation and weight. The algorithm does not take anything into account except for the existence of edges (nodes) between vertices (links).

13.4.2 Downtrace

Although the complete structural traces in the form of an algorithm already provide a computable structure reconstruction which may be a good basis for model comparison, the method has still a weak spot. An easy example will show how.

Figure 13.4 shows two very similar model structures and their adjacency structure, which will be used to derive the list of traces $\Gamma^{V}_{\upsilon,i}(V_v)$. If only the accountable traces of both models were compared, then similarity between models would be masked because only full paths would be compared, e.g., the following structures might not match although they have a structurally meaningful correspondence:

(VII) (A) – (B) – (C) – (D)
(VIII) (W) – (X) – (Y) – (Z)
(VIIa) (3) – (2) – (3) – (2) – (–3)
(VIIIa) (2) – (2) – (2) – (2) – (–2)

A minus ("–") is applied when the trace reaches a previously visited node, thus indicating a cycle. A cycle indicator can only occur at the end of a trace – the trace

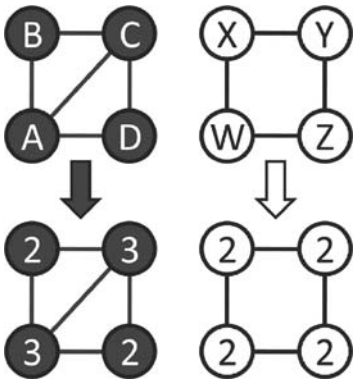


Fig. 13.4 Traces of similar model structures

stops after closing a cycle. Otherwise the trace would map the same structures more than once. When we count the matching traces (VIIa) and (VIIIa) later on to look for similarity, the structurally almost identical cycles would not match each other and therefore generate no tendency toward more similarity in the overall comparison between the two models. Thus, the additional structure between (A) and (B) within Fig. 13.4 should convey a difference between the two graphs but only to a certain ratio compared to the whole structure. More precisely: The structural similarity must correspond to the difference (one additional edge) and still preserve the consistency. This could of course easily be weighted when the structures are small and visible – in the example (Fig. 13.4) we actually *see* the cycle and want it to be accounted for. In the automated analysis, and also with larger graphs, this subjective plausibility cannot be generated. Therefore, the previous algorithm needs one more enhancement. After that, the problem of the masked structures will be solved.

$$\Gamma^{V}_{\upsilon,i}(V_{\nu})$$

is the list of traces of the whole graph. It will be augmented by the following algorithm and it will significantly expand the list of possible traces. For every single trace there is a *downtrace* with each individual degree being between $2 \leq |d| \leq |\nu|$, such that all possible combinations beneath the previously entailed degrees are generated by the function Ξ. d stands for each degree in the existing trace and ν is another stop criterion:

$$\Xi\left(\Gamma^{V}_{\upsilon,i}(V_{\nu})\right)$$

For (VIIIa) the downtrace function generates nothing but the initial trace, because all degrees are already 2. However, for (VIIa) six additional traces are generated:

(VIIb)(2) – (2) – (3) – (2) – (– 3)
(VIIc)(3) – (2) – (2) – (2) – (– 3)
(VIId)(3) – (2) – (3) – (2) – (– 2)
(VIIe)(2) – (2) – (2) – (2) – (– 3)
(VIIf)(3) – (2) – (2) – (2) – (– 2)
(VIIg)(2) – (2) – (2) – (2) – (– 2)

Only (VIIa) really exists on the surface of the graph, while (VIIb) to (VIIg) are generated by the downtrace function. They do, however, resemble the previously masked structures which are also contained in the graph. In this case, (VIIg) would match (VIIIa) whereas there would be no match without the downtrace. The stop criterion ν limits the downtrace function in a way that degrees above ν would not get a complete downtrace but rather a downtrace from ν down to 2. For example, if there is a degree of 20 connected to a degree of 3 in a graph and ν is set to 7, the following would be generated: (20) – (3) → (7) – (3) → (6) – (3) →... → (3) – (3) → (3) – (2) → (2) → (2). ν is needed in most practical applications because otherwise computing would be too greedy (demanding) due to the

exponential complexity of the downtrace function. Study 1 of this chapter also investigates the empirical differences between the complete and the constraint downtrace function.

13.4.3 Structural Matching Similarity Measure

At the end of the whole process a single structural similarity measure between pairs of graphs (A and B) is desired. To compute this measure, the similarity index of Tversky (1977) is used. ΞA and ΞB are two downtraces

$$\Xi_{\mathrm{A,B}}\left(\Gamma^{V}_{\upsilon,i}(V_{\nu})\right)$$

from the two graphs. $(\Xi_{\mathrm{A}} \cap \Xi_{\mathrm{B}})$ is the intersection between the two downtraces. $(\Xi_{\mathrm{A}} - \Xi_{\mathrm{B}})$ and $\Xi_{\mathrm{B}} - (\Xi_{\mathrm{A}})$ are difference sets. The former difference set contains all of the traces from graph A which cannot be found in graph B and the latter contains the traces from graph B which cannot be found in graph A.

$$s = \frac{f(\Xi_{\mathrm{A}} \cap \Xi_{\mathrm{B}})}{f(\Xi_{\mathrm{A}} \cap \Xi_{\mathrm{B}}) + \alpha \cdot f(\Xi_{\mathrm{A}} - \Xi_{\mathrm{B}}) + \beta \cdot f(\Xi_{\mathrm{B}} - \Xi_{\mathrm{A}})}$$

Depending on the research question and the methodology in which the structural matching index is used, α and β are either equal ($\alpha = \beta = 0.5$) or weighted differently. α and β may be proportionally weighted if different types of knowledge constructs are being compared, e.g., if a learner has half an hour to construct a re-representation, text, or graph and the result is later compared to a huge concept map which took several experts weeks to develop. In this case, researchers may want to control for the different sizes of the model, e.g., linearly:

$$\frac{\alpha}{\beta} = \frac{f(V_{\mathrm{B}})}{f(V_{\mathrm{A}})}$$

In this case, the expert map would be systemically larger than the learner's map due to the different tasks during the externalization process. Thus, the difference set for the experts $(\Xi_{\mathrm{A}} - \Xi_{\mathrm{B}})$ would also be systematically larger than the difference set for the learners $(\Xi_{\mathrm{B}} - \Xi_{\mathrm{A}})$. To control for this, α and β may be weighted as suggested above to allow a better tracking of a learner's progress. If the available theory or the methodological constraints provide other, e.g., nonlinear, dependencies between the difference sets, then they can easily be used to control the effect of the difference sets. If, however, like in most research designs, the same procedure and the same amount of time is implemented during externalization, then α and β should be equal even if the models turn out to be completely different in size. The sum of $\alpha + \beta$ also should always be 1 when used for structural comparison.

13.5 Studies 1 and 2: Trace-Based Structural Complexity Measure

Graphical representations can have different internal complexities. A side effect of the availability of the structural matching similarity measure is the availability of the traces which are constructed in the process.

An interesting and easily accessible value is the number of traces yielded by the downtrace function.

$$f\left(\Xi\left(\Gamma^{V}_{\upsilon,i}(V_{v})\right)\right)$$

In two identical validation studies on different subject domains, I investigated the structural complexity measure as a potentially selective empirical index on the basis of the complete structural traces and the downtrace function. Hence, the hypotheses for both studies are:

H_0: The trace-based structural complexity measure for non-experts is higher than or equal to that for experts:

$$f_{\mathrm{N_E}}\left(\Xi\left(\Gamma^{V}_{\upsilon,i}(V_{v})\right)\right) \geq f_{\mathrm{E}}\left(\Xi\left(\Gamma^{V}_{\upsilon,i}(V_{v})\right)\right)$$

H_1: The trace-based structural complexity measure for non-experts is lower than that for experts:

$$f_{\mathrm{N_E}}\left(\Xi\left(\Gamma^{V}_{\upsilon,i}(V_{v})\right)\right) < f_{\mathrm{E}}\left(\Xi\left(\Gamma^{V}_{\upsilon,i}(V_{v})\right)\right)$$

13.5.1 Methods

To validate the structural complexity measure, I conducted two studies where aggregated knowledge structures (models) from groups of experts were compared to models from groups of non-experts. The models were assessed and aggregated with MITOCAR (see Chapter 6 for a description of this tool).

The software was initially built to test for hypotheses between groups. To do so, the aggregation keeps the number of propositions which are represented in the aggregated graph constant. This means that all of the four presented and compared models consist of 30 propositions (surface = 30). Only the 30 strongest links are represented in the externalized graph within MITOCAR (this does not apply for the individual model assessment tool T-MITOCAR, see Chapter 6). Thus, if there are differences between the structural complexities, it would not be the number of propositions which makes the expert models more complex but the internal structure. The hypotheses are therefore not tested against whether there is just more

content in the model but rather against how the structure is composed in itself. For the two studies, this means that the surface is a controlled variable.

Both the expert models and the non-expert models were assessed in the same environment and under the same conditions. The main hypothesis of the study was about semantic and structural differences between experts and non-experts. All within-group homogeneity measures were tested for each group and yielded no significant derivation of the individuals from the group aggregated model. No compensation was paid for participation in the study in either group. The demographics of the two studies are reported for each study separately.

13.5.2 Study 1: In the Field of Learning and Instruction

In this study a group of 18 experts (13 female and 5 male) from the field of learning and instruction was assessed for their consensus model on knowledge transfer. The aggregated model of the experts was compared to a model constructed by a group of 15 non-experts (6 female and 9 male). The trace-based structural complexity measure between both models revealed differences.

Figure 13.5 shows the difference in complexity between the non-experts and experts in the field of learning and instruction. The differences are statistically significant against the null hypothesis of an equal distribution of complexity between the models ($\chi 2 = 68{,}831$; $df = 1$; $p < 0.05$). The ratio between the experts and the novices is Rn,e = 3.00: The experts have three times as many traces as the model of non-experts. For the first study, the H1 can be accepted.

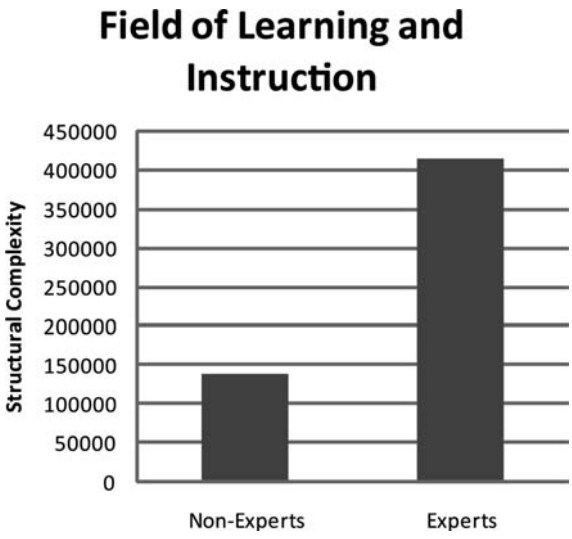


Fig. 13.5 Trace-based structural complexity within-group models, field of learning and instruction

13.5.3 Study 2: In the Field of Economics

An identical study was conducted with a group of 19 experts (13 female, 6 male) from the field of economics. The experts were compared to a group of 35 non-experts (26 female, 9 male). The difference in the sample sizes is a bit unfortunate. Due to organizational constraints, the non-experts were assessed first. Unfortunately, only 19 experts could be acquired for the study. Like in study 1, the aggregation of the group models was used for the comparison. Thus, the sample size could have only affected the homogeneity or variances of the output model. This was, however, tested and no significant differences in the homogeneity or variances were found (Pirnay-Dummer, 2006, pp. 181–185, 198–201). Both groups were assessed for their understanding of economic cycles. The main hypothesis of the study was also about semantic and structural differences between experts and non-experts. As mentioned, all within-group homogeneity measures were tested for each group and yielded no significant derivation of the individuals from the group aggregated model. No compensation was paid for participation in the study in either group. The trace-based structural complexity measure between the two models again showed remarkable differences.

Figure 13.6 shows the difference in complexity between the non-experts and experts in the field of economics. The differences are statistically significant against the null hypothesis of an equal distribution of complexity between the models ($\chi 2 = 711{,}476$; $df = 1$; $p < 0.001$). The ratio between the experts and the novices is Rn,e = 17.49: The expert model has more than 17 times as many traces as the model of the non-experts. Thus, the H1 hypothesis can also clearly be accepted for the second study.

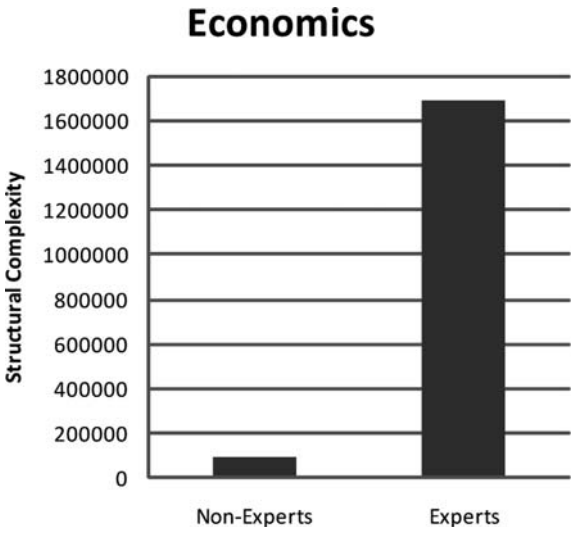


Fig. 13.6 Trace-based structural complexity within-group models, field of economics

13.5.4 Discussion of Studies 1 and 2

In general, it will not come as a stunning surprise that experts impart higher complexities to their externalizations. In the methods description I pointed out the specific feature of MITOCAR of keeping the number of edges constant. The differences therefore have to be generated by different internal structures, which is the point of complete structural comparison. The results of both studies can be interpreted as an additional construct validity measure (selectivity) and as an empirical criterion to support the applicability of the structural trace measures. It can also be seen as empirical evidence that experts structure their knowledge in a completely different way: Not only do they have more knowledge, their knowledge is integrated differently. Critics might still suspect that this refers to a kind of inner integratedness or connectedness of the expert structures, and that this could have also been discovered with a density measure, like γ (see Chapter 6 for details). However, the density of all of the models was almost the same (ranging from $\gamma = 0.32$ to 0.38) and did not yield any general or specific differences in this case.

13.6 Study 3: Technological Study on the Sensitivity of Structural Matching

To test the sensitivity of the structural matching index, I conducted a technological study. The research question of this study addresses how the structural matching index will react to random distortion of an initial model. The hypothesis of this study states that the structural similarity index will be lower depending on the number of distortions applied to a model. According to the theory which led to the structural matching algorithm, a model with more random distortion should decrease in similarity when compared to the original basic model.

H_0: The number of distortions does not correlate or correlates positively with the structural matching index when a distorted model is compared to its original model ($r_d \geq 0$)

H_1: The number of distortions correlates negatively with the structural matching index when a distorted model is compared to its original model ($r_d < 0$).

13.6.1 Methods

In order to test the hypothesis of this study, 12 basic model types of linked pairs (propositions linking two concepts each) were computer generated, including two simple model types (sequence and circle) and ten complex model types (hierarchies). For every model type, basic models were generated for a number of propositions, e.g., a sequence is built for a surface of $\Theta = 5$ to $\Theta = 41$ propositions, as were all of the other model types ($np = 37$ different surfaces for each model

type). Thus, there were $M = 450$ basic models. For each basic model a distortion was generated afterward by randomly replacing existing links with new ones. The number of distortions varied from $d = 1$ to 15. Additionally, each random distortion track (1–15) was repeated 10 times, thus leading to different distortion structures at every trial. This was done to assure that there was no effect from a single random track which may or may not have yielded atypical structures. Every resulting model was compared to its original basic model, which leads to $n = 64{,}800$ model comparisons. Afterward, the mean of the structural matching index for each level of model distortion was correlated to the corresponding level of distortion ($nm = 4{,}320$ for each cell), and the degrees of freedom were assumed on the level of aggregation ($df = 13$). This is methodologically more conservative since the amount of data in the dataset would make almost any correlation statistically significant.

13.6.2 Results

The main hypothesis aims at a correlation between the structural matching index and the randomized distortions.

Figure 13.7 shows the almost linear dependency between the level of distortion and the structural index. The randomized distortion of the model correlates negatively with the structural similarity index ($r = –0.0.99$). This correlation is statistically significant ($t\,(13) = –26.1271$, $p < 0.001$).

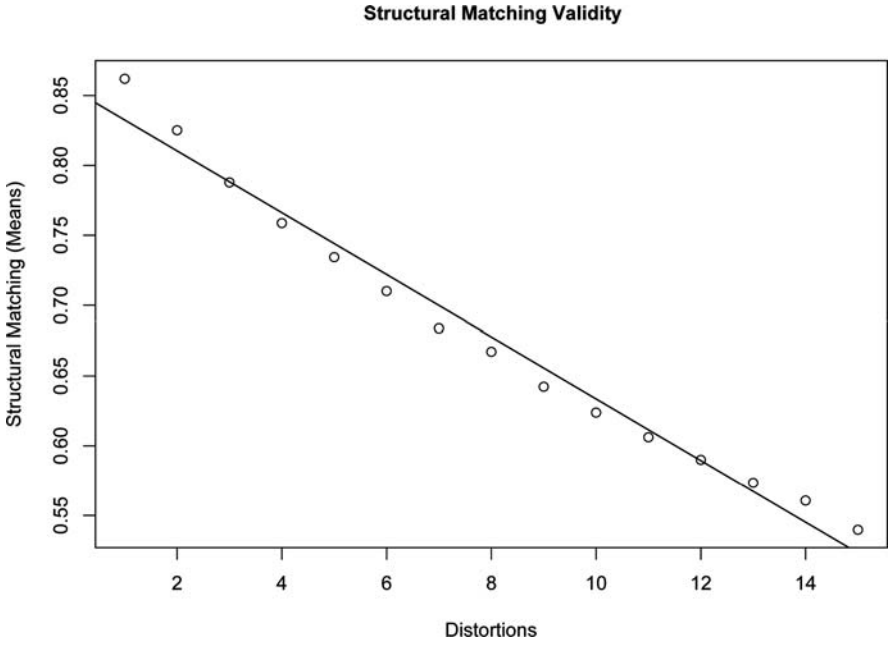


Fig. 13.7 Structural matching validity as dependency from randomized model distortions

13.6.3 Post Hoc Analysis

The data revealed two more interesting aspects. The first one seems almost obvious a posteriori. There is also a considerable correlation between the number of propositions and the structural matching index.

Figure 13.8 shows the effect of the surface: The larger two compared models are, the more likely it is that they share structure. The correlation is $r = 0.72$ for surfaces between $\Theta = 5$ and $\Theta = 41$ propositions. The correlation is also statistically significant ($t\,(35) = 6.1782$, $p < 0.001$). Nonlinear correlations (such as a saturation curve) did not yield results which were distinguishably different from the linear correlation.

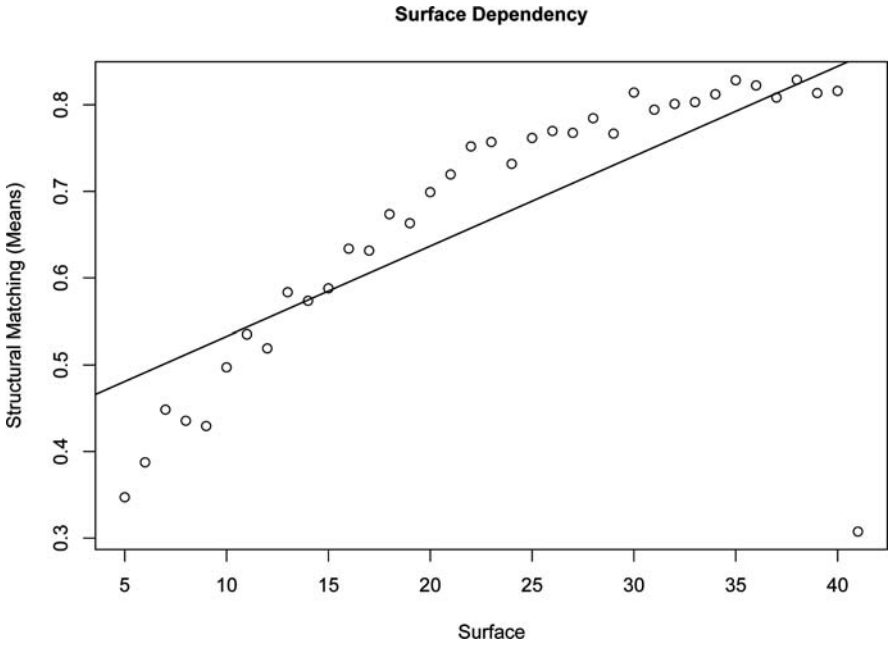


Fig. 13.8 Dependency between the surface (number of links) and the structural matching algorithm

The study also tested the difference between the complete and the constraint downtrace function. The internal convergent validity between the complete (no ν) and the constraint downtrace function ($\nu = 5$) was $r = 0.962$.

13.6.4 Discussion of Study 3

The distortions of the models triggered the structural similarity index almost linearly in the predicted way throughout both simple and complex model structures.

Therefore, I consider the algorithm to be sensitive to structural differences. The high correlation between the surface (number of propositions) and the structural matching shows limits in that the algorithm may not be sufficient if it is applied to very large graphs. Larger graphs should thus be analyzed by means of different methods. This also corresponds to the technical limitations of the algorithm: In the current form it has exponential complexity, which makes it very slow for the analysis of graphs with more than 150 propositions. This problem is known and the structural graph comparison is generally considered an NP-hard problem even for acyclic graphs (see Gentner & Markman, 2006). The constraint downtrace function which uses constraint path traces is sufficient for analysis and also opens the algorithm for real-time analysis, e.g., in web applications like T-MITOCAR.

13.7 Study 4: Empirical Study on the Semantic Interference with Structural Matching

Linguists assume on the basis of many multilingual studies that structure is not independent of semantics. A good overview of the discussion on autonomy can be found in Jackendoff (2007) or Taylor (2007). An introduction is provided by Langacker (2008). Clearly, this study and setup will not be able to contribute to this research. The focus here lies on the algorithm and its limits. However, the interdependency or interference of the semantic context may be of interest for researchers who want to use structural matching for their analysis. The algorithm is in itself only directed at structure. As shown, the content is not taken into account in any way. From a solely methodological and formal point of view, we would therefore assume that:

H^0: There is no semantic interference with structural matching: Structural matching is independent of content $\left(\frac{\sum_1^n \Delta\mu_n}{n} = 0\right)$.

The differences that may occur on the structural matching between contents should explained as resulting from general interdependencies between structure and semantics. Following the dependency theory, we would have to expect:

H_1: There is semantic interference with structural matching: Structural matching is content-dependent $\left(\frac{\sum_1^n \Delta\mu_n}{n} > 0\right)$.

13.7.1 Methods

To test the hypotheses, 13 diverse topics were selected. They included: iceberg lettuce (1), intrinsic motivation (2), constitution (3), astronautics (4), school (5), statistics (6), social sciences (7), perl (8), geodesy (9), refraction (10), chess (11), mental models (12), and reliability (13). The numbers appear again in the results.

For each topic, we took the first 60 documents yielded by the Google search engine. The topic was entered as search term/phrase. We used the Google engine because the search algorithms of Google have shown high semantic reliability in several studies (e.g., Bar-Ilan, 2001; Cilibrasi &Vitanyi, 2007; Janetzko, 2008). Model structures were generated for every document individually for documents which contained more than 350 words – the same way as in the third study. If the associations were again strong enough in the text (meaning: not random), the model was taken into the pool for its topic. Afterward, the models were compared in two ways: First, all the models within a topic were compared to each other ($N_1 = 14{,}421$ individual comparisons). Second, all the models between the topics were compared to each other in order to see whether there is less similarity when models with different content (topics) are compared structurally ($N_2 = 174{,}384$ individual comparisons). The latter comparisons may not be cross-validated by semantic means because they would always yield no similarity – for obvious reasons.

13.7.2 Results

Starting with the first corpus of comparisons, the results show that the means of the within topic comparisons differ between topics.

Although the interferences within topics differ, the variances are very high within each group as can be seen from the boxplot in Fig. 13.9. The differences are statistically significant ($F(1, 14{,}419) = 97.896$, $p < 0.001$), but the effect from the content

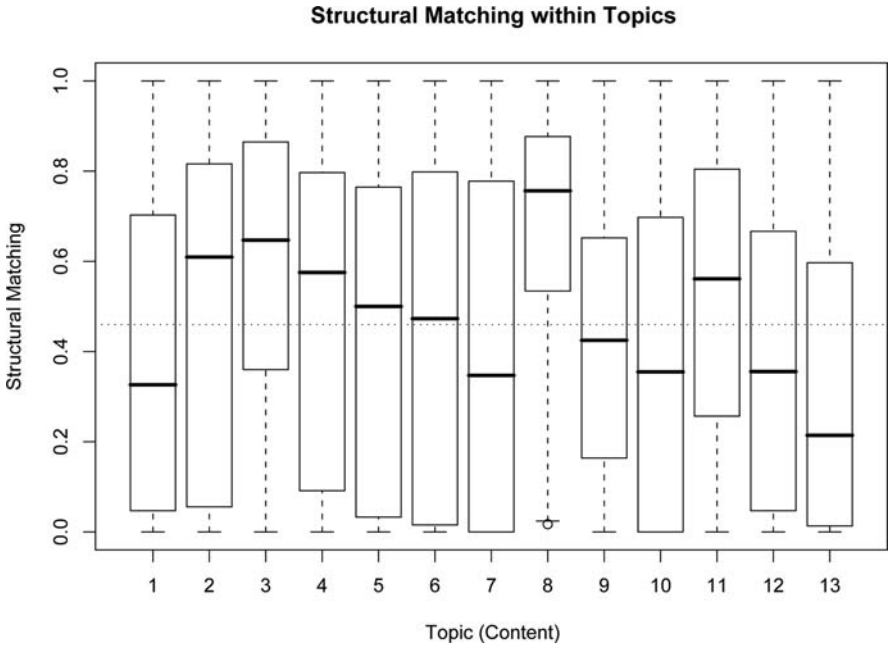


Fig. 13.9 Structural matching within each different topic ($N = 14{,}421$)

is so low ($\eta^2 < 0.0001$) that it does not make any sense to interpret it at all. Therefore, and despite the statistical significance, I will keep the null hypothesis at this point: There is no semantic interference with structural matching: Structural matching is independent of the context of the content.

Given the results from the within analysis, the results of the comparison between the different content groups will not be much of a surprise.

Figure 13.10 shows the pairwise comparisons of the individual models between the different topics. The coding at the x-axis is combined by the indices from the two groups under comparison: 102 means that topic 1 was compared to topic 2, and 1,012 means that topic 10 was compared to topic 12. The overall structural matching has a mean of 0.45 and hence is the same as that within the topics. Again, the differences are statistically significant ($F(1, 174{,}382) = 97.896$, $p < 0.001$) but the effect is not visible ($\eta^2 < 0.0001$). Thus, I will also keep the null-hypothesis for the between topics comparison.

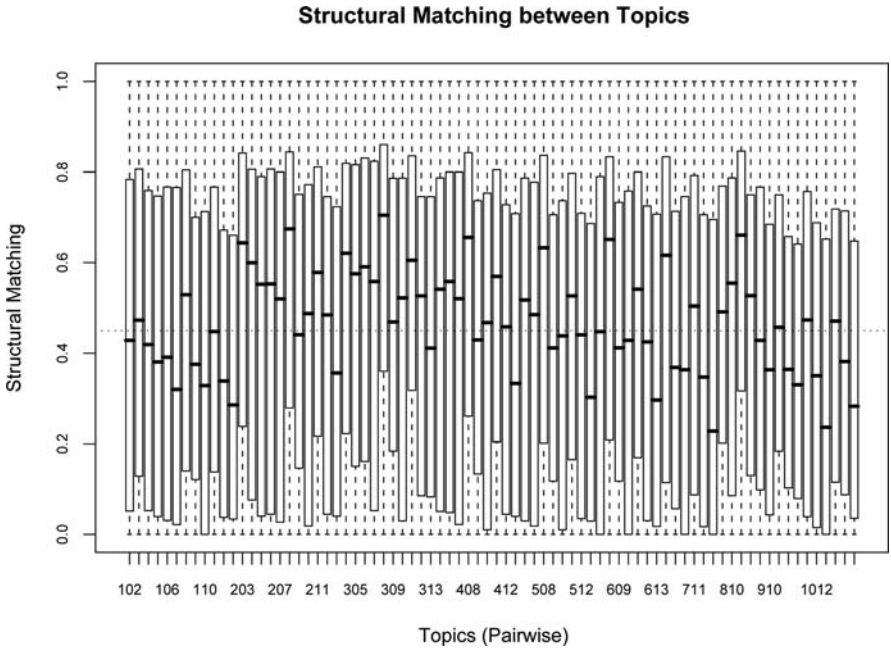


Fig. 13.10 Structural model matching between topics ($N = 174{,}384$)

13.7.3 Post Hoc Analysis

In the post-hoc analysis, I checked for the correlations between the available semantic indices and the structural matching index within each topic. Please refer to Chapter 6 of this book for details about the semantic indices.

Table 13.1 Correlations between semantic indices and the structural matching index

	Concept matching	Propositional matching	Balanced semantic matching
Structural matching	0.25[a]	0.14[a]	–0.01

[a] The correlations are statistically significant against the assumption of no correlation ($\rho \neq 0$).

There are correlations between the structural matching index and the concepts and propositions within this study. The only considerable correlation would be the one on the level of concepts (concept matching). This can also be interpreted as an indicator for divergent validity between the different measures. In Chapter 6 there is an overview on convergent and divergent validity which contains all currently available measures from SMD and MITOCAR.

13.7.4 Discussion of Study 4

The structure within each topic had too much variance to derive any meaningful effect from it. The fact that the structural matching index did not correspond to the different topics does not mean that there is independence between structure and semantics – the correlations of the post hoc analysis show a different perspective on this issue. This only means that the structural matching index does not correspond to semantic context. One might ask whether this is good or not. The results simply mean that semantic measures and the structural matching index aim at different constructs. This is to be assumed at least at the level of how structure and semantics may heuristically be measured within the graph-oriented approaches so far.

13.8 Comparison to Heuristic Measures of Structure

Complete structural comparison is complex in terms of the computing resources it needs. Not every time is a complete structural comparison needed. In some cases heuristics may suffice. Ifenthaler (2008) investigated several graph features and their correspondence to knowledge and learning. The methodologically strongest among them come from the SMD Technology (Ifenthaler, 2006): *surface matching* and *graphical matching* (also see Chapter 6 for details). They also measure structural properties in two different ways. However, research questions about knowledge are rarely simple to investigate. In this case, it makes sense to use multiple methods and measures. A closer look at the convergent validity supports this strategy. The validation study was conducted on a coherent text corpus from a pharmaceutical

Table 13.2 Convergent validity measures (correlations)

$N = 1{,}849{,}926$	Surface matching	Graphical matching
Structural matching	0.63	0.48
Surface matching		0.79

company. $N = 1{,}849{,}926$ pairwise model comparisons were analyzed for this study. Please refer to Pirnay-Dummer et al. (2009) for the full study.

Table 13.2 shows the convergent validities between structural matching, surface matching, and graphical matching. There is consistency between the measures, but they do not measure the same things. They cannot be considered as being parts of a coherent scale. All three of the measures aim at structure: Surface matching counts the number of vertices within a graph. Graphical matching is the diameter of the spanning tree (Kruskal, 1957) and thus like the "width" of a graph. Structural matching compares the building blocks of a graphical structure by analyzing them separately.

13.9 Conclusion

Structural matching is about graphical structure only. It accounts for structure as the way in which the whole is composed of simple and complex substructures: structural parts or "puzzle pieces." It does not necessarily correspond to natural language syntax, although associatedness will be detected by it. The results of the first two studies in this chapter indicate that the index is capable of predicting expertise. To fully answer this question, more studies would of course be needed. The third study showed that the algorithm is almost linearly sensitive to structure and will find structural differences as long as they are in the graph. The fourth study investigated the interdependence between the structural matching and semantic indices. If at all, low dependencies were found, and the context of the topic had no interpretable effect on structural matching.

In future research, the index may help to tell whether a specific assessment method is suited to reliably reconstruct knowledge on a graph. Also, hypotheses about general structuredness of expertise may be investigated with the algorithm. This will be particularly interesting for research questions which focus on knowledge structures between domains. It needs to be clearly stated that the transition process from thought and knowledge onto a graph, be it by means of natural language or somehow directly, will have to be methodologically sound. Only then will the algorithm work properly. It cannot transcend the method of assessment and it cannot complement any possible methodological weaknesses of such methods. Like very often, it makes sense to use multiple measures on the same constructs. To fully capture structure, more than one index should be applied.

The index which I discussed in this chapter is integrated into several toolsets: MITOCAR, T-MITOCAR, HIMATT, and AKOVIA. Within these tools, the measure is not applied or reported alone but always alongside other structural and semantic measures to provide a better description of the complex construct which we call *knowledge*.

References

Al-Diban, S. (2002). *Diagnose mentaler modelle* [*Assessment of mental models*]. Hamburg: Kovac.

Bar-Ilan, J. (2001). Data collection on the Web for informetric purposes: A review and analysis. *Scientometrics*, *50*(1), 7–32.

Cañas, A. J., Hill, G., Carff, N., Suri N., Lott, J., Eskridge, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak, & F. M. González (Eds.), *Concept maps: Theory, methodology, technology, proceedings of the first international conference on concept mapping*. Pamplona, Spain: Universidad Pública de Navarra.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity distance. *IEEE Transaction on Knowledge and Data Engineering*, *19*(3), 370–383.

Eckert, A. (1998). *Kognition und Wissensdiagnose. Die Entwicklung und empirische Überprüfung des computergestützten wissensdiagnostischen Instrumentariums Netzwerk-Elaborations-Technik (NET)*. Lengerich: Pabst.

Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2003). *GraphViz and dynagraph. static and dynamic graph drawing tools.* Florham Park, NJ: AT&T Labs – Research.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

Forbus, K. D., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998). Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Journal of experimental and Theoretical Artificial Intelligence*, *10*(2), 231–257.

Gentner, D., & Bowdle, B. (2008). Metaphor as structure-mapping. In R. Gibbs (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 109–128). New York: Cambridge University Press.

Gentner, D., & Markman, A. B. (2006). Defining structural similarity. *Journal of Cognitive Science*, *6*, 1–20.

Gilovich, T., & Griffin, D. (2002). Introduction – Heuristics and Biases: Then and now. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 1–18). Cambridge, UK; New York: Cambridge University Press.

Glaser, R. (1992). Expert knowledge and process of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics*. Hillsdale, NJ: Erlbaum.

Glaser, E. M., Abelson, H. H., & Garrison, K. N. (1983). *Putting knowledge to use: Facilitating the diffusion of knowledge and the implementation of planned change* (1st ed.). San Francisco, CA: Jossey-Bass.

Gruber, H. (1994). *Expertise Modelle und empirische Untersuchungen* [*Expertise. Models and empirical studies*]. Opladen: Westdt. Verl.

Helbig, H. (2006e). *Knowledge representation and the semantics of natural language*. Berlin, New York: Springer.

Ifenthaler, D. (2006). *Diagnose lernabhängiger Veränderung mentaler Modelle Entwicklung der SMD-Technologie als methodologisches Verfahren zur relationalen, strukturellen und semantischen Analyse individueller Modellkonstruktionen*. Freiburg: FreiDok.

Ifenthaler, D. (2008). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development.* doi: 10.1007/s11423-008-9087-4

Jackendoff, R. (2007). Linguistics in cognitive science: The state of the art. *The linguistic review*, *24*(4), 347–401.

Janetzko, D. (2008). Objectivity, reliability, and validity of search engine count estimates. *International Journal of Internet Science*, *3*(1), 3–77.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: L. Erlbaum.

Johnson-Laird, P. N. (1983). *Mental models. Toward a cognitive science of language, inference and language*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove: Lawrence Erlbaum.

Kant, I. (1787). *Critik der reinen Vernunft* [*The critique of pure reason*]. Riga: J. F. Hartknoch.

Kruskal, J. B. (1957). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceeding of the American Mathematical Society*, *7*, 48–50.

Kubose, T. T., Holyoak, K. J., & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of memory and language, 47*(3), 407–435.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford, New York: Oxford University Press.

Link, G. (1979). *Montague-Grammatik*. München: Wilhelm Fink.

Montague, R. (1974). *Formal philosophy: Selected papers of Richard Montague*. New Haven, CT: Yale University Press.

Nückles, M. (2004). *Mind maps und concept maps: Visualisieren, Organisieren, Kommunizieren.* München: Deutscher Taschenbuch-Verlag.

Pirnay-Dummer, P. N. (2006). *Expertise und Modellbildung – MITOCAR*. Freiburg: FreiDok

Pirnay-Dummer, P. N. (2008). Rendezvouz with a quantum of learning. Effect metaphors, extended design experiments and omnivariate learning instances. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction. Essays in honor of Norbert M. Seel* (pp. 105–143). New York: Springer.

Pirnay-Dummer, P. N., Ifenthaler, D., & Spector, J. M. (2009). Highly integrated model assessment technology and tools. *Educational Technology Research and Development.* doi:10.1007/s11423-009-9119-8

Preece, P. F. W. (1976). Mapping cognitive structure: A comparison of methods. *Journal of Educational Psychology*, *68*, 1–8.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., et al. (1985). Measuring the structure of expertise. *International Journal of Man-Maschine Studies*, *23*, 699–728.

Seel, N. M. (1991). *Weltwissen und mentale Modelle* [*World knowledge and mental models*]. Göttingen: Hogrefe Verl. für Psychologie.

Seel, N. M. (2003). Model centered learning and instruction. *Technology, Instruction, Cognition and Learning, 1*(1), 59–85.

Seel, N. M., Al-Diban, S., & Blumschein, P. (2002). Mental models and instructional planning. In J. M. Spector (Ed.), *Integrated and holistic perspectiveson learning, instruction and technology: Understanding complexity.* Dordrecht: Kluwer.

Snow, R. E. (1990). New approaches to cognitive and conative assessment in education. *International Journal of Educational Research, 14*(5), 455–473.

Taylor, J. R. (2007). Cognitive linguistics and autonomous linguistics. In D. Geeraerts & H. Cuyckens (Eds.), *Cognitive linguistics* (pp. 566–588). New York: Oxford University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method. A practical guide to modelling cognitve processes.* London: Academic Press.

Part IV
Application of Obtained Results

Intermezzo 4 – Using Knowledge to Support Knowing

Dirk Ifenthaler and Pablo Pirnay-Dummer

Good theories and sound research have a great chance of leading to practical improvements. The process may take time, but eventually when things are explained properly, the process succeeds; slower but usually more stable than by the use of intuitive approaches. But sometimes the odds are even more optimistic. These are the cases where the investigation itself is *part* of the improvement. The need for assessment strategies which support the process under assessment at the same time is not new. However, with new technologies at hand, at least parts of this demand can be better fulfilled. In the preceding three parts of the book the authors started with knowledge constructs, representations, and assessment methods and moved on to decisions on specific measures and reasoning. What still needs to be shown at this point is the impact the assessment, the interpretation, the aggregation, and methodological decisions have on knowing and the learning process itself. As diverse as they may be, the methods and technologies described so far have one common advantage: They use the cognitive facilities and assess them at the same time. Moreover, they all use them in the way in which they are used in everyday situations. Even when used for assessment only, these methods do not create an artificial assessment situation which leads too far away from the usual reflection. We thus come back to the beginning at this point, where we stated that the investigation of knowledge is recursive – and that the recursion may very well be infinite in theory. In the fourth and conclusive part we turn the tables on this fact: If knowledge describes knowledge by means of models about knowledge, then the observation and interpretation process is not the necessary evil but the initial practical *reason* for the whole research field:

To use knowledge to support knowing to construct knowledge to promote knowing...

In the fourth part, the authors have selected best practice examples and widely applicable interpretation patterns from their research which applies some of the available methods. On one hand, the chapters present unique research which could just as well stand on its own. On the other hand, the authors demonstrate carefully how the available technologies for computer-based diagnostics and systematic analysis of knowledge may be applied. The examples are multifaceted in order to

provide a wide range of inspiration on how far the practical applications reach. Other researchers would and will of course come up with different ideas on how to apply the technologies. Maybe the different solutions will lead to even more designs in the future. We look forward to that development and hope that we will be able to support it with new methodologies, new approaches, and new applications.

Chapter 14
Computer-Based Feedback for Computer-Based Collaborative Problem Solving

Harold F. O'Neil, San-hui Sabrina Chuang, and Eva L. Baker

14.1 Introduction

Collaborative problem solving is considered a necessary skill for success in today's world and schooling. Collaborative learning refers to learning environments in which small groups of students work together to achieve a common goal, and problem solving is "cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver" (Mayer & Wittrock, 1996, p. 47). Thus, collaborative problem solving is defined as problem solving activities that involve interactions among a group of individuals. Figure 14.1 shows the components and their relationships to each other in collaborative problem solving.

As seen in Fig. 14.1, collaborative problem solving is first divided into two components: collaborative learning and problem solving. According to O'Neil, Chung, and Brown (1997), collaborative learning or teamwork can be further assessed by six collaborative skills: adaptability, coordination, decision making, interpersonal, leadership, and communication. This study used the teamwork processes model developed by CRESST researchers as measurement of collaborative learning processes. The CRESST model consists of six skills. They are "(a) adaptability – recognizing problems and responding appropriately, (b) coordination – organizing group activities to complete a task on time, (c) decision making – using available information to make decisions, (d) interpersonal – interacting cooperatively with other group members, (e) leadership – providing direction for the group, and (f) communication – clear and accurate exchange of information" (O'Neil et al., 1997, p. 413).

H.F. O'Neil (✉)
CRESST, University of Southern California, Los Angeles, CA, USA
e-mail: honeil@usc.edu

The work reported herein was supported under U.S. Department of Education Award Number R305C80015. The work was also supported by the Office of Naval Research under Award Number N000140810126. The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the U.S. Department of Education or the Office of Naval Research.

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_14,

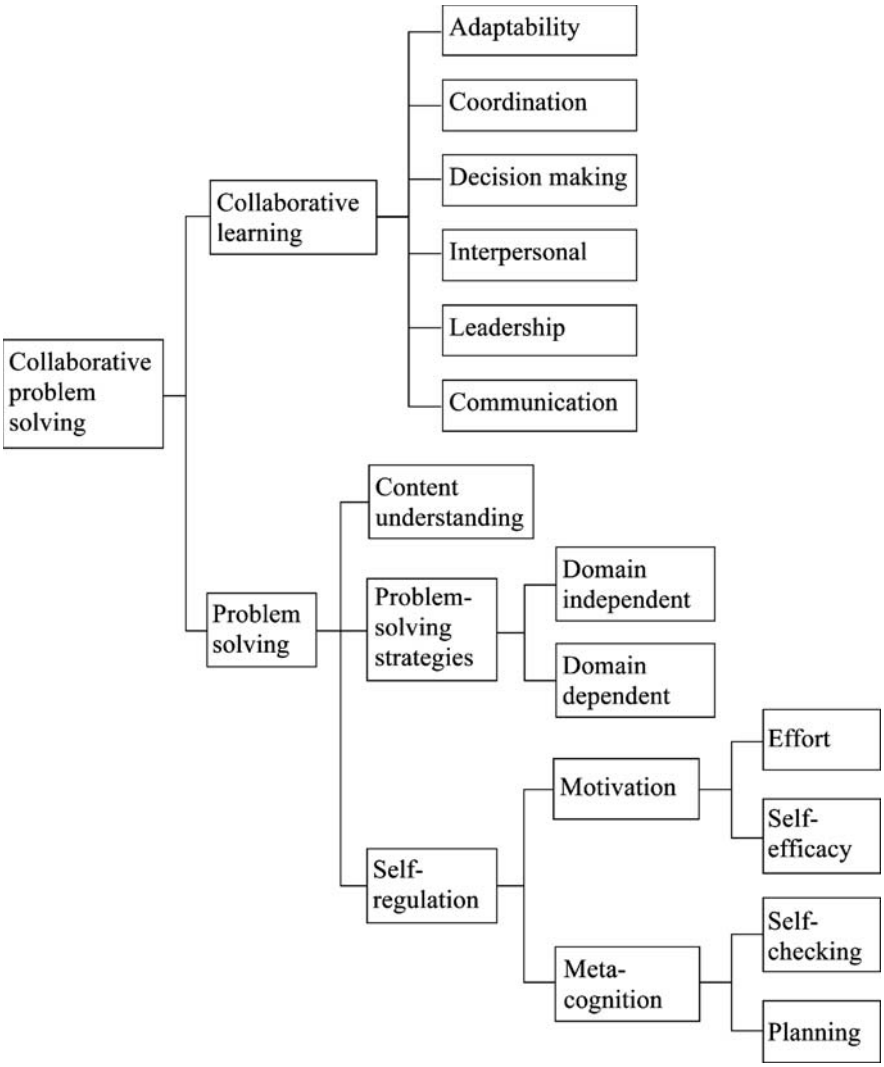


Fig. 14.1 Components of collaborative problem solving

According to O'Neil (1999), problem solving has three components: content understanding, problem solving strategies, and self-regulation. Content understanding is the domain knowledge required to solve a problem. Problem solving strategies can be domain-dependent or domain-independent. Self-regulation has two main components (motivation and metacognition) and each of them has two components. Motivation consists of effort and self-efficacy, and metacognition consists of self-checking and planning.

Several studies have shown the effectiveness of using computer technology to capture problem solving processes. For example, O'Neil, Wang, Chung, and Herl

(2000) and Hsieh and O'Neil (2002) used a computer-simulated teamwork task to evaluate problem solving and measure the collaboration processes involved. These processes were recorded by a computer and through predefined messages that participants used to communicate with their team members.

The feedback in Hsieh and O'Neil (2002) was divided into two categories: knowledge of response feedback and adapted knowledge of response feedback. Both types of feedback were based on comparing students' knowledge map performance to that of experts' map performance. Knowledge of response feedback provided students with information on how their map looked like when compared to experts' maps. Each concept on the map was categorized into three types of problems: needed (a) a little improvement, (b) some improvement, or (c) a lot of improvement. Hsieh and O'Neil's adapted knowledge of response feedback consisted of a knowledge of response feedback plus suggestions on improving the group map. For example, the feedback consisted of "You have improved 'food chain' from 'a lot of improvement' category to 'some improvement' category." This group performed the best and was significantly better than the control group.

However, the Hsieh and O'Neil (2002) study showed that one of the problem solving strategies, searching, was significantly negatively related to team performance. These findings were unexpected as most of the evidence indicates a positive effect by searching when the students are trained in searching techniques (Kuiper, Volman, & Terwel, 2005). By teaching searching and by providing different types of feedback, the Chuang and O'Neil (2006) study explored in further detail the effects of students' teamwork and problem solving processes on students' knowledge mapping performance. Chuang and O'Neil's (2006) feedback provided participants a direction about "what" area to improve for search and task performance, but also about "how" to improve the performance by using Boolean search operators. Thus, the effects of two types of feedback (adapted knowledge of response feedback and task-specific adapted knowledge of response feedback) with Boolean search strategies were also investigated. Their study showed that task-specific adaptive knowledge of responsive feedback solving was significantly more effective for improving problem solving. The collaboration between team members and individual students' problem solving process and strategies were positively related to which they attribute the training in searching and feedback on students' problem solving processes.

The purpose of this chapter was to further investigate the role of feedback in collaborative problem solving, in particular, the effect of narration plus on-screen text versus on-screen text only after-action review on team performance in collaborative problem solving tasks, i.e., a computer-based searching and knowledge mapping task. The "after-action review" (AAR) is a method for providing feedback to learners commonly used in military team training, e.g., following a simulated tactical exercise, the "action" (Morrison & Meliza, 1999). An excellent review of this military research is provided by Meliza and Goldberg (2008). The procedure is also labeled after-event review in the civilian sector (Ellis, Mendel, & Nir, 2006). In this chapter, we will use the term after-action review.

Because it is not possible to interrupt the training exercise to provide feedback on specific responses, the AAR is necessarily delayed. The AAR reviews what was

supposed to happen, identifies what actually happened during the execution, and stimulates team discussion on why it happened. During the discussion, team members learn from their mistakes and benefit from the lessons learned by other team members. The AAR becomes the bridge between the completed training event and the next training event, providing learning on "how to improve" that enables leaders to fix training weaknesses (Brown, Nordyke, Gerlock, Begley, & Meliza, 1998).

The military considers AAR an effective form of feedback, and although the literature on delay of feedback reviewed by Kulik and Kulik (1988) indicates that delay sometimes weakens the effectiveness of feedback, there are indications that when the learning task requires effortful cognitive processing, e.g., as in conceptual learning or problem solving, delay may be beneficial. For example, Lee and Zalatimo (1990) found that for students learning to solve analogy problems, those given delayed feedback scored higher on the posttest than the immediate feedback group. King, Young, and Behnke (2000) found evidence for an interaction of feedback type (immediate or delayed) with the nature of the learning task (rote or requiring effortful processing), immediate feedback being more effective with verbal learning tasks and delayed feedback more effective with concept learning or problem solving tasks. Similar findings were found by Hattie and Timperley (2007), who reported an effect size of 0.28 for the delayed group. King et al. (2000) found delayed feedback more effective with concept learning or problem solving tasks. Clariana, Wagner, and Murphy (2000) found among items with differential difficulty, retention of initial learning responses was greater for delayed feedback compared to immediate feedback across all items, but the result was more pronounced with difficult items (effect size of 1.17). The presumed mechanism for the effect is that a delay allows more time for metacognitive activities, identifying and filling knowledge gaps, and restructuring knowledge, activities that make the AAR feedback more effective.

Given the military's success with AAR in team training, and the suggestion that delayed feedback may be more effective with learning tasks requiring relatively effortful processing, we suggest that AAR-like delayed feedback may be effective in computer-based learning for the training of teams on a task with high cognitive load. There are different kinds of cognitive load (intrinsic, extraneous, and germane). The nature of the feedback (visual versus auditory) can become extraneous cognitive load (Paas, Renkl, & Sweller, 2003, 2004; van Merriënboer & Sweller, 2005). According to Mayer (2005) and Sweller, van Merriënboer, and Paas (1998), two important cognitive theoretical assumptions of multimedia learning were dual-channel assumption and limited capacity assumption. Dual-channel assumption means human beings process visual materials and audio materials in different channels (Mayer, 2005). Limited capacity assumption means during a given time each channel can only process a limited rather than unlimited amount of information (Kalyuga, Chandler, & Sweller, 2002; Ngu, Low, & Sweller, 2002; Vekiri, 2002).

Another assumption of cognitive load theory is that schemata are cognitive structures that allow information temporarily stored in working memory to be transferred into long-term memory and thus reduce working memory load (Sweller, Chandler, Tierney, & Cooper, 1990). Thus, it would be expected that when presenting more

information in the feedback than the working memory can handle, the information is lost rather than being used by the learners. This situation is called cognitive overload. Cognitive load theory and multimedia theory together have been widely used to explain feedback presentation research.

For example, Robinson and Molina (2002) designed two experiments in which students read a chapter-length text accompanied by either outlines or graphic organizers. Their results provided evidence that graphic organizers encoded in a visual format were more effective in assisting students than outlines which were encoded in a more auditory format. Their results were supportive of both cognitive load theory and multimedia theory.

14.1.1 Effects of Visual and Verbal Feedback

Mousavi, Low, and Sweller (1995) demonstrated that by mixing auditory and visual representation modes, cognitive load was reduced for mathematics learning. Their research also revealed that a visual−audio presentation mode promotes a deeper understanding of materials than a visual–visual presentation mode. Rieber (1996) experimented with the effects of animated graphical feedback and textual feedback on 41 undergraduates in a computer-based simulation program concerned with the laws of motion. They found that when given animated graphical feedback, subjects performed better, completed the game task in less time and were less frustrated.

Knowing that visual feedback is more successful than verbal feedback in general, Park and Gittelman (1992) found that within visual displays, animated visual display feedback was more effective than static visual display feedback for college students learning electronic troubleshooting skills. Similarly, Lalley (1998) demonstrated that video representation feedback, a visual–verbal presentation mode of feedback, led to better learning outcomes on the computerized biology multiple choice tests than textual feedback only.

Likewise, O'Neil, Mayer et al. (2000) examined training applications of a virtual environment simulation. They chose understanding an F-16 aircraft's fuel system as the learning task. They fixed most basic instructional design variables and only allowed the feedback representation to vary (e.g., narration versus on-screen text). The same information was in both versions. Only the mode of feedback delivery differed. They have shown that the narration group did better than the on-screen pop-up text group of the same information on three measures: the transfer test, the matching test, and the knowledge mapping test. These findings provide support for Mayers's modality principle (Clark & Mayer, 2003; Mayer, 2005, 2001).

In designing feedback, one must avoid cognitive load for both the task and the feedback if very intensive, and thus increased extraneous cognition load (Paas et al., 2003, 2004; Sweller et al., 1998; van Merriënboer & Sweller, 2005) and thus decreased performance. Therefore, given the mapping task and complexity of the feedback, it is likely that students would be cognitively challenged to process the total amount of information presented to them. To reduce the cognitive load in the

visual channel, the present study put some of the feedback into an audio channel so that the visual channel and audio channel were both engaged with neither channel overloaded with information at a given time. In other words, the study reported here investigated the effect of narration plus on-screen text AAR versus on-screen text only AAR on team performance in an online searching and mapping task. Our research question was:

Do teams which receive a narration plus on-screen text after-action review perform better on computer searching and mapping tasks than teams with an on-screen text after-action review only?

14.1.2 Methodology

The research design was an experimental one, as both groups and treatments were randomly assigned. The task was to improve a knowledge map about environmental science by searching a database for information on the topic. Each team consisted of a person whose role was to create the map, the other person searched the database. Students were randomly assigned into two-person groups as either searcher or mapper, and the groups were randomly assigned into one of two treatment groups. The difference in the two groups was the treatment group received after-action review feedback partly in narration and partly in on-screen text format while the control group received feedback in on-screen text format only.

14.1.3 Networked Knowledge Mapping System

Table 14.1 lists the specification for the networked knowledge mapping system that was used in the study.

The Java system used in this study was similar to Chuang and O'Neil's (2006) system which was based on the Schacter, Herl, Chung, Dennis and O'Neil (1999) study on individual problem solving and Chung, O'Neil, and Herl's (1999) study on collaborative assessment. Mappers in each group added concepts to the knowledge map via concept selection on the menu bar and linked concepts to other concepts via link selection on the menu bar respectively. In addition, mappers moved and erased concepts as well as links. Searchers in each group sought information from the simulated Web environment. The mappers could not search the simulated Web environment and the searcher could not construct the concept map. Thus each member in the group had to collaborate to successfully perform the task.

Participants received a paper handout that listed the 30 messages, grouped by common functions. The categories included (a) add concepts and links, (b) information from the web, (c) help and feedback seeking, (d) keeping track of progress, (e) messages about the group, and (f) quick responses. In addition to messages grouped by common functions listed on the handout, a complete list of concepts and links was provided on the handout as well.

Table 14.1 Domain specifications embedded in the software

General domain specification	This software
Scenario	Create a knowledge map on environmental science by exchanging messages in a collaborative environment and by searching for relevant information from a simulated World Wide Web environment.
Participants	Student team (two members).
Mapper	The one who does the knowledge mapping.
Searcher	The one who accesses the simulated World Wide Web environment to find relevant information and ask for feedback.
Knowledge map terms (Nodes)	Predefined – 18 important ideas identified by content experts: atmosphere, bacteria, carbon dioxide, climate, consumer, decomposition, evaporation, food chain, greenhouse gases, nutrients, oceans, oxygen, photosynthesis, producer, respiration, sunlight, waste, and water cycle.
Knowledge map terms (Links)	Predefined – 7 important relationships identified by content experts: causes, influences, part of, produces, requires, used for, and uses.
Simulated World Wide Web environment	Contains over 200 web pages with over 500 images and diagrams about environmental science and other topic areas.
Training	All students went through the same training section in video format. The training included the following elements: • how to construct the map (mapper) • how to search (searcher) • how to communicate with the other group member
Task feedback Three categories of AAR feedback	1. Feedback on the map performance by comparing group's knowledge map performance to that of expert's map performance. 2. Feedback on communication. 3. Feedback on search strategies.
Timing of feedback	Feedback was given to the experimental group at the end of the first 10-min session.
Experiment manipulation	For the experimental group, feedback on map performance was given in on-screen text format while communication and search feedback were given in both on-screen feedback and narration audio format. The control group received the same feedback except all three kinds of feedback were given in on-screen text format only.
Type of learning	Collaborative problem solving.
Problem solving measures	
Knowledge map	Content understanding and structure, semantic content score.
Information seeking	Browsing and searching.
Self-regulation	Planning, self-checking, self-efficacy, and effort.
Teamwork processes	Adaptability, coordination, decision making, interpersonal, leadership, communication.

14.1.4 Simulated World Wide Web Environment

The simulated World Wide Web ran on PC under Windows NT ™. The same World Wide Web environment used in Hsieh and O'Neil's (2002) study was used in this

study as well. The Web environment contained over 200 Web pages with over 500 images and diagrams about environmental science. Ninety percent of the information was downloaded from the Internet and 10% of the information was adapted from science textbooks and other science unit materials.

14.1.5 Feedback

AAR feedback was based on the previous study by Chuang and O'Neil (2006) who used task-specific adapted knowledge of response feedback. The task-specific adapted knowledge of response feedback included tips for Boolean search strategies.

Based on the previous studies (Hsieh & O'Neil, 2002; Chuang & O'Neil, 2006), the feedback of the after-action review feedback was divided into three categories. Category (1) was individualized according to each team's map performance. Category (2) and (3) used standard feedback for both groups.

Category (1). Feedback on student map performance was provided by comparing students' knowledge map performance to three experts' maps. The feedback showed students four things. First, it showed students which propositions matched at least two experts' propositions. Second, it showed students which propositions matched a single expert's. Third, it showed students which propositions did not match any of the experts. Last, it showed students concepts that were not properly included or had not been included at all in the map.

Category (2). The second category of feedback was searching feedback. The computer recorded how many searches had been conducted and displayed that number as partial feedback. This number was potentially unique for each group. However, information on how to conduct Boolean searches using the "and" operation was the same and independent of the number of searches.

Category (3). The third category of feedback was communication feedback. The computer calculated how many total messages were sent between the team members and displayed it as partial feedback. This number was potentially different for each group. However, the general communication tips were the same and independent of the number of messages.

14.1.6 Participants

Participants were 188 college students in Southern California. All of the participants were 18 years or older and possessed basic computer skills including the ability to use a laptop computer with a touch pad mouse and computer keyboard. The study started with 188 participants participating in 17 sessions. Students were paid $20 each for participating in the study. The 188 participants were randomly paired with a partner to form 94 teams, and the teams were randomly assigned to either the treatment group or control group. Teams in the treatment group received after-action

review feedback partly in audio and partly in text format on their performance. Teams in the control group received feedback in text format only.

The final data analyses were based on 160 participants. Fourteen sets of data were lost to computer technology failures or participants who were team members that were no-shows.

14.2 Measures

The measures in this study were adapted from the Chung et al. (1999), Hsieh and O'Neil (2002), Chuang and O'Neil (2006), and O'Neil et al. (1997) studies.

14.2.1 Group Outcome Measures

Group outcome measures were computed by comparing the semantic content score of a group's knowledge map to the semantic score of a set of three experts (Chuang & O'Neil, 2006). The following descriptions showed how these outcomes were scored. First, the semantic score was based on the semantic propositions in experts' knowledge map and was calculated by categorized map scoring (Herl, Baker, & Niemi, 1996). Using this method, all seven links were categorized into four classifications. First, "causes" and "influences" were classified as the "casual" category and were further marked as string "1." Second, links such as "requires," "used for," and "uses" were put in the "conditional" category and were marked as string "2." Third, "part of" and "produces," the remaining two links, were classified on their own and were marked as string "3" and "4" respectively. Every proposition in a student map was compared against each proposition in the three experts' maps. One match was scored as one point. The average score across all three experts was the semantic score of the map. For example, if a student group made a proposition such as "Photosynthesis produces oxygen," this proposition was first categorized into "Photosynthesis 4 oxygen" and then was compared with three experts' propositions. A score of one meant this proposition was the same with the proposition in the map of an expert. A score of zero meant this proposition was not the same as an expert's proposition. These scores were averaged across the three experts.

14.2.2 Information Seeking and Feedback Behavior Measures

Information seeking and feedback behavior measures consisted of two measures: (a) browsing and (b) searching. Browsing was measured by how many times the searchers selected the Web pages from the hypertext directory or glossary, or clicked on any hypertext within the Web pages. Each time the searchers selected the Web pages from the hypertext directory or glossary, or clicked on any hypertext within

the Web pages, a point was awarded to browsing score. In this manner, a browsing score was calculated.

For searching, one point was awarded for simple searching. For example, when a student typed "oxygen" as the search string, one point is awarded to the group. A group was awarded an additional point if the search involved Boolean search strategies. We considered Boolean searches a more complex task, thus an additional point was given for this activity. For example, when a student typed "oxygen and sunlight," in addition to the one point for simple searching, an additional point was awarded for using the Boolean operator "and."

14.2.3 Teamwork Process Measures

This study used the CRESST teamwork model developed by CRESST researchers as a measure of collaborative learning processes. The CRESST teamwork model consists of six processes. They are "(a) adaptability – recognizing problems and responding appropriately, (b) coordination – organizing group activities to complete a task on time, (c) decision making – using available information to make decisions, (d) interpersonal – interacting cooperatively with other group members, (e) leadership – providing direction for the group, and (f) communication – clear and accurate exchange of information" (O'Neil et al., 1997, p. 413).

Teamwork measures were used to evaluate the engagement of the group in each of the team processes (i.e., adaptability, coordination, decision making, interpersonal, leadership, and communication). Teamwork process measures were counted by adding the number of messages both members in a group sent from each teamwork process category. If the mapper sent 5 messages from the adaptability category and the searcher in a group sent 7 messages from the adaptability category, the adaptability score was 12. In addition to each team process measures, a count of all messages sent was used as the overall teamwork score. This overall teamwork score was termed "communication score" to the participants in the study.

14.3 Procedure

Data were collected in a large room with 12 laptops set up in pairs. There were altogether 80 sets of teams. Each participant was randomly assigned to a group and to a role (mapper or searcher). Because the present study intended to evaluate the effects of different after-action review feedback formats on map performance, all groups were randomly assigned either to receive narration/on-screen text AAR feedback or on-screen text feedback only.

The entire session required approximately 65 min, which included: (a) 10 min for the teamwork questionnaire; (b) 10 min for task instructions and training; (c) 10 min for the collaborative group task; (d) 5 min of AAR and a 5-min break for

treatment group; (e) 10-min break for the control group; (f) 10 min for the second collaborative group task; and (g) 5 min for debriefing.

14.3.1 Teamwork Questionnaire

All participants completed the 35-item Teamwork Questionnaire (Marshall et al., 2005) on the computer. The purpose of the questionnaire was to gather information about participants' knowledge and skills in teamwork. These data will be reported elsewhere. Since the teamwork questionnaire was administered before the treatment variations began, there was no expectation that there would be an impact of treatment on teamwork.

14.3.2 Task Instructions and Search Strategies Training

The searching training for the treatment group and the training for the control group were the same. Everyone watched a training video on a computer. The video showed both the searcher and the mapper how to construct a map, how to conduct a search, and how to communicate to his/her member via communication box with predefined messages. The video also explained the responsibilities for each role to mappers and searchers.

14.3.3 Collaborative Group Task 1

After the training, each pair was logged onto the computer network one at a time and then began working. The computer started recording the beginning of the team's first 10-min session when the team made its first map. At the end of the first 10-min session, the computer automatically saved the current map as the first map and a score was calculated, then the treatment group received after-action review feedback partly in narration and partly in on-screen text format while the control group received feedback in on-screen text format only. Feedback was given on improving the map, search strategies, and group teamwork skills. The feedback for improving the map and search strategies was important for improving the map and thus mainly cognitive in nature, whereas the feedback on teamwork processes was mainly affective in nature.

14.3.4 After-Action Review Feedback

The treatment group received AAR feedback at the end of the first session, which included a summary of feedback on the map in on-screen text format, and feedback on searching and communication in narration format. The control group received

exactly the same kinds of feedback in on-screen text format only. Figure 14.2 shows examples of feedback on the map, searching, and communication, respectively. The time for the AAR feedback was 5-min for both groups. The feedback on the map construction was individualized according to each individual map while the search and communication feedback were standard. The narration (audio) feedback on communication and on the search strategies was pre-recorded by the programmer and the participants listened to the feedback through a computer headphone. The audio feedback started automatically when the user finished reading the text feedback and clicked the audio feedback link. The time for reading the on-screen text feedback on the map construction was 3-min and the time for reading/listening to the feedback on search and communication was 2-min maximum.

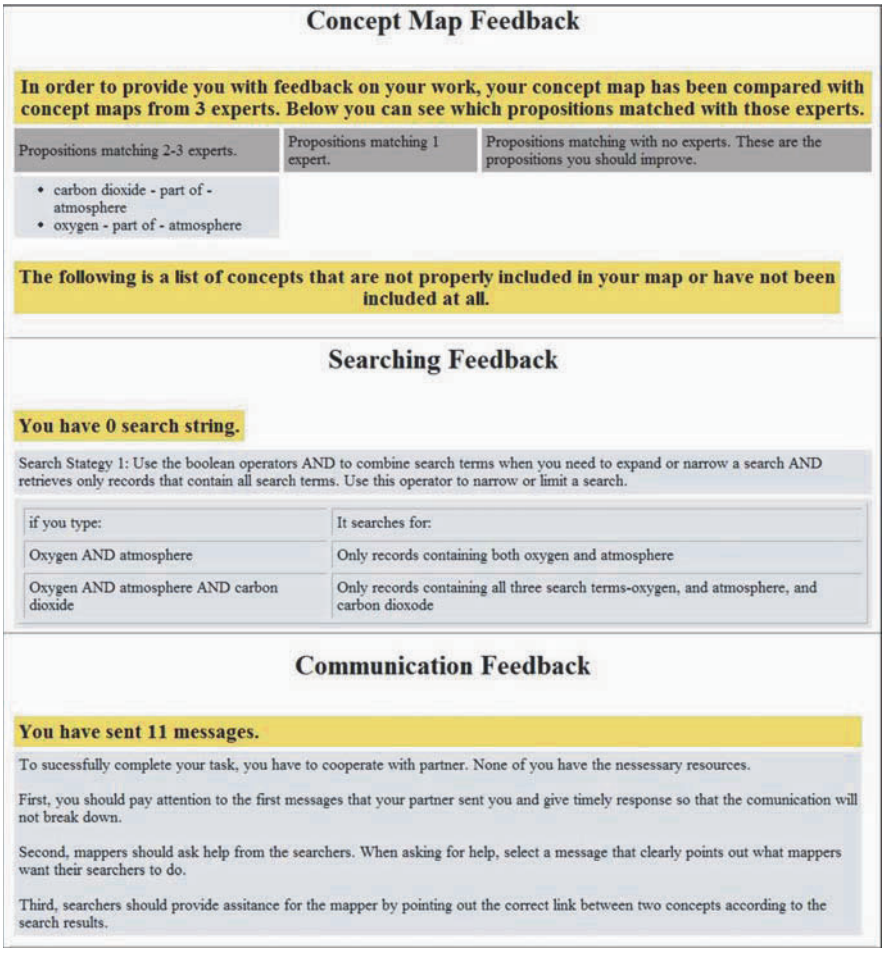


Fig. 14.2 An example of AAR feedback

14.3.5 Collaborative Group Task 2

During the break, the participants were allowed to go outside and use the bathroom, and most of them chose to talk on their cell phones. After the break, the second 10-min session began. The researcher retrieved the saved map and the team started working again. At the end of this second 10-min session to improve their maps, the computer automatically saved the second map as the final map.

14.3.6 Debriefing

The study ended after the second 10-min session. The participants were given a 5-min debriefing session for any questions they had.

14.4 Data Analysis

Each team received two sets of scores for each task: a map score (content understanding), search scores (total searches, total Boolean counts), and total teamwork scores or "communication" score. The content of the team map was compared to that of a set of three expert maps and a score was assigned to each of the maps created by the participants. The total search scores were the total number of searches with Boolean counts. Boolean count was a count of Boolean searches. The communication score was the sum of all messages sent between the mapper and the searcher. Independent sample *t*-tests and analysis of covariance were conducted to examine the difference between treatment group and control group teams during the two sessions, and to determine if the different AAR treatment had a significant effect on map scores, search scores, and communication scores. Statistical significance was defined as probability less than 0.05 ($p < 0.05$). So, for example, for *t*-tests, probability was less than 0.05 and was two-tailed comparisons.

14.5 Results and Discussion

14.5.1 The Effect of AAR on Team Map Scores

The pre- and post-map scores for both the treatment and the control groups were collected, and descriptive analysis was conducted (see Table 14.2). For the pretest, the mean group outcome for the control group was 10.97 while the mean group outcome for the experiment group was 12.77. A *t*-test indicated the mean differences were not significant. For the posttest, the mean group outcome for the knowledge of response feedback control group was 22.13 while the mean group outcome for the adapted knowledge of response feedback group was 29.08. The mean differences of

Table 14.2 Pre- and posttest map scores for narrative plus on-screen text treatment and on-screen text control group

Map score	Treatment group ($n = 34$) *M* (*SD*)	Control group ($n = 34$) *M* (*SD*)
Pretest	12.77 (9.12)	10.97 (9.03)
Posttest	29.08 (12.46)	22.13 (11.32)

the group outcome for two feedback treatment in this study were statistically significant. The effect size was medium (d = 0.58). Here we followed Cohen's (1988) definition on the effect size. Cohen categorized three effect sizes: small, medium, and large depending on the d-value. He defines "small effect size: d = 0.2" (p. 25), "medium effect size: d = 0.5" and "large effect size: d = 0.8" (p. 26). Although the pretest scores were not significantly different, an ANCOVA was also computed using the pretest as a covariate and indicated that the treatment effect was significant. Therefore, the after-action review had significant effect on improving the teams' content understanding.

14.5.2 The Effect of AAR on Search Scores

Participants' search skills were measured by (a) total number of searches conducted, and (b) number of Boolean searches (Boolean count) conducted. Table 14.3 lists the descriptive analysis results for both the treatment group (on-screen text plus narration) and control group (on-screen text only) teams for the search scores.

T-tests indicate that there was no statistically significant difference between the two groups on their search scores for the pretest, but there was significant difference in the posttest. The effect size for posttest was small (d = 0.47).

Table 14.3 Descriptive results of search scores for narrative plus on-screen text treatment and on-screen text control group

Search score	Treatment group ($n = 34$) *M* (*SD*)	Control group ($n = 34$) *M* (*SD*)
Pretest	28.02 (16.52)	27.64 (15.57)
Posttest	62.06 (21.25)	51.97 (28.19)

Table 14.4 lists the descriptive analysis results for both the treatment group and control group teams on their Boolean score. A *t*-test indicates that there was no statistically significant difference between the two groups on their Boolean scores for the pretest but there was significant difference in the posttest. The effect size was medium (d = 0.51).

Table 14.4 Descriptive results of Boolean scores for treatment and control groups

Boolean score	Treatment group ($n = 34$) *M (SD)*	Control group ($n = 34$) *M (SD)*
Pretest	7.91 (0.84)	7.00 (1.01)
Posttest	10.77 (7.36)	7.90 (5.62)

14.5.3 The Effect of AAR on Communication Scores

With the sum of total message frequency counts calculated for all six teamwork processes, we then calculated the group level (two-person team) teamwork process measures. The frequency counts were calculated by adding the number of usage for the individual messages left in each teamwork process and then divided by the number of teams. For example, for adaptability process, the number of usage for messages added together (messages 1, 2, 25, 26, and 27) were 650. When we divide by the number of teams ($N = 78$), the mean was calculated to be 8.33. In the same manner, we calculated the mean for adaptability process for adapted knowledge of response feedback group to be 7.30 ($SD = 3.01$), and for task-specific adapted knowledge of response feedback group to be 9.36 ($SD = 5.18$).

Table 14.5 presents the descriptive statistics of teamwork process. As may be seen in Table 14.5, adaptability messages for the on-screen Text Feedback Only group ($n = 39$) had a mean of 7.31 with $SD = 3.01$. For this group ($n = 39$), leadership messages (e.g., "Let's work on [C]") were most often used whereas decision-making messages (e.g., "Feedback shows we should work more on [C]") were the least used. Independent sample *t*-tests were used to denote statistical significance. We used a Bonferroni correction that reset the critical *p*-value from $p < 0.05$ to $p < 0.008$. Using this conservative criterion, decision making was significantly higher in the text plus audio feedback group ($T = 8.20$, $p < 0.0001$). Unexpectedly, leadership was significantly higher in the on-screen text control group ($T = 4.38$, $p < 0.0001$). No other comparisons were significantly different.

Table 14.5 Descriptive statistics of message counts for teamwork processes (on-screen text only group) ($n = 39$)

Teamwork process	Mean	*SD*
Adaptability	7.31	3.01
Coordination	6.51	3.22
Decision making	4.69	1.54
Interpersonal	8.71	3.46
Leadership	9.30	3.89
Communications	52.36	28.28

Table 14.6 presents the descriptive statistics of teamwork process for on-screen text plus audio feedback group. As may be seen in Table 14.6, adaptability messages for the text plus audio feedback group ($n = 39$) had a mean of 9.36 with $SD = 5.18$.

Table 14.6 Descriptive statistics of message counts for teamwork processes (on-screen text plus narration feedback group) ($n = 39$)

Teamwork process	Mean	*SD*
Adaptability	9.36	5.18
Coordination	7.30	3.30
Decision making	10.84	4.42
Interpersonal	7.31	2.04
Leadership	6.03	2.56
Communications	41.87	9.73

For this group ($n = 39$) in contrast to the on-screen text only group, leadership messages (e.g., "Let's work on [C]") were least often used whereas decision-making messages (e.g., "Feedback shows we should work more on [C]") were the most used.

14.6 Summary and Conclusions

In summary, the results of this study indicated that the on-screen text plus narration feedback had a significant effect on the content understanding as measured by their increased map scores compared to the on-screen text only group. Moreover, the on-screen text plus narrative feedback group did significantly increase the number of searches and number of Boolean operators than the on-screen text feedback only group. These results were expected since the narrative plus on-screen text feedback was on increasing the team's searches and Boolean searches. In general, the on-screen text plus narration AAR groups had minimal effects on their teamwork scores except for the decision-making teamwork scores.

Even though both groups were presented with the same information on search and communication, the different formats of presentation seem to be the reason why there was a difference in the group outcome. As Mayer (2001) pointed out, cognitive processing differs cognitively for spoken words versus printed words. In our task, the feedback requires a lot of cognitive capacity to process, to retain, and to transfer. For example, the feedback contains all the information on the concept maps with 18 concepts and the links they formed with each other plus feedback on communication and search. Students only have 5-min to process all the feedback. It is highly likely that there is a cognitive overload happening for the students in the on-screen text only version. In effect one overloads the visual channels with presentation of on-screen text feedback. Therefore, putting some feedback in an audio channel via narration, a different channel from the visual channel, might decrease some of the cognitive load and thus help the retention of the feedback and the transfer of the feedback into actual practice. These results provide support for the modality principle (Clark & Mayer, 2003; Mayer, 2005).

14.7 Summary of Chapter

Collaborative problem solving was defined as problem solving activities that involve interactions among a group of individuals. Collaborative problem solving is considered a necessary skill for success in today's world and schooling. The purpose of this chapter was to further investigate the role of computer-based feedback in collaborative problem solving, in particular, the effect of a narration plus on-screen text versus an on-screen text only version after an after-action review feedback on team performance in collaborative problem solving task, i.e., a computer-based searching and knowledge mapping task. The "after-action review" (AAR) is a method for providing delayed feedback to learners commonly used in military team training, e.g., following a simulated tactical exercise, the "action." Common in military training following the AAR, a different type of task is performed. For example, if the task before the AAR was an offensive scenario, then the task following the AAR might be a defensive scenario. Our study was the first to our knowledge to focus on the effects of an AAR on a subsequent task of the same nature. The research literature in this topic was also reviewed and the results of a study supporting the AAR intervention were discussed in terms of cognitive load theory.

Acknowledgments The authors would like to thank Ali Abedi for his programming assistance.

References

Brown, B. R., Nordyke, J. W., Gerlock, D. L., Begley, I. J., & Meliza, L. L. (1998, May). *Training analysis and feedback aids (TAAF Aids) study for live training support* (Study Report, Army Project Number 2O665803D730). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Chuang, S. H., & O'Neil, H. F. (2006). Role of task-specific adapted feedback on a computer-based collaborative problem-solving task. In H. F. O'Neil & R. Perez (Eds.), *Web-based learning: Theory, research, and practice* (pp. 239–254). Mahwah, NJ: Lawrence Erlbaum.

Chung, G. K. W. K., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior*, *15*, 463–493.

Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research and Development*, *48*(3), 5–22.

Clark, R. C., & Mayer, R. E. (2003). *E-learning and the science of instruction*. San Francisco: Pfeiffer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Ellis, S., Mendel, R., & Nir, M. (2006). Learning from successful and failed experience: The moderating role of kind of after-event review. *Journal of Applied Psychology*, *91*, 669–680.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of U.S. history knowledge. *Journal of Educational Research*, *89*(4), 206–219.

Hsieh, I. G., & O'Neil, H. F., Jr. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior*, *18*(6), 699–715.

Kalyuga, S., Chandler, P., & Sweller, J. (2002). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology*, *92*(1), 126–136.

King, P. E., Young, M. J., & Behnke, R. R. (2000). Public speaking performance improvement as a function of information processing in immediate and delayed feedback interventions. *Communication Education*, *49*(4), 365–374.

Kuiper, E., Volman, M., & Terwel, J. (2005). The web as an information resource in K-12 education: Strategies for supporting students in searching and processing information. *Review of Educational Research*, *75*(3), 285–328.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79–97.

Lalley, J. P. (1998). Comparison of test and video as forms of feedback during computer assisted learning. *Journal of Educational Computing Research*, *18*(4), 323–338.

Lee, W., & Zalatimo, S. (1990). Computer-assisted instruction with immediate feedback versus delayed feedback in learning to solve analogy items. *International Journal of Instructional Media*, *17*(4), 319–329.

Marshall, L., O'Neil, H. F., Chen, A., Juehl, M., Hsieh, I., & Abedi, J. (2005). Teamwork skills assessment and instruction. In J. M. Spector, C. Ohrazda, A. Van Schaak, & D. A. Wiley (Eds.), *Innovations in instructional technology: Essays in honor of M. David Merrill* (pp. 131–149). Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge, NY: Cambridge University Press.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: Simon & Schuster Macmillan.

Meliza, L. L., & Goldberg, S. L. (2008). Impact of after-action review on learning in simulation-based U.S. Army training. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 255–272). New York: Lawrence Erlbaum Associates.

Morrison, J. E., & Meliza, L. L. (1999). *Foundations of the after-action review process* (Special Report 42). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, *87*(2), 319–334.

Ngu, B., Low, R., & Sweller, J. (2002). Text editing in chemistry instruction. *Instructional Science*, *30*, 379–402.

O'Neil, H. F., Jr. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior*, *15*(3–4), 225–268.

O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411–452). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., Mayer, R. E., Herl, H. E., Niemi, C., Olin, K., & Thurman, R. A. (2000). Instructional strategies for virtual aviation training environments. In H. F. O'Neil Jr. & D. H. Andrews (Eds.), *Aircrew training and assessment* (pp. 105–130). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., Wang, S., Chung, G. K. W. K., & Herl, H. E. (2000). Assessment of teamwork skills using computer-based teamwork simulations. In H. F. O'Neil Jr. & D. H. Andrews (Eds.), *Aircrew training and assessment* (pp. 245–276). Mahwah, NJ: Lawrence Erlbaum Associates.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Science*, *32*(1–2), 1–8.

Park, O., & Gittelman, S. S. (1992). Selective use of animation and feedback in computer-based instruction. *Educational Technology Research and Development*, *40*(4), 27–37.

Rieber, L. P. (1996). Animation as feedback in a computer-based simulation representation matters. *Educational Technology Research and Development*, *44*(1), 5–22.

Robinson, D. H., & Molina, E. (2002). The relative involvement of visual and auditory working memory when studying adjunct displays. *Contemporary Educational Psychology*, *27*, 118–131.

Schacter, J., Herl, H. E., Chung, G., Dennis, R. A., & O'Neil, H. F., Jr. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem solving. *Computers in Human Behavior*, *15*, 403–418.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load and selective attention as factors in the structuring of technical material. *Journal of Experimental Psychology: General*, *119*, 176–192.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177.

Vekiri, L. (2002). What is the value of graphical displays in learning? *Educational Psychology Review*, *14*(3), 261–312.

Chapter 15
Modeling, Assessing, and Supporting Key Competencies Within Game Environments

Valerie J. Shute, Iskandaria Masduki, Oktay Donmez, Vanessa P. Dennen, Yoon-Jeon Kim, Allan C. Jeong, and Chen-Yen Wang

15.1 Introduction

> Human beings, viewed as behaving systems, are quite simple. The apparent complexity of our behavior is largely a reflection of the complexity of the environment in which we find ourselves. (Herbert A. Simon, 1996, p. 53)

A critical challenge for any successful instructional-learning system involves accurately identifying characteristics of a particular learner or group of learners – such as the type and level of specific knowledge, skills, and other attributes. This information can then be used to improve subsequent learning (Conati, 2002; Park & Lee; 2003; Shute, Lajoie, & Gluck, 2000; Snow, 1994). But *what* are the most valuable competencies needed to succeed in the twenty-first century, and *how* can we assess them accurately and support their development? These questions comprise the crux of our research, with a focus on the "how" part of the story in this chapter.

To put our research issues in context, the demands associated with living in a highly technological and globally competitive world require today's students to develop a very different set of skills than their parents (and grandparents) needed. That is, when society changes, the skills that citizens need to negotiate the complexities of life also change. In the past, a person who had acquired basic reading, writing, and calculating skills was considered to be sufficiently literate. Now, people are expected to read critically, write persuasively, think and reason logically, and solve increasingly complex problems in math, science, and everyday life. The general goal of education is to prepare young people to live independent and productive lives. Unfortunately, our current educational system is not keeping pace with these changes and demands of today's more complex environment.

V.J. Shute (✉)
Florida State University, Tallahassee, FL, USA
e-mail: shute@mail.coe.fsu.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_15,

15.1.1 Purpose

This chapter will describe our ideas and tools for modeling, assessing, and supporting key competencies (e.g., systems thinking, creativity, and collaboration) via formative assessment embedded within immersive games. Through an extensive literature review described elsewhere (Shute, Dennen, Kim, Donmez, & Wang, 2008), we have identified and have begun modeling a set of educationally valuable attributes, or *competencies*, that are currently being ignored in our schools (locally and globally), but we believe should not be – especially with an eye toward the near future. Our modeling efforts extend an existing evidence-centered design (ECD) approach formulated by Mislevy, Steinberg, and Almond (2003) and employ Bayesian networks (Pearl, 1988). That is, inferences – both diagnostic and predictive – are handled by Bayes nets and used directly in the student models to handle uncertainty via probabilistic inference to update and improve belief values on learner competencies. To make these ideas more concrete, we present an analysis (or worked example) of an existing 3D immersive game called Quest Atlantis: *Taiga Park* (e.g., Barab, 2006; Barab, Zuiker et al., 2007; Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007), and demonstrate how evidence is gathered and interpreted in relation to one of our targeted competencies: systems thinking skill.

The longer term goal of our research, outside the scope of this chapter, is to fully develop, refine, pilot test, and ultimately validate our evidence-based approach using stealth assessment embedded within immersive learning environments (e.g., games, simulations, scenarios) that can elicit data from learners, make inferences about competency levels at various grain sizes, and use that information as the basis for targeted and immediate support. The motivation for this research is the belief that certain attributes of people, such as insulating against opposing views, reducing complex issues to black-and-white terms, and failing to question entrenched ideas will likely *not* move us – citizens of the world – in the direction necessary to flourish in the twenty-first century. Our research goals are toward ensuring that current and future *worldizens* can learn to systematically and creatively think, communicate, question, collaborate, solve difficult problems, reflect on decisions and solutions to problems, and adapt to rapidly changing circumstances.

There are many obstacles that need to be overcome before education is truly effective for the future and for the masses (e.g., shortage of well-qualified teachers, inadequate financial resources for poor schools, delivery of content in ways that do not engage students, reliance on tests to get numbers instead of insight). One obstacle that is not usually included in the various lists – but should be – concerns a lack of clear vision about what exactly we are preparing our kids for. We can readily identify trends, such as the *shrinking world* phenomenon that occurs as we become progressively more interconnected. And we know that in the long run, it is less important to memorize information than to know how to locate and make sense of credible information. But do our schools alter their curricula to accommodate these emergent needs? No. Are we adequately preparing our students for the realities of their future? No. Students are still pushed to memorize and repeat facts,

and consequently they are graduating high school ill-prepared to tackle real-world, complex problems. We cannot directly adjust the wind (the future), but we *can* adjust the sails (competencies). To do so effectively, we need to have a good sense of bearings – where we are, and where we are heading.

15.1.2 Where We Are

This section briefly overviews two major problems confronting us today: (a) disengaged students, and (b) an effectively shrinking world, commensurate with increased communication technologies (e.g., Barab, Zuiker et al., 2007; Gee, 2004a, 2004b; Shute, 2007). It provides the basic rationale for our moving toward authentic, engaging learning activities and related stealth assessment to support learning.

15.1.2.1 Disengaged Students

There is a huge gulf between what kids do for fun and what they are required to do in school. School covers material that we deem important, but kids, generally speaking, are unimpressed. These same kids, however, are highly motivated by what they do for fun (e.g., play interactive games). This mismatch between mandated school activities and what kids choose to do on their own is cause for concern regarding the motivational impact (or lack thereof) of school, but it need not be the case. Imagine these two worlds united. Student engagement is strongly associated with academic achievement; thus, combining school material with games has tremendous potential to increase learning, especially for lower performing, disengaged students. The logic underlying the research is as follows. Compelling storylines (narratives) represent an important feature of well-designed games. Well-designed games tend to induce *flow* (Csikszentmihalyi, 1990), a state in which a game player loses track of time and is absorbed in the experience of game play. Flow is conducive to engagement, and engagement is conducive to learning. The problem is that immersive games lack an assessment infrastructure to maximize learning potential. Furthermore, typical assessments are likely to disrupt flow in good games. Thus, there is a need for embedded (i.e., *stealth*) assessments that would be less obtrusive and hence less disruptive to flow.

15.1.2.2 The Shrinking World

The second problem motivating our research is that the world is effectively shrinking. We are currently confronted with problems of enormous complexity and global ramifications (e.g., the massive meltdown on Wall Street, nuclear proliferation, global warming, a plastic island the size of Texas in the Pacific, antibiotic resistant microbes, destruction of the rain forests, and poverty). The people who will be making and managing policy decisions in the near future need to be able to understand, at the very least, how research works and how science works because solutions are going to be highly technical and highly complex. When confronted by problems,

especially new issues for which solutions must be created out of whole cloth, the ability to think creatively, critically, collaboratively, systemically, and then communicate effectively is essential. Learning and succeeding in a complex and dynamic world is not easily measured by multiple-choice responses on a simple knowledge test. Instead, solutions begin with re-thinking assessment, identifying new skills and standards relevant for the twenty-first century, and then figuring out how we can best assess students' acquisition of the new competencies – which may in fact involve the teacher, the computer, the student, one's peers, and so on. Moreover, the envisioned new competencies should include not only cognitive variables (e.g., critical thinking and reasoning skills) but also noncognitive variables (e.g., teamwork, tolerance, and tenacity) as the basis for new assessments to support learning (Abedi & O'Neil, 2005; Farkas, 2003).

15.1.3 Where We Should Be Heading

The primary goal of this chapter is to figure out *how* to accomplish the design and development of valid and reliable assessments for critical competencies. As a preliminary step, we have begun to identify key competencies (see Fig. 15.1). This is not a comprehensive list; additional competencies will be identified and modeled as our research evolves. In this chapter we will model systems thinking skill to demonstrate how evidence-based assessments might be developed and embedded within games and simulation environments. Modeling, assessing, and supporting students in relation to our set of skills is intended to allow students to grow in a number of important new areas, function productively within multidisciplinary teams, identify and solve problems (with innovative solutions), and communicate effectively.

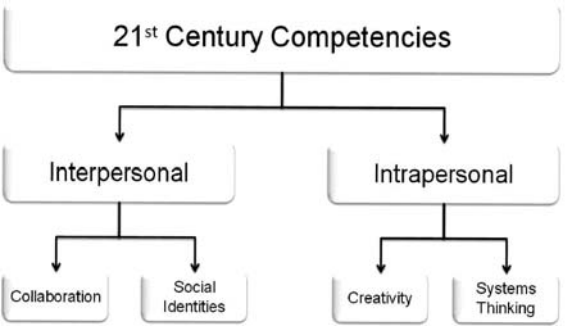


Fig. 15.1 Current set of key competencies for the twenty-first century

To accomplish our goal of developing really good assessments that can also support learning, we turn now to the "how" part of the story; namely, an overview of evidence-centered design (ECD) which supports the design of valid assessments. ECD entails developing competency models and associated assessments. We extend ECD by embedding these evidence-based assessments within interactive environments – comprising stealth assessment. Afterward, we present (a) a literature review

and comprehensive model associated with the systems thinking competency and (b) a description of how these ideas would actually play out within an existing immersive game – Quest Atlantis: Taiga Park.

15.2 Assessment Methodology: Evidence-Centered Design

> The nature of the construct being assessed should guide the selection or construction of relevant tasks, as well as the rational development of construct-based scoring criteria and rubrics. (Sam Messick, 1994, p. 17)

The fundamental ideas underlying ECD came from Messick (1994; see quote above). This process begins by identifying what should be assessed in terms of knowledge, skills, or other attributes. These variables cannot be observed directly, so behaviors and performances that demonstrate these variables should be identified instead. The next step is determining the types of tasks or situations that would draw out such behaviors or performances. An overview of the ECD approach is described below (for more on the topic, see Mislevy & Haertel, 2006; Mislevy, Almond, & Lukas, 2004; Mislevy et al., 2003).

15.2.1 ECD Models

The primary purpose of an assessment is to collect information that will enable the assessor to make inferences about students' competency states – what they know, believe, and can do, and to what degree. Accurate inferences of competency states support instructional decisions that can promote learning. ECD defines a framework that consists of three theoretical models that work in concert. The ECD framework allows/requires an assessor to: (a) define the claims to be made about students' competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of tasks that will elicit that evidence. These three actions map directly onto the three main models of ECD shown in Fig. 15.2.

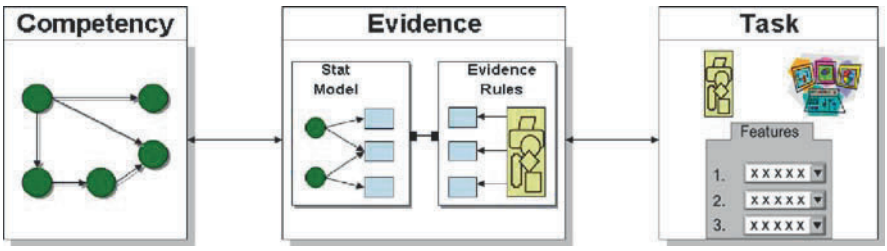


Fig. 15.2 Three main models of an evidence-centered assessment design

A good assessment has to elicit behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence in terms

that suit the purpose of the assessment. Working out these variables, models, and their interrelationships is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design.

15.2.1.1 Competency Model

What collection of knowledge, skills, and other attributes should be assessed? This can also be phrased as: What do you want to say about the person at the end of the assessment? Variables in the competency model (*CM*) are usually called "nodes" and describe the set of person variables on which inferences are based. The term "student model" is used to denote a student-instantiated version of the CM – like a profile or report card, only at a more refined grain size. Values in the student model express the assessor's current belief about a student's level on each variable within the CM. For example, suppose the CM for a science class that valued the general competency of systems thinking contained a node for "Create a causal loop diagram." The value of that node – for a student who was really facile at understanding and drawing causal loop diagrams – may be "high" (if the competency levels were divided into low, medium, and high), based on evidence accumulated across multiple, relevant tasks.

15.2.1.2 Evidence Model

What behaviors or performances should reveal differential levels of the targeted competencies? An evidence model expresses how the student's interactions with, and responses to a given problem constitute evidence about competency model variables. The evidence model (EM) attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies; and (b) What's the connection between those behaviors and the CM variable(s)? Basically, an evidence model lays out the argument about why and how observations in a given task situation (i.e., student performance data) constitute evidence about CM variables. Using the same node as illustrated in the CM section above, the evidence model would clearly indicate the aspects of causal loop diagrams that must be present (or absent) to indicate varying degrees of understanding or mastery of that competency. The same logic/methods apply to noncognitive variables as well – stating clearly the rubrics for scoring aspects of creativity, teamwork, etc.

15.2.1.3 Task Model

What tasks should elicit those behaviors that comprise the evidence? A task model (TM) provides a framework for characterizing and constructing situations with which a student will interact to provide evidence about targeted aspects of knowledge or skill related to competencies. These situations are described in terms of: (a) the presentation format (e.g., directions, stimuli), (b) the specific work or response products (e.g., answers, work samples), and (c) other variables used to describe key features of tasks (e.g., knowledge type, difficulty level). Thus, task specifications

establish what the student will be asked to do, what kinds of responses are permitted, what types of formats are available, and other considerations, such as whether the student will be timed, allowed to use tools (e.g., calculators, dictionaries), and so forth. Multiple task models can be employed in a given assessment. Tasks are the most obvious part of an assessment, and their main purpose is to elicit evidence (which is observable) about competencies (which are unobservable).

15.2.1.4 Design and Diagnosis

As shown in Fig. 15.2, assessment design flows from left to right, although in practice it is more iterative. Diagnosis (or inference) flows in the opposite direction. That is, an assessment is administered, and the students' responses made during the solution process provide the evidence that is analyzed by the evidence model. The results of this analysis are data (e.g., scores) that are passed on to the competency model, which in turn updates the claims about relevant competencies. In short, the ECD approach provides a framework for developing assessment tasks that are explicitly linked to claims about student competencies via an evidentiary chain (i.e., valid arguments that connect task performance to competency estimates), and are thus valid for their intended purposes. New directions in educational and psychological measurement promote assessment of authentic activities and allow more accurate estimations of students' competencies. Further, new technologies let us administer formative assessments during the learning process, extract ongoing, multi-faceted information from a learner, and react in immediate and helpful ways, as needed.

The following section describes our ideas for embedding assessments within multimedia environments, such as games and simulations.

15.2.2 Stealth Assessment

When embedded assessments are so seamlessly woven into the fabric of the learning environment that they are virtually invisible, we call this stealth assessment (see Shute, Ventura, Bauer, & Zapata-Rivera, in press). Such assessments are intended to support learning, maintain flow, and remove (or seriously reduce) test anxiety, while not sacrificing validity and reliability (Shute, Hansen, & Almond, 2008). In addition, stealth assessment can be accomplished via automated scoring and machine-based reasoning techniques to infer things that are generally too hard for humans (e.g., estimating values of competencies across a network of skills via Bayesian networks).

In learning environments with stealth assessment, the competency model accumulates and represents belief about the targeted aspects of knowledge or skill, expressed as probability distributions for CM variables (Almond & Mislevy, 1999; Shute, Ventura, et al., in press). Evidence models identify what the student says or does that can provide evidence about those skills (Steinberg & Gitomer, 1996) and express in a psychometric model how the evidence depends on the CM variables (Mislevy, 1994). Task models express situations that can evoke required evidence.

One big question is not about how to collect this rich digital data stream, but rather how to make sense of what can potentially become a deluge of information. Another major question concerns the best way to communicate student-performance information in a way that can be used to easily inform instruction and/or enhance learning. A good solution to the issue of making sense of data, and thereby fostering student learning within immersive environments, is to extend and apply ECD. This provides (a) a way of reasoning about assessment design, and (b) a way of reasoning about student performance in gaming or other learning environments.

We now turn our attention to a literature review and model of a particular key competency – systems thinking skill. Subsequently, we present an example of how to assess this competency within a Quest Atlantis environment (i.e., Taiga Park).

15.2.3 Systems Thinking

> The whole is more than the sum of its parts. (Aristotle)

As noted earlier, rapid changes in today's world have revealed new challenges to and requests from our educational system. Problems facing today's citizens (e.g., global warming, racial and religious intolerance) are complex, dynamic, and cannot be solved unilaterally. Furthermore, many of these problems are ill-structured in that there is not just one correct solution. Instead, we need to think in terms of the underlying system and its subsystems to solve these kinds of problems (Richmond, 1993). The ability to act competently in such complex situations requires competence in systems thinking (ST) (Arndt, 2006).

15.2.3.1 Definitions of Systems Thinking

Definitions of systems thinking tend to focus on the relationships between elements in a given environment. Barak and Williams (2007) define ST as the ability to describe and analyze structures and phenomena in natural, artificial, and social environments. Similarly, Salisbury (1996) defines ST as being able to consider all of the elements and relationships that exist in a system, and know how to structure those relationships in more efficient and effective ways. In general, a system can be defined as a group of parts or components working together as a functional unit (Ossimitz, 2000; Salisbury, 1996). A system can be physical, biological, technological, social, symbolic, or it can be composed of more than one of these (Barak & Williams, 2007). Furthermore, many systems are quite complex (e.g., the ecosystem of the world and the human body). To understand the behavior of such complex systems, we must understand not only the behavior of the parts, but also how they act together to form the behavior of the whole. Thus, complex systems are difficult to understand without describing each part and each part must be described in relation to other parts (Bar-Yam, 1997).

Each system consists of closed-loop relations, and system thinkers use diagramming languages and methods to visually represent the relations and feedback

structures within the systems. They also use simulations to run and test the dynamics to see what will happen (Richmond, 1993). The National Science Education Standards (National Research Council, 1996) identifies systems as an important and unifying concept that can provide students with a "big picture" of scientific ideas which can then serve as a context for learning scientific concepts and principles. Thus, a strong background in systems thinking is critical to understanding how the world works.

15.2.3.2 Systems Thinking and Its Role in Education

Traditional teacher-centered approaches to education may be less suitable than learner-centered approaches for teaching and bolstering ST skills, especially skills related to considering, understanding, and solving complex problems (Arndt, 2006). This is because in many teacher-centered classrooms students try to assimilate content that is presented by the teacher (Brown, 2003). Students are typically not engaged in ST beyond perhaps repeating back the teacher's thoughts and interpretations. Although students encounter much content, they do not often learn what to do with it. Thus, this type of learning really does not help much when confronted with novel, complex problems (Arndt, 2006; Richmond & Peterson, 2005). Furthermore, this approach is poorly suited for the transfer of solutions to similar classes of problems. It comes as no surprise that most facts taught and learned via the traditional approach are quickly forgotten (Arndt, 2006). As a consequence, the expectations and needs for a twenty-first century educational system are being inadequately met in settings where students have minimal control of their own learning.

Alternatively, learner-centered approaches are based on the notion that learning is primarily a construction rather than an assimilation process. To learn, the student must construct or reconstruct what is being taken in (Richmond, 1993; Shute, 2007). Students who engage in ST have to actively construct functional relations among relevant components, either mentally or externally.

15.2.3.3 The Competency Model of Systems Thinking

To assess and support ST within a school environment, it is possible to construct indicators for important aspects of systems thinking (Assaraf & Orion, 2005). Having a good competency model should permit educators to collect data about students' knowledge of and performance on a set of tasks requiring the application of ST skills. This information could then be used to make inferences about students' current ST competency levels, at various grain sizes, for diagnostic, predictive, and instructional purposes. Our proposed ST competency model consists of three first-level variables: (1) specifying variables and problems in a system, (2) modeling the system, and (3) testing the model via simulation (see Fig. 15.3). Each of these first-level variables has a number of "progeny" and each will now be described in turn.

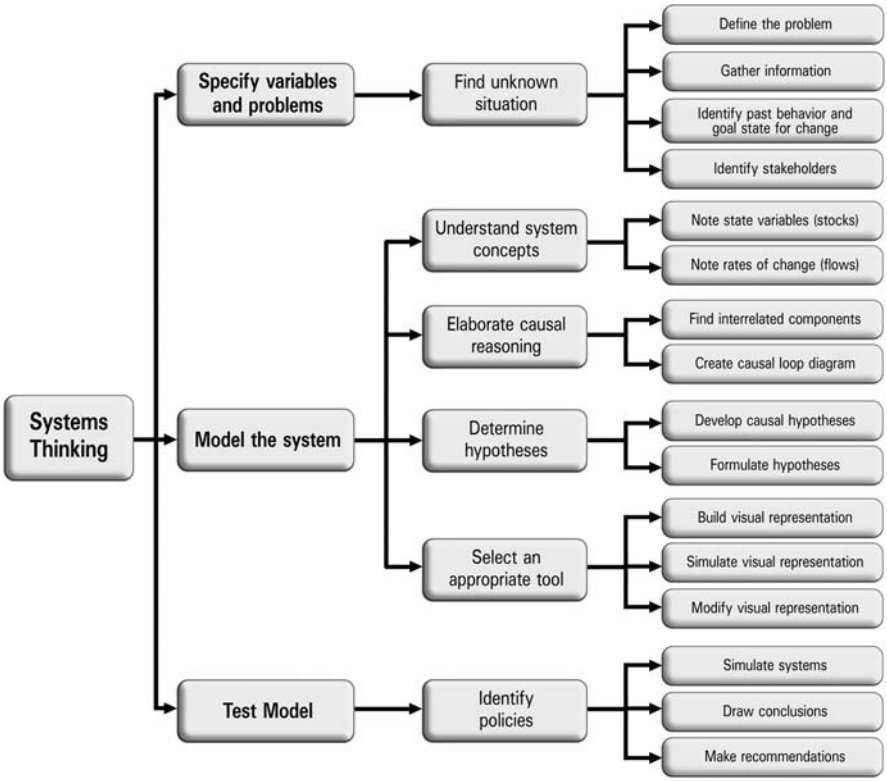


Fig. 15.3 Competency model of systems thinking

Specify Variables and Problems

We believe that the ST process begins by defining problems, formulating and testing potential solutions, and distinguishing fundamental causes of problems (Walker, Greiner, McDonald, & Lyne, 1998). So what exactly is a problem? Jonassen (2004) defines at least two critical features of a problem. The first relates to an unknown entity within some context (i.e., the difference between a goal state and a current state). The second aspect relates to finding or solving the unknown, which must have social, cultural, or intellectual value. Finding the unknown within a problem is important because if no one perceives an unknown, or even a need to determine an unknown, then there is no perceived problem. After defining a problem, system components can be specified in relation to that problem. The best way to determine system components is to answer questions about causality such as: "What causes overpopulation?" Some relevant answers may include: poverty, lack of education, inadequate birth control resources, etc.

Model the System

Conceptual modeling is one of the main tools used to show thinking about a system. The intent of a model is to identify the feedback structures that control behavior. By making these structures explicit, the process helps us share our thoughts with others and simplify complex things. That is, because many elements of a system cannot be observed directly, models help us to visualize and externalize those elements (Jonassen, Strobel, & Gottdenker, 2005; Salisbury, 1996). Fortunately, today's computer technologies allow us to simulate almost any complex situation that we might want to study. Computer simulations also highlight and make visible otherwise hidden processes such as planning, decision making, and evaluation processes (Dörner, 1997). One of the most well-known ST tools is called STELLA (Systems Thinking in an Experiential Learning Laboratory with Animation; see Mills & Zounar, 2001; Salisbury, 1996). Other software applications that are appropriate for creating system diagrams and models in educational settings include: Powersim, Vensim, Modus, Dynasis, and CoLab.

A particularly difficult part of modeling complex systems concerns *interactions* because no action is unilateral in its impact. When one element of a system is changed, it in turn influences other elements of the system. Thus, ST requires an understanding of the dynamic, complex, changing nature of systems (Salisbury, 1996). To illustrate, consider the *butterfly effect* in Chaos Theory, which describes how very small changes, like the flapping of a butterfly's wings in Miami, can affect extremely large systems, like weather patterns in Paris (for more, see Lorenz, 1995). The focus on interactions within ST contrasts with traditional analysis which typically separates the whole into constituent parts (Aronson, 1996). To understand the whole system and its dynamic interactions, the concepts of stocks and flows are crucial (Mills & Zounar, 2001; Sterman, 2000). *Stocks* can be defined as state variables (or accumulations) which hold the current, snapshot state of the system. Stocks completely explain the condition of the system at any point in time and do not change instantaneously. Rather, they change gradually over a period of time. Stocks can represent concrete materials, such as the amount of water in a lake, or abstract concepts, such as level of happiness. *Flows* represent changes, or rates of change. Flows increase or decrease stocks not just once, but at every unit of time (Martin, 1997). For example, the total accumulation of water within a lake is decreased by evaporation and river outlets while it is increased by precipitation and river inlets. Consequently all system changes through time can be represented by using only stocks and flows.

In addition to fully understanding relevant system terms (i.e., stocks and flows, as well as inputs, processes, and outputs), system thinkers must also be concerned with *feedback loops*. Feedback loops are the structures within which all changes occur (Ossimitz, 2000), a closed chain of casual relationships that feeds back on itself (Georgiou, 2007). In other words, feedback represents information about results that supports the system so that the system can modify its work (Salisbury, 1996). The idea of feedback in systems is the most important concept in understanding a problematic situation in a holistic manner, and it also opens the door for quite

complex understanding. In interrelated systems we have not only direct, but also indirect effects which may lead to feedback loops. Every action, change in nature, etc. is located within an arrangement of feedback loops.

Feedback loops are represented by causal loop diagrams, and there are two types of feedback: positive (reinforcing) and negative (balancing) (Ossimitz, 2000; Sterman, 2006). Negative feedback intends to achieve some steady state. Positive feedback is self-reinforcing, either in terms of growth (regenerative dynamics) or deterioration (degenerative dynamics). Both growth and deterioration eventually collapse the system in the absence of negative feedback (Georgiou, 2007). World population and birth rate have a positive feedback relationship because large populations cause large numbers of births, and large numbers of births result in a larger population. Each may view the other as a cause (Richmond, 1993), reminiscent of the old chicken-or-egg conundrum. Adding another factor into the equation (e.g., death rate) would be an example of a negative feedback loop influencing population. As a final point on the feedback issue, a proper understanding of feedback loops requires a *dynamic* perspective, in order to see how things appear and then change over time (Ossimitz, 2000).

Another distinction that is made in systems thinking is between open-loop and closed-loop systems. Most people tend to think in a linear manner and use linear thinking (i.e., one cause, one effect) to achieve their goals. Such thinking represents an open-loop system (see Fig. 15.4), where you see a problem, decide on an action, expect a result, and the loop ends (Forrester, 1996).

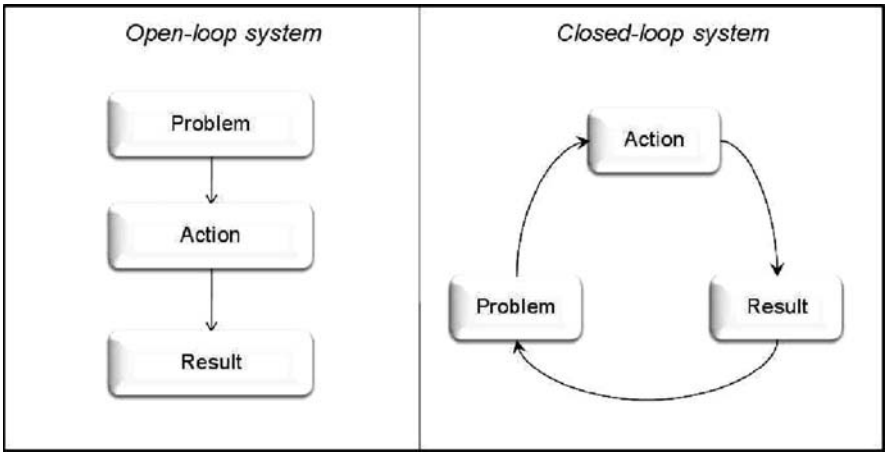


Fig. 15.4 Comparing open-loop and closed-loop systems

However, the real-world does not consist of simple linear relations but of complex relations that are highly interconnected and dynamic. Consequently, the behavior of real systems is often difficult to anticipate because it may be counter-intuitive, nonlinear, and irreversible. As a result, linear thinking applied to complex systems is likely to fail (Senge, 1994; Sterman, 2000). To illustrate, think about

the factors effecting gasoline prices in the United States. Increasing and decreasing gasoline prices depend on a whole host of factors (e.g., value of the US dollar, supply, demand, OPEC capacity, war effects, Wall Street crises, etc.) and these factors have complex relations with one another. To solve complex problems (like predicting gas prices or tracking hurricane trajectories), people need to think in terms of the "big picture" and about how variables are related to each other rather than in terms of discrete, detailed facts. ST requires knowing about the individual parts of a system, the role each part plays, and how these parts interact to function as a whole (Assaraf & Orion, 2005). In real-life, after gathering information about a problem, this usually leads to some action that produces a result. But in actuality, there is no beginning or end. Instead, the process is iterative (i.e., a closed-loop system; see the right side of Fig. 15.4). So, systems are never totally open. If a system *were* totally open, then it would have no orderly interaction with its environment.

Test the Model

After conceptually modeling the system, the next step involves actually testing out the model. This entails simulating the system (via computational models), running the model, and then drawing conclusions and making decisions based on the obtained results (Richmond & Peterson, 2005). The actual results are compared with the expected results and significant differences must be examined carefully. Differences can be described by computer models. The examination process of unexpected simulation results contains significant opportunities for learning because it requires intensive reflection by the student, as well as adaptation of one's mental model (Sterman, 2000).

15.3 Application of the Stealth Assessment Approach

> Reason does not work instinctively, but requires trial, practice, and instruction in order to gradually progress from one level of insight to another. (Immanuel Kant)

The purpose of this worked example of the systems thinking competency is to test the viability of our stealth assessment approach within an existing immersive game. In the example that follows, we first briefly describe the game (Quest Atlantis: Taiga Park), an immersive, role-playing game set in a modern 3D world (see Barab, Sadler et al., 2007). Next, we present an ECD formulation relating to systems thinking skill as applied and assessed during game play. Finally, we compare a hypothetical player at two different points in time (at the beginning and more advanced stages of learning) in relation to her ST skill.

15.3.1 Quest Atlantis: Taiga Park

Taiga is the name given to a beautiful virtual park with a river running through it (Barab, Zuiker et al., 2007; Zuiker, 2007). The park is populated by several groups of people who use or depend on the river in some capacity. Although the groups

are quite different, their lives (and livelihoods) are entwined, demonstrating several levels of "systems" within the world (e.g., the ecological system comprising the river and the socio-economic system comprising the groups of stakeholders in the park). In addition to the park ranger (Ranger Bartle), the three stakeholders include: (a) the Mulu (indigenous) farmers (e.g., Norbe and Ella); (b) Build-Rite Timber Company (e.g., Manager Lim, Lisa, and Hidalgo); and (c) the K-Fly Fishing Tour Company (e.g., Markeda and Tom). There are also park visitors, lab technicians, and others with their own sets of interests and areas of expertise.

The Taiga storyline is about how the fish population in the Taiga River is dying. Students participate in this world by helping Ranger Bartle figure out how he can solve this problem of the declining fish population and thus save the park. Students begin the series of five missions by reading an introductory letter from Ranger Bartle. In the letter, Ranger Bartle pleads for help and states his need for an expert field investigator (i.e., you, the player/student) who can help him solve the declining-fish-population problem. As part of the first mission, a student has to interview 13 different characters throughout the park. Each of them is affiliated with one of the park's main stakeholders. By interviewing the various characters, students "hear" from each one of them about what causes the fish decline in the river – consisting of both opinions and facts about the problem. It soon becomes obvious that the three main stakeholders blame each other, and also that there are more complex problems than just the declining fish problem. At the end of the first mission, students are required to formulate and state an initial hypothesis about the fish-decline problem. This hypothesis is not based on scientific evidence, but on what was heard from the different stakeholders.

For the second mission, students collect water samples from three different sites and analyze the water quality based on six indicators, such as pH level, temperature, and turbidity. Students must submit their interpretation of the water quality data, and also explain which human activities (e.g., fishing, farming, and logging) at each of the three water collection sites cause the problem and how they are interrelated. After completing the second mission, students receive a message from Jesse, Ranger Bartle's intern, which initiates the third mission. The third mission is similar to the second, but focuses on reasoning about the data that has been collected, and drawing a preliminary scientific conclusion based on the hypothesis rendered in the preceding mission.

The fourth mission is set 2 years in the future. It starts with the student being required to name one of the stakeholders as the key culprit in terms of the fish-decline problem. Using a time machine (woven neatly into the narrative), and exploring Taiga 2 years in the future, students can see that ignoring the larger picture (i.e., interrelationships among the stakeholders) and focusing on a simple causal hypothesis and ensuing solution does not work. For instance, suppose that a student blamed the loggers for the fish-decline problem (i.e., logging causes erosion that increases the river's turbidity which leads to gill damage and ultimately death in fish). On the basis of this hypothesis, the park ranger "solves" the problem by ridding the park of the loggers. The future results of the logger-removal decision show that the problem has yet to be solved. Erosion continued because nobody replanted

trees, the farmers had to increase farming activities to offset lost revenue from the rent no longer received from the loggers, the fish population continued to suffer and decline, and the park found itself on the brink of disaster. To complete this mission, the student has to explore the future park and explain what has occurred, answering the following questions: (a) Why does blaming just one group create a whole set of different problems? and (b) How can the set of problems be resolved?

The fifth and final mission in Taiga requires students to think of the park as a system, and generate a more coherent hypothesis in relation to the problem, on which the park ranger will act. Students then again employ the time machine to travel 5 years into the future where they view the new version of Taiga Park based on their systemic solution to the problem (i.e., involving both environmentally and economically sustainable solutions). By interviewing different people in Taiga in the future, students identify which changes occurred and how they reflect a socio-scientific solution. In terms of the various *levels* of systems mentioned earlier, students should understand (a) local level systems; i.e., the fragile and interconnected nature of our various ecological systems, like in and around rivers; and (b) socio-economic level systems, like those shown by the entwined relationships among the Taiga stakeholders.

The Taiga Teacher's Guide for this unit notes that activities have been designed around formalized scientific understanding and science learning standards. The five core scientific concepts in the unit include: erosion, eutrophication, water quality indicators (e.g., turbidity, dissolved oxygen), watersheds, and formulating and evaluating hypotheses. Also, through participating in this unit, students are expected to develop valuable skills such as socio-scientific reasoning, scientific inquiry, and scientific decision making. From their experiences in Taiga, students are expected to develop an appreciation for the complexities involved in scientific decision making by balancing ethical, economic, political, and scientific factors (e.g., the best solution from a scientific perspective can be conflicting with political or economic perspectives). Eventually, students are expected to develop deep environmental awareness by appreciating the complexity of environmental problems.

15.3.2 ECD Models Applied to Taiga

Taiga Park, with its requirement for socio-scientific inquiry as well as continuous reflection and revision of current understanding, is an ideal environment to demonstrate the use of ECD for systems thinking. In their role as an expert assistant to the park ranger, students interview stakeholders, collect data, and develop hypotheses about why the fish population in Taiga is declining. Eventually (i.e., in their final mission), the students are expected to recommend a systems-based solution to the park ranger based on their final hypothesis concerning all of the variables affecting the decline in Taiga's fish population.

As described earlier, one important aspect of systems thinking requires a person to conceptualize a model of the system. The main purpose of conceptual modeling

is to help a person visualize and externalize elements and relations within a system, and to improve understanding of the dynamic interactions among the different components of a system (i.e., the stocks and flows). To view a problem in a holistic manner, students need to understand how feedback works within a particular system. For instance, feedback loops demonstrate the direct and indirect effects within systems, and causal loop diagrams demonstrate students' understanding of how component changes affect other parts of the system. Once the causal relationships and feedback loops have been established, students should be able to form hypotheses about the relationships within the system. To determine whether a hypothesis is correct, some form of simulation is needed to demonstrate the stated relationships between system components. This process enables students to then modify the original hypothesis. Fortunately, in Taiga Park, there is a time machine. This clever narrative device permits one to simulate consequences of particular actions at various points of time in the future.

Figure 15.5 shows a conceptualization of the ECD models for a fragment of the ST competency (i.e., *Model the System*). Notice that "competency model" and "evidence model" are the same terms as we used in the previous ECD discussion. However, when extending to game environments, we use the term "action model" instead of task model. An action model reflects the fact that we are dynamically modeling students' actions within the particular game. These actions form the basis for gathering evidence and rendering inferences and may be compared to simpler

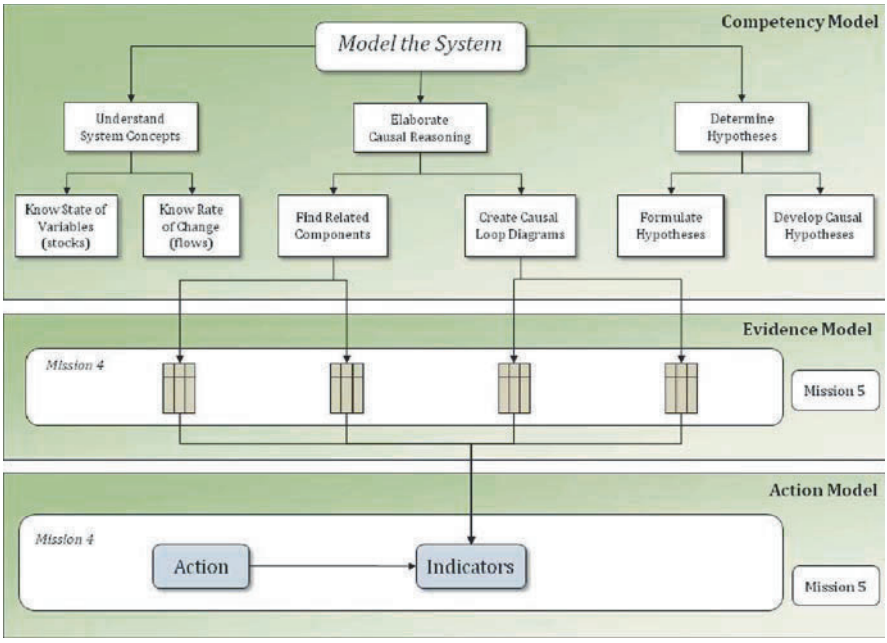


Fig. 15.5 Conceptualization of ECD models applied to Taiga

task responses as with typical assessments. The lined boxes shown within the evidence model denote what are called conditional probability tables (CPTs). These CPTs represent the statistical relations (or "glue") between the indicators (observable) and competencies (unobservable). Finally, note that "mission" is used to define a set of required actions within a particular quest.

Competency Model: By the time students reach Mission 4 in Taiga, they have (a) interviewed a variety of people who have a stake in the park, (b) collected water samples from three different points along the river, and (c) taken snapshots at five observation posts located along the river. Thus in mission 4, students need to demonstrate an understanding of how the water quality indicators (e.g., turbidity, pH level, temperature) relate to the activities along the river – specifically in relation to their effects on the fish population. Additionally, students should be able to draw a causal diagram that shows the stocks and flows of the components that are reducing the population of fish in the river.

Evidence Model: This model is established to determine how the observable aspects of the students' actions in the game may be used (i.e., collected and aggregated) as evidence for the competency variables. The evidence model contains: (a) outcomes from the assigned tasks such as diagrams created or short answers provided to specific questions, (b) rules for scoring the student submissions, and (c) indicator weights in relation to associated competencies.

Action Model: Similar to the task model, the action model in a gaming situation defines the sequence of actions, and each action's indicators of success. Actions represent the things that students do to complete the mission. Some of the required actions are sequential in nature and must be completed in order to proceed within the mission. Other actions can occur at any point in time, and as often as desired. Table 15.1 lists a few representative actions and their indicators relevant to various Taiga missions.

In the current version of Taiga, students write and submit short essays to their teachers as a required part of the missions. The teacher then reviews the essays, using a set of rubrics to score them. For example, a student may receive maximum points (and earn a badge) for an essay answer that demonstrates: (a) an ability to interpret water-quality indicators, (b) an understanding of ecological processes, and (c) the capability to integrate evidence (obtained during missions) and the associated processes. Students falling short of the criteria are advised to visit the water expert at Taiga to discuss the water indicators and ecological processes again. They are also told to revise and resubmit their essays if they wish to receive the badge of completion.

In addition to the essays, students can create and submit *causal loop diagrams* (demonstrating the stocks and flows within the system and their cause-effect relationships). In the current version of the game, such diagrams may be uploaded as an attachment to student essays, but they are optional. One problem with the current implementation is the large burden it places on teachers to not only monitor their students' game play, but additionally to carefully read and score all essays, interpret and assess the quality of all submitted causal diagrams, as well as provide feedback to support students' learning. Also, there may be ambiguity in diagrams and

Table 15.1 List of actions and associated indicators

Action	Indicators
Summarize water quality indicators along the river	Accurately note water quality indicators for 3 points along the river Accurately note whether indicators signify good or bad water quality
Explain how water-quality data account for fish death	Correctly explain how the indicators are symptoms of erosion and eutrophication Correctly link these ecological processes to the population of fish in Taiga River
Explain how the various stakeholders contribute to the fish-decline problem	Correctly identify stakeholders and their main activities near the river Correctly relate these activities to erosion and eutrophication
Create causal loop diagram	Include complete set of variables and links in the diagram Accurately identify relationships among variables (positive or negative)
Evaluate a hypothesis	Correctly identify one group responsible for the problem at Taiga Accurately explain and/or depict how this group's activities lead to ecological processes detrimental to the fish

subjectivity in assessing, on the teachers' parts. Moreover, crafting causal diagrams, we believe, should be an integral (not optional) part of the game.

15.3.2.1 Tools to Automatically Assess Causal Diagrams

If causal diagrams were required in the game, how could we automate their assessment? Solving this issue would reduce teachers' workload, increase the reliability of the scores, and clearly depict students' current mental models (or conceptualizations) of various systems operating within Taiga. Students' causal diagrams can be created using one of several computer-based tools designed for this purpose (e.g., CmapTools, by Cañas et al., 2004; freeware which can be downloaded from: http://cmap.ihmc.us/conceptmap.html). There are currently quite a few tools and technologies emerging whose goal is to externalize and assess what are otherwise internal conceptions (e.g., see Shute, Jeong, Spector, Seel, & Johnson, in press). The tool that we focus on in this illustration is an Excel-based software application called jMap (Jeong, 2008; Shute, Jeong, & Zapata-Rivera, in press), designed to accomplish the following goals: (1) elicit, record, and automatically code mental models; (2) visually and quantitatively assess changes in mental models over time; and (3) determine the degree to which the changes converge toward an expert's or the aggregated group model (for more information about the program, including links and papers, see: http://garnet.fsu.edu/~ajeong).

With jMap, students create their causal maps using Excel's autoshape tools. Causal links are used to connect a collection of variables together, and link strength

may be designated by varying the thicknesses of the links (not relevant in the following worked example). In jMap, comparisons between a student's and a target map[1] begin by automatically coding/translating each map into a transitional frequency matrix. For instance, if the target map contained eight variables comprising a complete causal diagram, this would translate to an 8×8 frequency matrix representing all pairwise linkages (see Table 15.2). Each observed link within the student's map is recorded into the corresponding cell of the matrix.

Table 15.2 Example of a transitional frequency matrix

Transitional Frequency Matrix	Taiga Park income	Need more logging	Cutting trees	Soil erosion	Sediment in water	Temperature of water	Dissolved oxygen	Fish population
Taiga Park income								
Need more logging								
Cutting trees								
Soil erosion								
Sediment in water								
Temperature of water								
Dissolved oxygen								
Fish population								

Once all (i.e., student and expert maps) have been automatically tabulated into transitional frequency matrices, jMap can be used to superimpose: (a) the map of one learner produced at one point in time over a map produced by the same learner at a later point in time; (b) the map of one learner over the map of a different learner; or (c) the map of a learner over the map of an expert. jMap can also be used to aggregate all the frequencies across the frequency matrices of multiple learners to produce an aggregate frequency matrix representing the collective group. As a result, the resulting collective group map can also be superimposed over an individual learner's map or an expert map. Users (e.g., teachers, researchers, students,

[1]The target map is usually an expert's map, but may be another student map (e.g., the same student at different times, a different student, or even a group of students). See Shute, Jeong, and Zapata-Rivera (in press) for examples.

etc.) can toggle between maps produced over different times to animate and visually assess how maps change over time and see the extent to which the changes are converging toward an expert or group map. Additional jMap tools enable users to compile raw scores to compare quantitative measures (e.g., the percentage of shared links between the compared maps).

In this proposed scenario, and as part of their gaming mission, students would draw their causal diagrams using jMap, which would contain a collection of relevant system concepts or stocks. Students would choose relevant variables from the collection, and link them together, similar to completing a puzzle, into a causal diagram. This activity would (a) take place within the Taiga narrative (e.g., as part of a task assigned to the student by Ranger Bartle), and (b) demonstrate students' emerging understanding of the interrelatedness of relevant concepts. The submitted maps would then be automatically compared in terms of propositional structure with an expert (or target) map. Higher similarity indices between the two would lead to higher estimates for the relevant competency.

15.3.2.2 Adding Stealth Assessment to Taiga

To illustrate this automated, evidence-based assessment methodology within Taiga, we implemented a part of the ECD model relating to systems thinking skill, and focused on the competency: *Model the System*.

Figure 15.6 shows the initial state of the network. When a student performs an action in the game (e.g., creates a causal loop diagram), relevant indicators are calculated. For this example, the indicators include (a) accuracy/completeness of the variables included in the diagram, and (b) accuracy of the links established (i.e., positive versus negative relations). These comprise the set of indicators associated with that particular node (see Table 15.1). The indicator data, derived from the jMap tool, are then automatically inserted into the Bayes net which is instantly updated with new probability values propagated throughout the network.

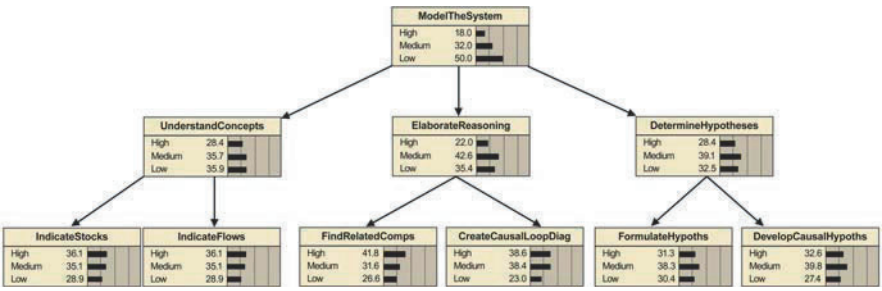


Fig. 15.6 Initial Bayesian model for a fragment of systems thinking skill

Consider a hypothetical student named Clara. Suppose we have two causal loop diagrams obtained from her at two different points in time: during an early mission in Taiga, and then during her final mission. During the early mission, Clara blamed

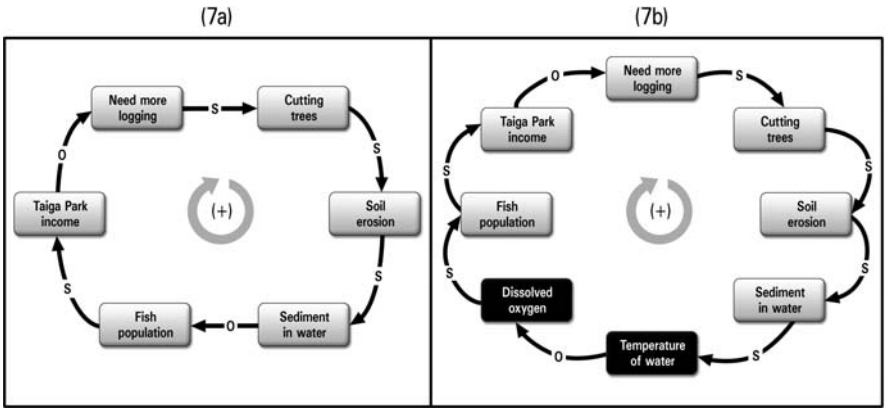


Fig. 15.7 Clara's causal loop diagram at Time 1 (**a**) and an expert diagram of the system (**b**)

the decline-in-fish-population problem solely on the loggers. Her causal loop diagram at that point is shown in Fig. 15.7a (see left panel). The full set of variables available in the jMap collection includes those shown in her diagram, as well as others such as dissolved oxygen in the water, temperature of the water, pH level of the water, and so on. The relationships between variables are also recorded directly in the diagram using an "S" (for same, denoting a positive function) or an "O" (for opposite, for an inverse function).

At this relatively early stage of learning, Clara appears to have a basic understanding of what is going on in the river relative to the logging business, but does not yet fully understand all of the variables that cause a decrease in the fish population. If her diagram was compared to an expert's (using jMap), her errors of omission would suggest that she believes sediment in the water directly and negatively affects the fish population. However, sediment in the water actually serves to increase water temperature, which in turn causes a decrease in the dissolved oxygen. Inadequate oxygen would cause fish to die. This provides the basis for valuable feedback to Clara, which could be automatically generated, or provided by the teacher (e.g., "Nice job, Clara – but you forgot to include the fact that sediment increases water temperature which decreases the amount of dissolved oxygen in the water. That is the reason the fish are dying – they do not have enough oxygen"). In addition, the lab technician (or another knowledgeable character in Taiga) could provide feedback in the form of a causal loop diagram, explicitly including those variables in the picture. That way, she can see for herself what she had left out. See the right panel in Fig. 15.7b for an example of an expert diagram, highlighting her omitted variables and links.

When she visits Taiga 2 years in the future, Clara would quickly realize that her simple conceptualization of the problem (i.e., blaming just a single group of Taiga stakeholders – the loggers) and the ensuing solution (i.e., Ranger Bartle's banning the loggers from Taiga Park) was in vain. That is, 2 years into the future, she sees

converging evidence that the fish population is still suffering – perhaps even worse than before. Over the course of additional actions and interactions in Taiga (e.g., comparing photos taken along the river at different times, interviewing people in the present and the same people again in the future), she gradually understands the ramifications of her previous solution. That is, because the loggers are gone, the Mulu farmers had to increase their farming operations to offset their lost income (from loggers' rent money). This increase in farming operations resulted in more nutrients from fertilizer running off into the river and affecting the ecosystem (negatively for the fish – positively for the algae); and more toxic waste running off into the river from increased use of pesticides. Many actions and interactions later, Clara eventually comprehends the functional relationships among all three stakeholders and sees how they all are to blame for the problem. This holistic (system) understanding can now provide the basis for an effective solution to the declining-fish-population problem that concurrently addresses all aspects of the issue (i.e., the effects of farming, logging, and fishing tournaments on the fish population). Consequently, she draws a more comprehensive causal diagram (see Fig. 15.8) and recommends various regulations on all three stakeholders to Ranger Bartle.

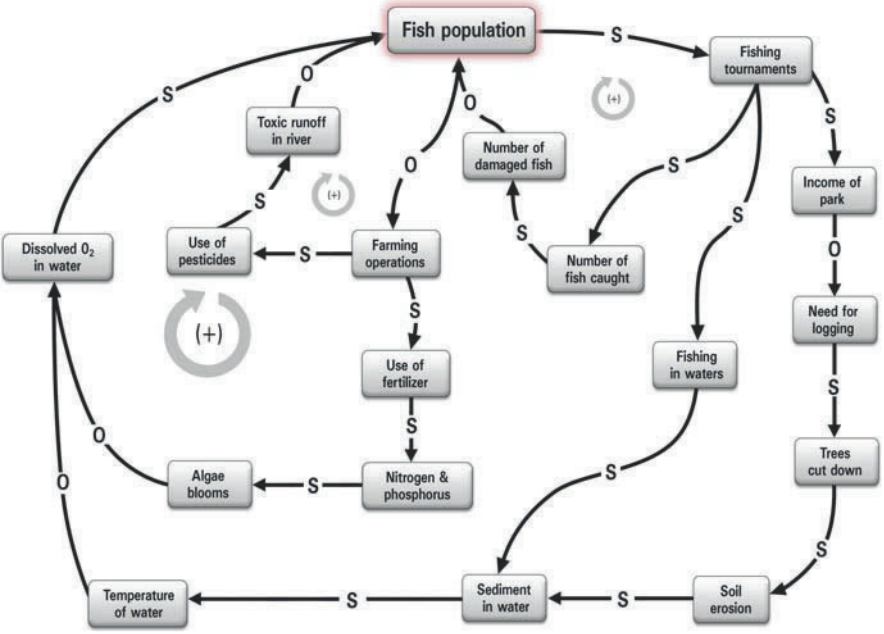


Fig. 15.8 Clara's causal loop diagram – Time 2

So how does jMap derive indicator values to feed into the Bayes net? Let us look at the jMap analysis comparing Clara's Time 1 map to an expert map. As shown earlier, Clara demonstrated incomplete modeling of the system based on her performance on relevant indicators. A screen capture from jMap is shown in Fig. 15.9. Here, jMap's generated diagram uses colored links to clearly and visually identify

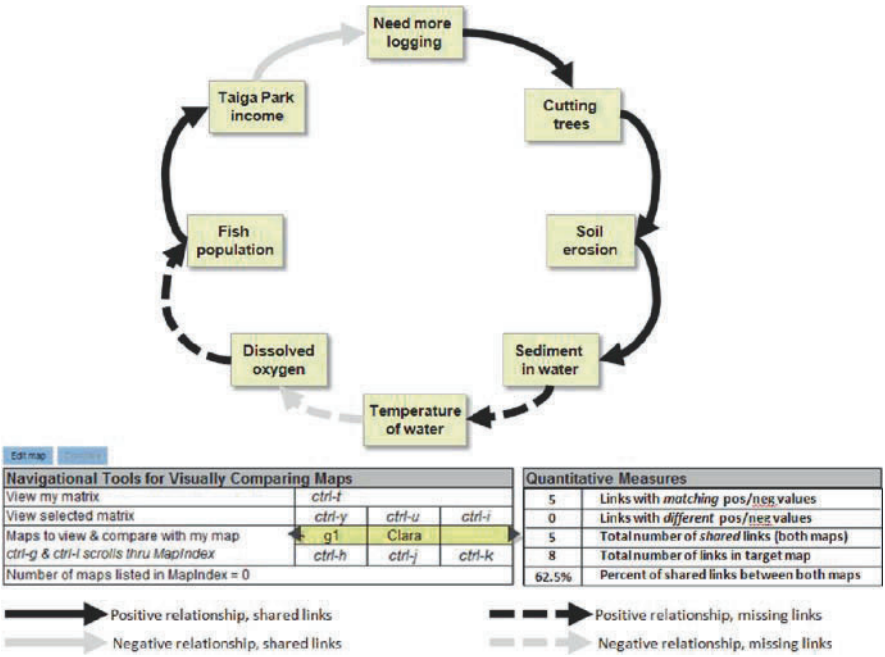


Fig. 15.9 jMAP interface showing a Clara's Time 1 map overlaid on the expert's map

differences between two selected maps – in this case between Clara's Time 1 map and the expert map. Dashed arrows denote *missing links* (i.e., links that are present in the expert map but missing in the student map), and solid arrows denote shared links, which match in terms of identical positive/negative assigned values. The color black represents positive relations and grey represents negative ones. jMap also has the option to represent link strengths (e.g., weak, medium, and strong influences), but we are ignoring link strength in this scenario to make the example easier to understand. By visual inspection, we can see that Clara has omitted three links (and two important variables) in her causal loop diagram relative to the expert's map (shown by the three dashed arrows).

In addition to the standardized maps, the jMAP interface includes two tables, as shown below the map in Fig. 15.9. The table on the left includes navigational tools. These allow the user (e.g., teacher, student, researcher) to easily move among all possible maps using control-key functions, showing the map, the matrix, or both, and compared to the expert model or another model, such as a group model. The table on the right labeled "Quantitative Measures" provides an indication of the similarity between the current map (in this case, Clara at Time 1) and the expert map. The percentage of shared links between the two maps is 62.5%.

If cut-off values were assigned (e.g., 0–33% = low; 34–66% = medium; 67–100% = high), then Clara's accuracy/completeness of her diagram would be classified as medium. Furthermore, because she had created the correct relations of

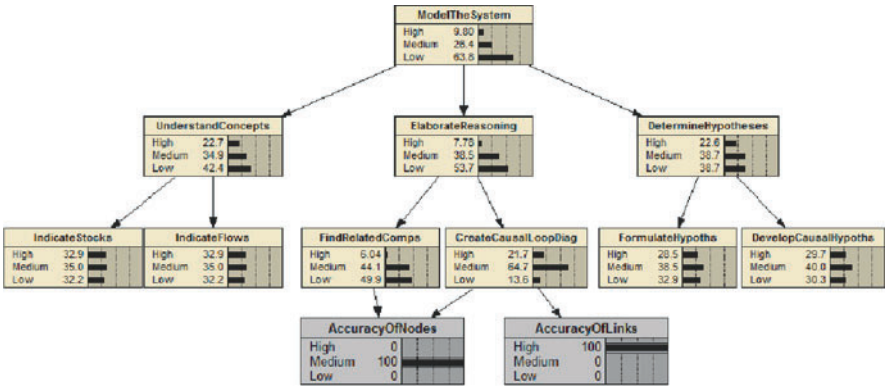


Fig. 15.10 Bayesian model for Clara at Time 1

the links in her diagram (i.e., positive versus negative functions), she would receive a score of "high" on that indicator. These indicator outcomes are then inserted into the Bayes net (see Fig. 15.10).

Once the information is inserted into the Bayes net, it is propagated throughout the network to all of the nodes, whose estimates are subsequently altered. For instance, her Time 1 estimate for the competency, "Create causal loop diagram" is medium; her "elaborate reasoning" competency, however, is estimated at low, as is her overall competency, "model the system." She has more work to do in Taiga, and this analysis and diagnosis targets particular areas for improvement.

By the final Taiga mission, as evidenced in her causal loop diagram shown in Fig. 15.8, Clara has acquired a good understanding of the various systems in Taiga. Her final causal diagram shows the interwoven processes of erosion and eutrophication taking place along the river from the three Taiga communities. The Bayesian model of Clara at Time 2 (not shown) provides evidence of her ability to understand the relationships among system components, with an overall estimate of her "model the system" competency to likely be "high" (i.e., $p(\text{high}) = 0.60$; $p(\text{medium}) = 0.36$; and $p(\text{low}) = 0.04$). This example shows how the outcomes of actions carried out within the game can be used to infer different levels for important competencies in a game environment.

15.4 Summary and Discussion

We presented an innovative approach for embedding evidence-based assessment within an immersive game environment to estimate students' evolving system thinking skills. The ongoing assessment information is intended to provide the basis for bolstering students' competency levels within the game, directly and indirectly. Our approach represents an extension of ECD, which normally entails assessment tasks (or games, simulations, etc.) being developed at the end of the ECD process. But in

this chapter, we illustrated how we can employ an evidence-based approach using an existing game.

The steps of this approach involve the following: (a) define the competency model for systems thinking, independently from the game, via an extensive literature review which is validated by experts (the validation is currently underway); (b) determine indicators of the low-level nodes in the CM relative to particular game actions; (c) specify scoring rules for the indicators; and (d) develop evidence models that statistically link the indicators to particular nodes in the CM via Bayes nets (or any other method for accumulating evidence). Our hypothesis is that the CM (stripped of specific "indicators") should be transferable across environments that require students to engage in systems thinking skill. This type of "plug and play" capability would make the CM scalable, which comprises part of our plans for future research. Finally, we presented just one example of automatically assessing a component of ST (i.e., creating causal loop diagrams). However, other nodes in the model can be easily and automatically assessed, like those that relate to acquiring relevant knowledge (e.g., water-quality indices like turbidity and alkalinity) and skill at gathering pertinent information in the environment (e.g., collecting water samples from different parts of the river and making sense of the data). Additional attributes (e.g., teamwork and communication skills) can similarly be assessed in the game, providing that a CM has been developed and indicators fully identified.

Another near-future research plan includes examining our stealth assessment approach under conditions where there are multiple, valid solutions to a problem (i.e., less-structured scenarios compared to Taiga Park). For instance, we are currently exploring and analyzing other worlds in Quest Atlantis and deriving assessments that pertain to (a) creative problem solving, and (b) multiple-perspective taking, both identified as key competencies for the twenty-first century. In less-structured environments, multiple solutions can be identified by experts in the content area, and each possible solution then converted to a Bayesian network. The higher level competency nodes (reflecting mastery of rules applicable to a wide range of problems within a content area) should be similar, while the lower level indicators reflect different approaches to problem solving (Conati, 2002).

The main problem that we seek to address with this research is that educational systems (in the US and around the world) are facing enormous challenges that require bold and creative solutions to prepare our students for success in the twenty-first century. Part of the solution will require a strong focus on students developing the ability to solve complex problems in innovative ways, as well as the ability to think clearly about systems. We need to identify ways to fully engage students through learning environments that meet their needs and interests (e.g., through well-designed educational games). When coupled with online collaboration with other students (locally and from around the world), such environments additionally have the potential to develop students' communication skills and creative abilities as they become exposed to diverse cultures and viewpoints.

We maintain that not only is it important to determine the skills needed to succeed in the twenty-first century, but also to identify particular methods for designing and developing assessments that are valid and reliable and can help us meet the

educational challenges confronting us today. One looming challenge, as mentioned earlier, concerns the need to increase student engagement. Thus, we have chosen to embed our stealth assessment approach and associated tools within the context of an immersive game (e.g., Quest Atlantis). Through such games, learning takes place within complex, realistic, and relevant environments (although even fantasy games, such as quests within legendary kingdoms involving nonhuman characters, can be used as the basis for assessment and support of valuable skills). Moreover, games can provide for social negotiation where students learn to communicate and collaborate with others on team quests. Such skills are integral parts of many games, and are crucial for players to complete missions. This design feature can help students consider and respect multiple perspectives from other team members who play different roles and have different strengths and backgrounds. Games can also engender ownership of learning since students can choose to complete a particular quest or explore less well-trodden paths to satisfy their curiosity.

The challenge for educators who want to employ games to support learning is making valid inferences about what the student knows, believes, and can do without disrupting the flow of the game (and hence student engagement and learning). Our solution entails the use of ECD which enables the estimation of students' competency levels and further provides the evidence supporting claims about competencies. Consequently, ECD has built-in diagnostic capabilities that permits a stakeholder (i.e., the teacher, student, parent, and others) to examine the evidence and view the current estimated competency levels. This in turn can inform instructional support.

So what are some of the downsides of this approach? Implementing ECD within gaming environments poses its own set of challenges. For instance, Rupp, Gushta, Mislevy, and Shaffer (in press) have highlighted several issues that must be addressed when developing games that employ ECD for assessment design. The competency model, for example, must be developed at an appropriate level of granularity to be implemented in the assessment. Too large a grain size means less specific evidence is available to determine student competency, while too fine a grain size means a high level of complexity and increased resources to be devoted to the assessment. In addition, developing the evidence model can be rather difficult in a gaming environment when students collaborate on completing quests. For example, how would you trace the actions of each student and what he/she is thinking when the outcome is a combined effort? Another challenge comes from scoring qualitative products such as essays, student reflections, and online discussions where there remains a high level of subjectivity even when teachers are provided with comprehensive rubrics. Thus a detailed and robust coding scheme is needed that takes into account the context of the tasks and semantic nuances in the students' submissions. Finally, for the task or action model, issues remain in terms of how the assigned tasks should be structured (or not). While particular sequences of actions (e.g., as in Quest Atlantis) can facilitate more reliable data collection, it might limit the students' ability to explore the environment or go down alternative paths that make games more interesting and promote self-learning. Therefore, when game designers build assessments into the game, they need to find the ideal balance between student exploration and structured data collection.

Currently, Quest Atlantis employs a system that enables teachers to view their students' progress during their missions via a web-based Teachers Toolkit panel. This enables teachers to receive and grade all of the student submissions (which, across the various missions, may start to feel like a deluge). In our worked example, instead of spending countless hours grading essays and diagrams, teachers instead could simply review students' competency models, and use that information as the basis for altering instruction or providing formative feedback (see Shute, 2008). For example, if the competency models during a mission showed evidence of a widespread misconception, the teacher could turn that into a teachable moment, or may choose to assign struggling students to team up with more advanced students in their quests. This harnesses the power of formative assessment to support learning.

In conclusion, our proposed solution using ECD, stealth assessment, and automated data collection and analysis tools is meant to not only collect valid evidence of students' competency states, but to also reduce teachers' workload in relation to managing the students' work (or actually "play") products. This would allow teachers, then, to focus their energies on the business of fostering student learning. If the game was easy to employ and provided integrated and automated assessment tools as described herein, then teachers would more likely want to utilize the game to support student learning across a range of educationally valuable skills. Our proposed ideas and tools within this worked example are intended to help teachers facilitate learning, in a fun and engaging manner, of educationally valuable skills not currently supported in school. Our future research plans include implementing a full systems thinking stealth assessment into the Taiga Park virtual world to test its efficacy in support of students as well as teachers.

References

Abedi, J., & O'Neil, H. F. (2005). Assessment of noncognitive influences on learning. *Educational Assessment*, *10*, 147–151.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223–237.

Arndt, H. (2006). Enhancing system thinking in education using system dynamics. *Simulation*, *82*(11), 795–806.

Aronson, D. (1996). *Overview of system thinking*. Retrieved January 8, 2009, from http://www.thinking.net/Systems_Thinking/OverviewSTarticle.pdf

Assaraf, O. B.-Z., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, *42*(5), 518–560.

Barab, S. A. (2006, Winter). From Plato's Republic to Quest Atlantis: The role of the philosopher-king. *Technology, Humanities, Education, and Narrative*, *2*, 22–53.

Barab, S. A., Sadler, T. D., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: Supporting consequential play. *Journal of Science Education and Technology*, *16*(1), 59–82.

Barab, S. A., Zuiker, S., Warren, S., Hickey, D., Ingram-Goble, A., Kwon, E.-J., Kouper, I., & Gerring, S. C. (2007). Situationally embodied curriculum: Relating formalisms and contexts. *Science Education*, *91*(5), 750–782.

Barak, M., & Williams, P. (2007). Learning elemental structures and dynamic processes in technological systems: A cognitive framework. *International Journal of Technology and Design Education, 17*(3), 323–340.

Bar-Yam, Y. (1997). *Dynamics of complex systems. Studies in nonlinearity*. Reading, MA: Addison-Wesley.

Brown, K. L. (2003). From teacher-centered to learner-centered curriculum: Improving learning in diverse classrooms. *Education*, *124*(1), 49–54.

Cañas, A. J., Hill, R., Carff, R., Suri, N., Lott, J., Eskridge, T., et al. (2004). CmapTools: A knowledge modeling and sharing environment. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), *Concept maps: Theory, methodology, technology*. Proceedings of the First International Conference on Concept Mapping (pp. 125–133). Pamplona: Universidad Pública de Navarra.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence*, *16*(7–8), 555–575.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.

Dörner, D. (1997). *The logic of failure: Recognizing and avoiding error in complex situations.* New York: Metropolitan Books.

Farkas, G. (2003). Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology*, *29*, 541–562.

Forrester, J. W. (1996). *System dynamics and K-12 teachers*. Retrieved August 8, 2008, from Massachusetts Institute of Technology (MIT), Systems Dynamics in Education Project Web site: http://sysdyn.clexchange.org/sdep/Roadmaps/RM1/D-4665-4.pdf

Gee, J. P. (2004a). *Situated language and learning: A critique of critical schooling.* London: Routledge.

Gee, J. P. (2004b). *What video games have to teach us about literacy and learning.* London: Palgrave Macmillan.

Georgiou, I. (2007). *Thinking through systems thinking.* London: Routledge.

Jeong, J. C. (2008). *Discussion analysis tool (DAT)*. Retrieved December 22, 2008, from http://garnet.fsu.edu/∼ajeong/DAT

Jonassen, D. H. (2004). *Learning to solve problems: An instructional design guide*. San Francisco, CA: Pfeiffer.

Jonassen, D., Strobel, J., & Gottdenker, J. (2005). Model building for conceptual change. *Interactive Learning Environments*, *13*(1), 15–37.

Lorenz, E. N. (1995). *The essence of chaos*. Seattle, WA: University of Washington Press.

Martin, L. A. (1997). *Road map 2: Beginner modeling exercise*. MIT System Dynamics in Education Project. Retrieved August 5, 2008, from Massachusetts Institute of Technology (MIT), Systems Dynamics in Education Project Website: http://sysdyn.clexchange.org/sdep/Roadmaps/RM2/D-4347-7.pdf

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23.

Mills, I. J., & Zounar, E. D. (2001). *On the application of system dynamics to the integration of national research and K-12 education*. Paper presented at the International Conference on Engineering Education, Oslo & Bergen, Norway.

Mislevy, R. J. (1994). Evidence and inference in educational assessment, *Psychometrika*, *12*, 341–369.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED483399)

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3–62.

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

Ossimitz, G. (2000). *The development of systems thinking skills using system dynamics modeling tools*. Retrieved August 13, 2008, from Universität Klagenfurt, Institut für Mathematik, Statistik und Didaktik der Mathematik Website: http://wwwu.uni-klu.ac.at/gossimit/sdyn/gdm_eng.htm

Park, O., & Lee, J. (2003). Adaptive instructional systems. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 651–685). Mahwah, NJ: Lawrence Erlbaum.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review*, *9*(2), 113–133.

Richmond, B., & Peterson, S. (2005). *An introduction to systems thinking: STELLA software*. Lebanon, NH: High Performance Systems, Inc.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (in press) Evidence-centered design of epistemic games: Measurement principles for complex learning environments.

Salisbury, D. F. (1996). *Five technologies for educational change: Systems thinking, systems design, quality science, change management, instructional technology.* Englewood Cliffs, NJ: Technology Publications.

Senge, P. M. (1994). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday/Currency.

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139–187). New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Shute, V. J., Dennen, V. P., Kim, Y. J., Donmez, O., & Wang, C.-Y. (2008). *21st century assessment to promote 21st century learning: The benefits of blinking.* Unpublished manuscript, Florida State University, Tallahassee.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it – Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education*, *18*(4), 289–316.

Shute, V. J., Jeong, A. C., Spector, J. M., Seel, N. M., & Johnson, T. E. (2009) Model-based methods for assessment, learning, and instruction: Innovative educational technology at Florida State University. In M. Orey (Ed.), *2009 Educational media and technology yearbook*, Westport, CT: Greenwood Publishing Group.

Shute, V. J., Jeong, A. C., & Zapata-Rivera, D. (in press).Using flexible belief networks to assess mental models. In B. B. Lockee, L. Yamagata-Lynch, & J. M. Spector (Eds.), *Instructional design for complex learning*. New York: Springer.

Shute, V. J., Lajoie, S. P., & Gluck, K. A. (2000). Individualized and group approaches to training. In S. Tobias & J. D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military* (pp. 171–207). New York: Macmillan.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: Theories and applications*. Philadelphia: Routledge/LEA.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). The MIT Press.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence* (pp. 3–37). New York: Cambridge University Press.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science*, *24*, 223–258.

Sterman, J. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Boston: Irwin/McGraw-Hill.

Sterman, J. (2006). Learning from evidence in a complex world. *American Journal of Public Health*, *96*(3), 505–514.

Walker, P. A., Greiner, R., McDonald, D., & Lyne, V. (1998). The tourism futures simulator: A systems thinking approach. *Environmental Modeling and Software*, *14*(1), 59–67.

Zuiker, S. (2007). *Transforming practice: Designing for liminal transitions along trajectories of participation.* Unpublished doctoral dissertation, Indiana University, Indiana.

Chapter 16
A Methodology for Assessing Elicitation of Knowledge in Complex Domains: Identifying Conceptual Representations of Ill-Structured Problems in Medical Diagnosis

Tiffany A. Koszalka and John Epling

16.1 Introduction

16.1.1 Assessing Learning in Complex Domains

Assessing learning progress is complicated at best. Identifying learning progress in complex domains that regularly require higher-order thinking (e.g., identification and shaping of the problem and problem goal, identification of hidden factors, analysis of situation and origin of facts) to solve ill-defined problems is even more difficult (Huber, 1995). An ill-defined problem is characterized by uncertainty with regard to (a) initial problem states or inputs, (b) desired output states, and/or (c) transformations that will guarantee success in attaining desired goals (Dörner, 1996; Gogus, Koszalka, & Spector, 2009; Jonassen, 2000; Spector & Koszalka, 2004). Lack of certainty in one or more of these three aspects can be due to external factors (e.g., there is no evidence of a particular item or existing data are fuzzy or vague) or to internal factors (e.g., the problem-solver is unsure about the effects of performing a particular action). Such problems often arise in complex domains that have many interrelated factors, nonlinear relationships among some factors, complex internal feedback mechanisms, and potentially delayed effects.

Medical diagnosis is an example of a domain that is comprised of complex and ill-defined problems that require higher-order thinking and a strong understanding of presented problems and mechanisms of disease to resolve successfully. Expertise in this type of domain can be defined as the ability to respond effectively and meaningfully to ill-defined problems (Dörner, 1996; Klein, 1998). The difficulty lies in determining whether and to what extent learners are making progress in their abilities to solve such complex problems that often lack clarity, have multiple goals,

T.A. Koszalka (✉)
Syracuse University, Syracuse, NY, USA
e-mail: takoszal@syr.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_16,

lack complete information, contain multiple interrelated variables, are dynamic in nature, and have multiple acceptable solutions and approaches to resolution (Funke, 2001; Funke & Frensch, 2007).

The type of multifaceted learning and problem-solving skills required to address problem situations encountered in complex domains such as medical diagnosis are not easily testable using traditional assessment tools. Some current educators and researchers use the time intensive think-aloud protocols to assess cognitive development and problem-solving abilities in medical diagnosis (Coderre, Mandin, Harasym, & Fick, 2003) where as others continue to evaluate medical students diagnostic problem-solving abilities using multiple choice, portfolio, simulated patient reports, and clinical simulations (Epstein, 2007). The training in medical diagnosis, for example, often includes putting medical students in situations where they are faced with clinical problems, prompted to identify patterns noted with the problems and potential diagnosis, use forward reasoning to identify diagnostic hypotheses based on patient signs, symptoms, and additional information, and backward reasoning to identify data to support and explain their hypothesis (Coderre et al., 2003). To be proficient in medical diagnosis, physicians typically must (a) construct a representation of the problem that accounts for multiple related pieces of evidence, (b) identify a variety of factors related to the problem and its solution, (c) explore and view the problem and evidence from multiple perspectives, and (d) create and implement a viable solution (Norman, 1994; Spector & Koszalka, 2004). Through think-aloud protocol research it was found that fourth year medical students and experienced physicians (novices and experts) who used diagnostic strategies of pattern recognition to organize situational knowledge and solve clinical problems were much more successful in accurately diagnosing clinical presentations (Coderre et al., 2003). This research concluded instruction in medical diagnosis and medical decision making should support the development of mental frameworks or schemes to support higher-order thinking processes that will help novices progress toward expertise.

However, there is a lack of efficient and reliable assessment methodologies that assess learner progress and development of higher-order thinking processes in complex domains. The most empirically established methodology for assessing relative levels of expertise in complex domains, as used during Coderre et al. (2003) research, involves think-aloud protocol analysis (Ericsson & Simon, 1993). The time-consuming nature of think-aloud protocol analysis methodologies, however, is not practical for applications in large-scale instructional or work situations (Spector & Koszalka, 2004). This study was aimed at developing and testing a methodology and tools suitable for assessing progress in developing problem-solving abilities for ill-defined problems, beginning with the identification problem conceptualization patterns. Such a methodology, and accompanying tools, could be used by educators and instructional designers to support learning assessment in complex domains thereby informing curriculum and instructional design solutions for a variety of small-scale or large-scale instructional contexts.

This chapter describes partial findings from a National Science Foundation funded project called “Dynamic Enhanced Evaluation of Problem Solving (DEEP)”

(Spector & Koszalka, 2004). DEEP is a methodology designed to assess learning progress in problem-centered learning environments for complex domains. This chapter presents findings with regard to one of the domains studied, medical diagnosis.

The underlying approach for DEEP is causal influence diagramming developed by system dynamicists to elicit critical system components from experts (Spector & Koszalka, 2004; Sterman, 1994). Causal influence diagramming was developed as a knowledge elicitation and knowledge representation tool and has been validated in several complex domains (Spector, Dennen, & Koszalka, 2005). In the DEEP model there is an assumption that as learners make progress toward expertise they become more like experts in their performance (Ericsson & Smith, 1991) and, as a consequence, more skilled in problem-solving and higher-order reasoning (Jonassen, 2000; Seel, 2003). To validate this assumption, the first step required is to determine whether expert participants exhibit recognizable patterns of problem conceptualizations in response to complex, ill-defined problem scenarios. The second step requires developing a set of measures of similarity between problem-solving patterns of novice respondents so as to show progress over time toward expertise.

The study (Spector & Koszalka, 2004) was aimed at investigating the practical utility of the DEEP methodology based on higher-order causal reasoning (Grotzer & Perkins, 2000). Higher-order reasoning was analyzed based on representations that expert and novice participants created using a simple concept mapping and annotation tool to conceptualize domain specific ill-defined problems in the domain of medical diagnosis. Problems presented to respondents were designed to be similar to authentic medical diagnosis in that they were ill-defined and had open goals. The scenarios were intentionally lacking in information that would be required to develop a complete solution; these problems were ill-defined due to incomplete data and information. The possible outcomes in the scenarios were also left sufficiently open so that the respondent could to determine the desired goal state. Concept maps and causal influence diagrams methods used to allow problem-solvers the type of environment to represent their thinking of ill-structured problems using a minimum of provided structure.

16.1.2 Assessing the Ability to Solve Ill-Defined Problems

Gagné (1985) defined problem solving "as a process by which the learner discovers a combination of previously learned rules which can be applied to achieve a solution for a novel situation (p. 155)." He emphasized that the problem-solving process yields new learning that "may be ways of solving problems in general – in other words, cognitive strategies which can guide the learners' own thinking behavior (p. 156)." However, problems can be described as consisting of distinct structures and complexity. The combination of these structures and complexities are used to define problems as routine and non-routine problems (Mayer & Wittrock, 1996), well-structured and ill-defined problems (Jonassen, 1997), or simple and complex

problems (Frensch & Funke, 1995). A typical mathematics problem for example is a well-defined, routine, or well-structured problem resulting in a single answer.

This study uses ill-defined problems where problem elements are not all known or well-defined (Chi & Glaser, 1985; Wood, 1983), there is no single best solution or typical solution (Kitchner, 1983; Spiro, 1987), and multiple representations of the problem scenarios will most likely emerge within a specific context (Jonassen, 1997). Since ill-defined problems do not have routine solution steps, it is difficult to assess ill-defined problem-solving skills. Problem-solving studies suggest that domain-specific strategies (e.g., Sternberg, 1985) and domain expert solution schema (e.g., Voss & Post, 1988) should be used to assess problem-solving skills. According to previous studies on assessing ill-defined problem-solving skills (e.g., Allison, Morfitt, & Demaerschalk, 1996; Anderson, 1982; Andrews & Halford, 2002; Baker & Schacter, 1996; Dabbagh, Jonassen, Yueh, & Samouilova, 2000; Ericsson & Smith, 1991; Herl, O'Neil, Chung, & Schacter, 1999; Jacobson, 2000; Jonassen, 1997; Spiro, 1987), a new assessment methodology for ill-defined problems is required.

Causal influence diagramming (Cunningham & Stewart, 2002; Dörner, 1987; Ifenthaler & Seel, 2005; Seel, 2003; Spector, Christensen, Siotine, & McCormack, 2001) and concept mapping (Herl et al., 1999; Liu & Hichey, 1996; McClure, Sonak, & Suen, 1999; Ruiz-Primo & Shavelson, 1997; Taricani & Clariana, 2006) were identified as promising techniques to support assessment of learning in complex domains. The DEEP methodology solicits expert and novice created annotated concept maps representing conceptual representations of given ill-defined problems as a technique to identify predictable patterns of understanding (Markham, Mintzes, & Jones, 1994; Spector & Koszalka, 2004). It was designed to measure learning progress after learners participate in instructional interventions (or domain practice events) designed to improve understanding of complex and ill-defined problems. Thus, it was developed to provide an efficient and reliable method of assessing learning in problem-centered learning environments in complex domains that typically involve analyzing many interrelated factors to resolve problems that do not have standard or single solutions (Spector & Koszalka, 2004).

16.1.3 Assessing Progress in Complex Problem Solving in DEEP

The fundamental assumption of the DEEP methodology is that it is possible to predict performance and assess relative level of expertise by examining a problem conceptualization that suggests likely solution alternatives for specific complex problems (Spector et al., 2005). Annotated concept maps can be analyzed to measure differences between expert and novice subjects in their representations conceptualizing an ill-structure problem (Goldsmith, Johnson, & Acton, 1991; Herl et al., 1999). The DEEP methodology requires the collection of responses from novices and experts. Patterns in expert participants' responses are identified and used as a baseline target for assessing novice individuals who are undergoing

instruction. DEEP can then provide the basis for assessing learning outcomes in complex, ill-defined, problem-solving domains as well as provide guidance for designing and sequencing learning and task activities appropriate to advance an individual learner's progress (Spector & Koszalka, 2004).

Learning assessment in DEEP involves providing participants with an ill-defined problem and eliciting their thoughts on how they will approach the development of a solution, a process called conceptualizing the problem space. When conceptualizing problems, learners are asked to: (a) identify and briefly describe key factors influencing the situation; (b) identify and describe how these factors are interconnected; (c) indicate additional information that would be required to resolve the problem situation; and (d) identify assumptions made in responding to the problem. The participants are not asked to solve the problem; they are asked to conceptualize the problem. The first two factors constitute an annotated concept map (problem conceptualization). The third and fourth factors are reflective in nature and help clarify how the respondent is thinking about the problem situation. The DEEP methodology, and the software developed to implement the methodology, concentrates on efficiently capturing each of these factors for both experts and novices.

16.2 Methods

16.2.1 Research Design and Questions

This study was conducted during a year-long project investigating how medical experts (highly experienced) and novices (less experienced) conceptualized ill-defined problems in medical diagnosis. A medical diagnosis expert was engaged to develop several complex medical diagnostic problems and prepare sample explanations of a problem-solving process for each scenario. These explanations were solicited in the form of annotated concept maps conceptualizing various aspects of the example problems. Study participants then responded to two representative problem scenarios developed by the domain expert. After some initial training in the concept mapping method and the software used in this project, participants were provided each of these medical diagnosis problems, one at a time, and asked to create an annotated concept map that conceptualized how they thought about the problem. They created their concept maps using an automated tool that prompted for the description of the nodes and links between the nodes. The concept maps, associated descriptions, and list of assumptions created by each participant were the primary data analyzed. The primary research questions included:

1. Do expert participants exhibit recognizable patterns of problem conceptualizations in response to complex problem scenarios?
2. Are the problem representations of experts recognizably different than those of novices?

There were two associated hypotheses linked to these questions. First, experts would exhibit recognizable patterns among themselves of thinking and problem solving when confronted with complex, challenging, and ill-defined problems. Second, novices would exhibit noticeably different problem-solving patterns from the experts.

16.2.2 Research Methodology

Board-certified family physicians in active clinical practice (experts) were situated in a computer lab, trained in the DEEP methodology and tools, and asked to respond to two separate medical diagnosis scenarios. Their responses were analyzed with regard to salient features in anticipation of establishing a basis of comparison with novice responses. Then, first and second year medical students (novices) were trained and asked to respond to the same problem scenarios. Their responses were analyzed and compared to those of the experts with regard to a number of salient features that fell into three categories: surface features (e.g., number of nodes and links; average number of words to describe each node and link), structural features (e.g., key node clusters; connectedness of the concept maps measured by the percentage of orphan nodes lacking connection back to other nodes), and semantic features (e.g., did experts and novices say the same kinds of things about similar nodes).

16.2.3 Problem Scenarios

Two common problem scenarios for family medicine physicians were used for the medical diagnosis test groups. Each was a written scenario less than 500 words, sufficiently rich so that straightforward responses were not obvious, and not so complex that the participant could not develop a sense of the problem space and document their thoughts within 3 h. Participants were asked to indicate factors (facts, concepts, variables, etc.) that may be relevant to a solution and provide a short description of each item along with a brief explanation of why and how each is relevant. Subjects were also asked to depict how these factors were interrelated. This information was assumed to represent how a respondent conceptualized the problem space and approached the problem-solving process (see Fig. 16.1).

16.2.4 Participants

Since this chapter reports partial findings of the "Dynamic Enhanced Evaluation of Problem Solving" project (Spector & Koszalka, 2004), it should be noted that the primary data collection phase involved 16 experts and 49 novices in three domains: medical diagnosis, engineering design, and environmental biology. This chapter

Medical diagnosis scenario 1 Intermittent chest pain	Medical diagnosis scenario 2 Acute knee pain
Mrs. B is a 45-year-old female who comes to your office complaining of chest pain. She first noticed the pain several days ago, and initially attributed it to heartburn because it went away after taking TUMS and lying down. She notes that the pain is both "sharp" and "like pressure." It is located just under her left breast, and does not radiate anywhere. She does not get short of breath with the pain, and has become nauseous with the pain only once. The pain comes and goes, and she does not have it right now. She is unsure if the pain gets worse on exercise, "it just seems to be random." She has no history of hypertension or diabetes mellitus, but has elevated cholesterol (TC-256, HDL – 45, Trig – 168, LDL – 178). She does not smoke, and walks once or twice a week. On physical examination, she is alert and in no distress. Her neck exam reveals no jugular venous distension, carotid artery pulsations are 2+ and no bruits are auscultated. Her chest is clear to auscultation in all lung fields. Her heart exam reveals a normal rate, normal S1 and S2, and no gallops or murmurs. Her point of maximal impulse is located at the 5th intercostal space in the midclavicular line. Her abdominal examination reveals no tenderness, masses, or hepatosplenomegaly. Her extremities are without edema, and her distal pulses are 2+ and equal bilaterally.	Ms. M is a 45-year-old female with a history of hypertension, fibromyalgia and chronic neck and back pain, who presents complaining of 2 months of right knee pain and 2 days of acute worsening of the knee pain after squatting. She denies any history of trauma or twisting injury to the knee. She has not had any problems with that knee before. The knee is swollen and she has a hard time bending it or bearing weight on it. She denies any fevers or chills. She does not smoke or use illegal substances. She does not have any wounds, any vaginal discharge or bleeding, or other swollen joints. On examination, she has a markedly swollen right knee compared with the left. There is no distinct erythema, but the right knee feels very slightly warmer than the other. She is very tender to palpation in the joint line, and even somewhat tender on palpation of the patella. She has no other bony tenderness. She cannot tolerate any further examination; especially range of motion beyond about 20 degrees of flexion/extension. The distal neurovascular examination of the right leg is equal to the left leg, and is normal.

Fig. 16.1 Medical diagnosis scenario 1 and scenario 2

reports on findings in the medical domain only. In the medical domain, there were 6 experts and 14 novices. The experts in medical domain were experienced family physicians who also taught medical students at an allopathic medical university in northeastern part of the United States. Each physician was board certified and had over 5 years experience in clinical practice. Physicians were solicited by the project's medical domain content expert through a distribution of emails to the faculty in the Department of Family Medicine at the university.

The novices were medical students, primarily in their first two preclinical years who had had some training in medical interviewing, physical examination, and basic anatomy, physiology and pathophysiology. Volunteer novices were solicited through a mass emailing. Both the experts and novices were compensated for their participation, and exemption from review by the institutional review board for the protection of human subjects was obtained.

Table 16.1 Participant demographic information

	Expert* Participants *n* (%)	Novice* Participants *n* (%)
Gender		
Male	5 (83.3)	7 (50.0)
Female	1 (16.7)	7 (50.0)
Age		
Under 30	0 (0.0)	13 (92.9)
Over 30	6 (100.0)	1 (7.1)
Race		
White	6 (100.0)	13 (92.9)
African-American	0 (0.0)	1 (7.1)
Work experience (yrs)		
$0 \leq 1$	0 (0.0)	14 (100.0)
$1 \leq 5$	0 (0.0)	0 (0.0)
$5 \leq 10$	2 (33.3)	0 (0.0)
Over 10	4 (66.7)	0 (0.0)
Work position		
Medical student	0 (0.0)	14 (100.0)
Physician/clinician	6 (100)	0 (0.0)

* Experts $n = 6$; ** Novices $n = 14$.

Table 16.1 represents the basic demographic information collected.

16.2.5 Data Collection Process

The Web-based DEEP tool, depicted in Fig. 16.2, prompted participant responses to demographic questions and the problem scenarios. All data were collected using this tool. The data collection process included:

1. Registration of respondents in the DEEP Problem Conceptualization Tool
2. Administration of a background survey (demographics and perceptions)
3. Presentation of the research project, expectations and an explanation of the data collection process
4. Training and practice with DEEP Problem Conceptualization Tool using non-medical diagnosis scenario
5. Presentation of the first problem scenario and collection of the responses (3–4 h)
6. Presentation of the second problem scenario and collection the responses (2–3 h)

In order to determine relevant characteristics of expert and novice participants, the background survey included items intended to identify the participants' perception of their preparation and expertise in the specialized domain, inclination to

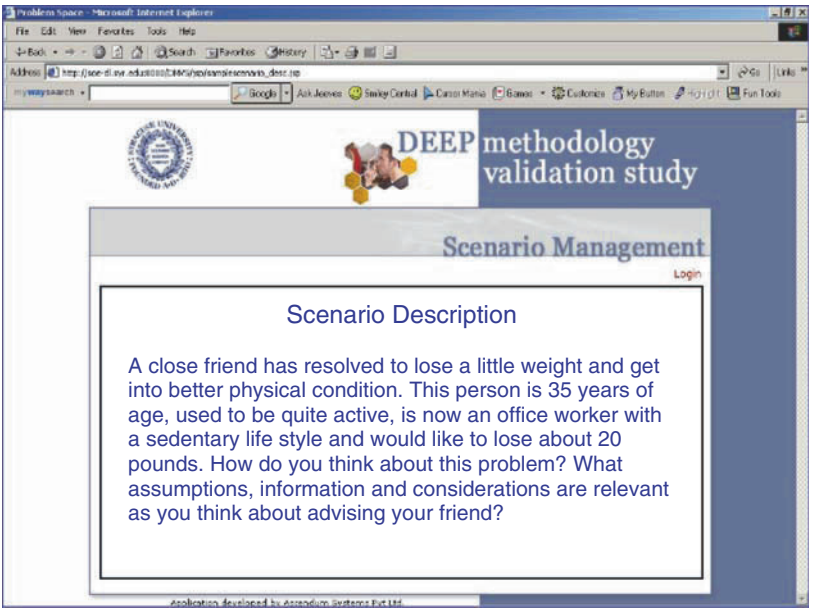


Fig. 16.2 The DEEP tool with a protocol process training problem scenario

engage in deliberate practice, and learning to improve their performance (Ericsson, 2001) as well as the length of their work experience (Spector & Koszalka, 2004).

For each problem scenario, respondents used the DEEP tool to record the concepts, variables, and relationships as a representation of how they were thinking about the problem (see Fig. 16.3). Respondents then recorded their assumptions along with information and issues that they considered relevant in actually developing a solution. Respondents were not asked to solve the problems; rather, the emphasis was on representing how they thought about the problem situations in terms of key factors and the relationships among them.

16.2.6 Data Analysis

Three levels of analysis were performed on the data: surface (level 1), structural (level 2), and semantic (level 3). At each level of analysis, differences and similarities were assessed between expert and novice respondents to identify patterns in problem conceptualization. Identifying such patterns using the DEEP methodology aids in the identification of learning progress and in the design of future instructional interventions (Spector & Koszalka, 2004).

The surface analysis (level 1) simply involved counting the number of nodes and links identified as relevant to the problem, assessing the density (number of words) of annotations of those nodes and links, and interpreting the general numeric

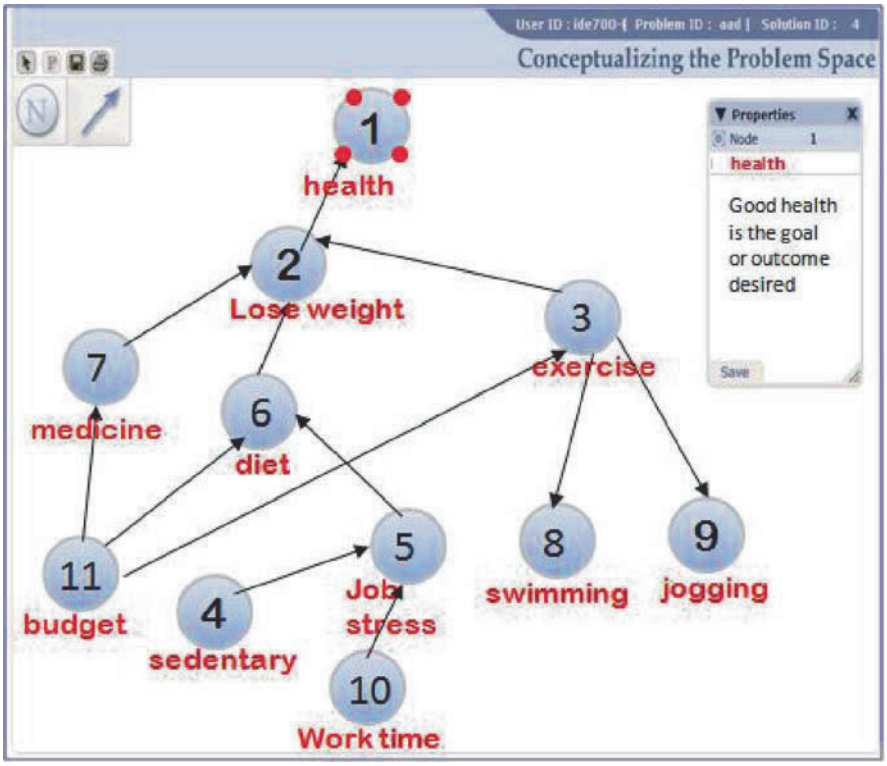


Fig. 16.3 A sample response from the protocol training scenario. *Black dots* indicate selected node

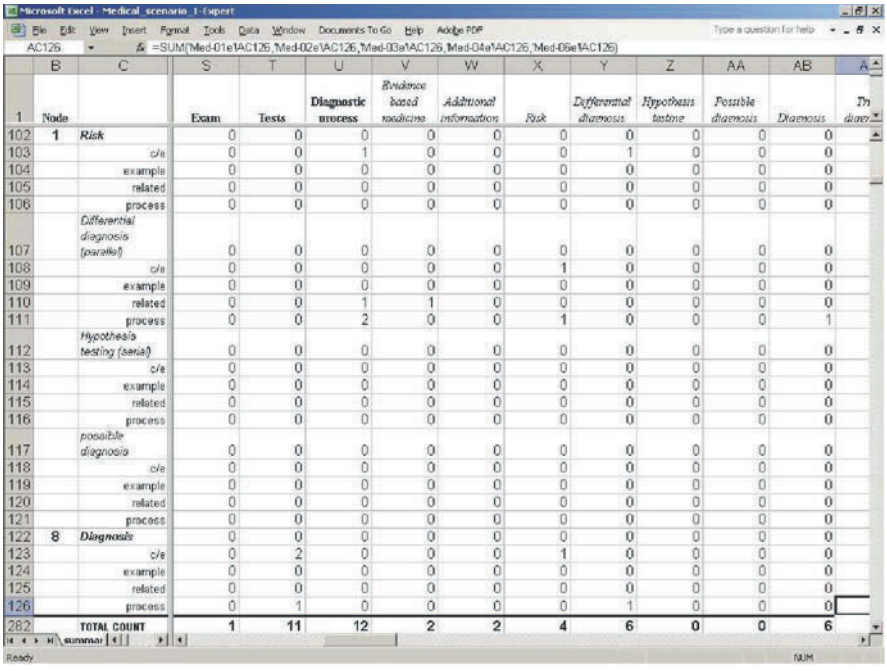
patterns of the representation. The responses were analyzed in terms of number of nodes, one-way links, two-way links, words per annotated node or link, and the first several nodes identified. The relationships between the nodes were defined and coded as the following:

- Cause–Effect (c/e): Results from, results in, as a result of, causes, influences, if-then, caused by, due to fact that, leads to, contributes to, plays a role, brings about, impacts.
- Example (e): A kind of, a part of, an example of, illustrated by.
- Correlation (c): Related factor, parallel relations, correspondence.
- Process (p): Next step, previous step, sequence.

The structural analysis (level 2) involved identifying the similarities and differences in responses in terms of the relationships among the various nodes. This analysis depended on the ability to say that nodes in different responses represented the same or similar ideas, which required a semantic analysis of the annotation

(words used in labels and descriptions). The semantic analysis (level 3) involved comparing two representations with regard to the number and percentage of the same or similar nodes and links (considering semantic analysis), and whether similar nodes and node clusters appeared in responses. Given the complexity of coding and analyzing such data, the following key steps were taken to create a reasonable data analysis procedure to interpret level 2 and 3 data (Spector & Koszalka, 2004):

1. Identify the key protocols for coding and analyzing each response based on the domain expert's comments and responses to the problem scenarios. These protocols included coding nodes and major types of links between nodes into cause–effect, correlation, example, and process schemes
2. Code a sample of three expert and five novice responses using the protocols to test the coding scheme
3. Conduct an expert review of the initial coding to determine the accuracy of interpretation of technical (medical diagnosis) concepts, terms, and explanations
4. Modify protocols and codes as required, complete coding of all responses and enter codes into protocol matrix spreadsheet (see Fig. 16.4)
5. Formulate similarity measures and comparative assessments based on the coded responses in spreadsheet



	B	C	S	T	U	V	W	X	Y	Z	AA	AB	A
1	Node		Exam	Tests	Diagnostic process	Evidence based medicine	Additional information	Risk	Differential diagnosis	Hypothesis testing	Possible diagnosis	Diagnosis	Tr diagn
102	1	Risk	0	0	0	0	0	0	0	0	0	0	
103		c/e	0	0	1	0	0	0	1	0	0	0	
104		example	0	0	0	0	0	0	0	0	0	0	
105		related	0	0	0	0	0	0	0	0	0	0	
106		process	0	0	0	0	0	0	0	0	0	0	
107		Differential diagnosis (parallel)	0	0	0	0	0	0	0	0	0	0	
108		c/e	0	0	0	0	0	1	0	0	0	0	
109		example	0	0	0	0	0	0	0	0	0	0	
110		related	0	0	1	1	0	0	0	0	0	0	
111		process	0	0	2	0	0	1	0	0	0	1	
112		Hypothesis testing (serial)	0	0	0	0	0	0	0	0	0	0	
113		c/e	0	0	0	0	0	0	0	0	0	0	
114		example	0	0	0	0	0	0	0	0	0	0	
115		related	0	0	0	0	0	0	0	0	0	0	
116		process	0	0	0	0	0	0	0	0	0	0	
117		possible diagnosis	0	0	0	0	0	0	0	0	0	0	
118		c/e	0	0	0	0	0	0	0	0	0	0	
119		example	0	0	0	0	0	0	0	0	0	0	
120		related	0	0	0	0	0	0	0	0	0	0	
121		process	0	0	0	0	0	0	0	0	0	0	
122	8	Diagnosis	0	0	0	0	0	0	0	0	0	0	
123		c/e	0	2	0	0	0	1	0	0	0	0	
124		example	0	0	0	0	0	0	0	0	0	0	
125		related	0	0	0	0	0	0	0	0	0	0	
126		process	0	1	0	0	0	0	1	0	0	0	
282		TOTAL COUNT	1	11	12	2	2	4	6	0	0	6	

Fig. 16.4 Excerpt of coding protocol summarizing expert responses to scenario 1

6. Complete coding for all (recoding initial responses) respondents and enter into the protocol data matrix
7. Conduct an inter-rater reliability check on coding, reconcile differences between independent coders
8. Prepare statistical summaries from coded data for each scenario with summaries for individuals and groups

It should be noted that a reliability checking process was conducted by raters to check the coders' results ensuring that codes were accurate given the complexity of the categories and the types of links involved. The initial coding was completed by two independent coders and the inter-rater reliability was checked by two additional individual raters. Both raters found 90% reliability in coding between the two coders for both scenario 1 and scenario 2. Any discrepancies noted in coding were discussed with the coders and resolved. Most of the disagreements involved the type of relationship between nodes. Table 16.2 presents the resulting rate of reliability among the two individual raters.

Table 16.2 Inter-rater reliability: percentage of agreement between codes for each scenario based on rater review

	Percentage of code agreement between 2 coders	
Rating event	Independent rater 1	Independent rater 2
Scenario 1 (2 coders)	90%	90%
Scenario 2 (2 coders)	90%	90%

16.3 Results and Analysis

Data were first organized and analyzed based on the three levels of analysis. An interesting finding is reflected in the Table 16.3 level one analysis. In the level 1 analysis of both scenarios, on average, expert medical practitioners provided fewer nodes than novices for scenario 1 and about the same number of nodes for scenario 2. However, experts provided denser elaborations (more words per node and link) and had fewer two-way links in both cases.

The level 2 structural analysis suggested that the types of links made by the experts and novices tended to be different. The highest percentage of links made by both experts and novices in scenario 1 were cause and effect. In scenario 2 the experts had almost an equal number of cause and effect and process links whereas novices had approximately equal numbers of cause and effect and correlation links. The percentage of process links was much higher for the experts than novices in both scenarios. The novices had higher percentage of links that were correlations as compared to the experts in both scenarios (see Tables 16.4 and 16.5).

Table 16.3 Surface (level 1) analysis by scenario

Scenario	Group	Avg. # nodes	Avg. one-way links	Avg. two-way links	Avg. words-node	Avg. words-node-name	Avg. words-link
Scenario 1	Novices	13.07	16.00	6.86	15.86	1.83	11.57
	Experts	10.06	17.60	4.00	26.60	2.24	21.20
Scenario 2	Novices	12.21	16.36	5.00	21.50	2.06	15.46
	Experts	12.25	15.50	1.50	31.75	2.35	14.63

Table 16.4 Medical scenario 1 summary data for used links between nodes (level 2 analysis)

Experts ($n = 6$)			Novices ($n = 14$)		
Type of link	No. of links	Percentage	Type of link	No. of links	Percentage
Cause/effect	73	68.9%	Cause/effect	185	58.2%
Example	0	0.0%	Example	24	7.5%
Correlation	8	7.5%	Correlation	84	26.4%
Process	25	23.6%	Process	25	7.9%
Total links	**106**	**100%**	**Total links**	**318**	**100%**

Table 16.5 Medical scenario 2 summary data for used links between nodes (level 2 analysis)

Experts ($n = 6$)			Novices ($n = 14$)		
Type of link	No. of links	Percentage	Type of link	No. of links	Percentage
Cause/effect	28	41.2%	Cause/effect	137	46.6%
Example	0	0.0%	Example	27	9.2%
Correlation	10	14.7%	Correlation	123	41.8%
Process	30	44.1%	Process	7	2.4%
Total links	**68**	**100%**	**Total links**	**294**	**100%**

During the level 3 analysis it was possible to identify factors around which the links tended to cluster in the various responses. These clusters were analyzed with the results from the level 2 analysis to ensure that clusters formed from nodes and links analyzed across individuals' representations did indeed have similar meaning.

The top factors identified across all experts and all novices are shown in the following tables for each scenario: Tables 16.6 and 16.8 show the top five nodes from data belonging to expert participants for scenario 1 and scenario 2, respectively. Tables 16.7 and 16.9 show the top five nodes from data belonging to novice participants for scenario 1 and scenario 2, respectively

Table 16.6 Medical novices – scenario 1 ($n = 14$; total links = 318) (level 3 analysis)

	Nodes	Links	From/to Node/Node	Percentage
1	History of present illness	67	31/36	21.07%
2	Hypothesis testing	55	17/38	17.30%
3	Chief complaint	49	14/35	15.41%
4	Patient's background/social history	45	27/18	14.15%
5	Stress	44	23/21	13.84%

Table 16.7 Medical experts – scenario 1 ($n = 6$; total links = 106) (level 3 analysis)

	Nodes	Links	From/to Node/Node	Percentage
1	Tests	19	8/11	17.93%
2	Differential diagnosis	16	10/6	15.09%
3	Diagnosis	14	8/6	13.21%
4	History	10	8/2	9.43%
4	Post-visit additional information	10	3/7	9.43%
4	Clinical knowledge	10	7/3	9.43%

Table 16.8 Medical novices – scenario 2 ($n = 14$; total links = 294) (level 3 analysis)

	Nodes	Links	From/to Node/Node	Percentage
1	History of present illness	107	50/57	36.40%
2	Past medical history	83	55/28	28.23%
3	Hypothesis testing	66	10/56	22.45%
4	Chief complaint	33	12/21	11.22%
5	Patient's concern/preference	31	18/13	10.54%
5	Physical exam	31	21/10	10.54%

Table 16.9 Medical experts – scenario 2 ($n = 6$; total links = 68) (level 3 analysis)

	Nodes	Links	From/to Node/Node	Percentage
1	Differential diagnosis	21	10/11	30.88%
2	Tests	10	4/6	14.71%
2	Referral	10	3/7	14.71%
2	Contingency management	10	5/5	14.71%
3	Hypothesis testing	9	6/3	13.24%

16.3.1 Comparisons of Expert and Novice Responses

In novices' responses to scenario 1, the nodes that have the most links are history of present illness, hypothesis testing, chief complaint, social history, and stress. In scenario 2 the nodes that have the most links are history of present illness, past medical history, hypothesis testing, chief complaint, patient's concerns, and physical exam.

The patterns noted in history, testing, and complaints could be explained in several ways. First, novices tend to represent the information presented to them from the scenario in their diagrams, so they used several nodes to illustrate the history of present illness of the patient and chief complaint. Second, as the literature suggests, novices tend to use hypothesis testing when making diagnostic decisions (Coderra et al., 2003; Wiener, 1996). In their responses, several nodes for hypothesis testing were noted. Third, in this particular scenario, many novices describe social history, diet, and stress, while the experts did not mention these factors. Thus, in response to one of the research questions, there is a recognizable difference in problem representation between experts and novices.

When implying a cause–effect relationship, respondents tended to make links from a factor that had an influence on another factor that they believed showed an effect. In novices' responses to scenario 1, for example, level 3 analysis suggested clusters were distinguished among nodes such as stress, diet, and social history to chief complaint and history of present illness (Fig. 16.5).

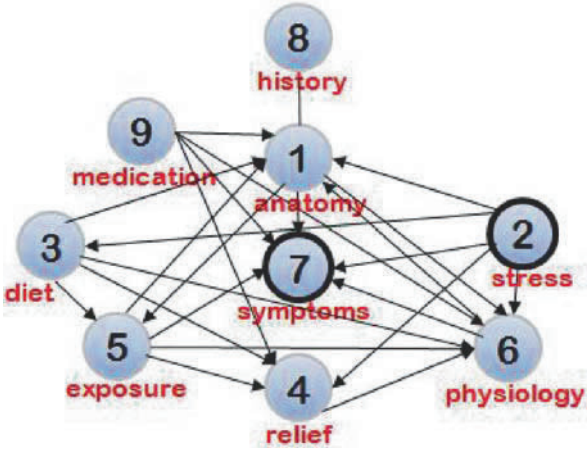


Fig. 16.5 Sample clusters for novices for medical scenario 1. *Highlighted nodes* are the most frequently chosen among respondents

In novices' responses to scenario 2, clusters were observed from nodes such as stress, diet, history of present illness, and hypothesis testing, which are usually various diagnostic possibilities (see Fig. 16.6).

As before, highlighted nodes in Figs. 16.5 and 16.6 reflect clusters that appeared most frequently among the respondents. In an individual diagram, a highlighted node may not appear to be a cluster but it does appear to be a cluster when more

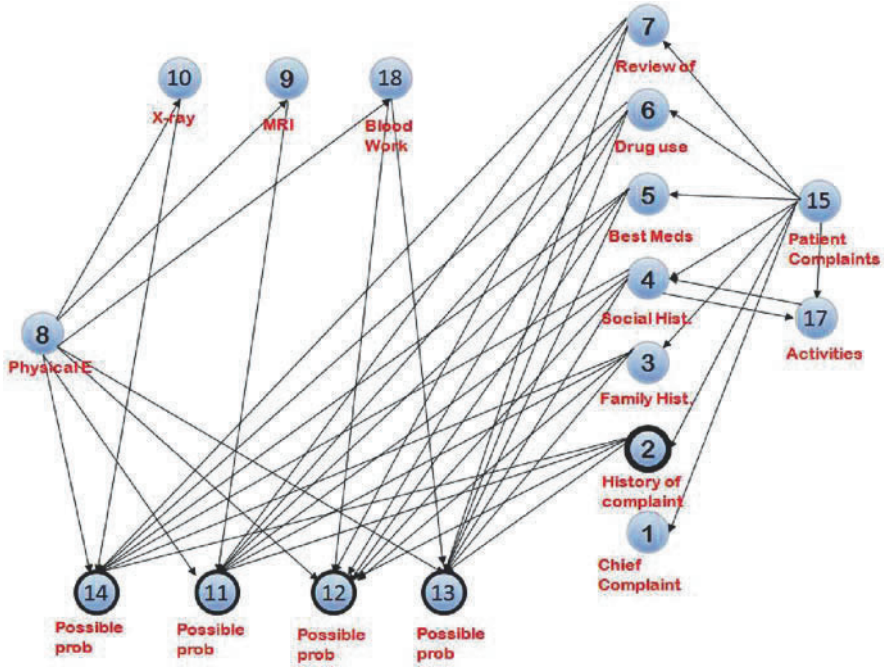


Fig. 16.6 Sample clusters for novices on medical scenario 2. *Highlighted nodes* are the most frequently chosen by respondents

responses are considered. Likewise, what appears to be a cluster in a particular figure may not show up as a cluster across a large number of respondents and is not highlighted in these cluster diagrams.

For scenario 1, the experts created the most links for these nodes: tests, differential diagnostic and diagnosis, history, post-visit additional information, and clinical knowledge, demonstrating a recognizable pattern of problem conceptualization, as predicted.

For scenario 2 expert responses were similar, clustering around differential diagnosis, tests, referrals, contingency management, and hypothesis testing (see Fig. 16.7). In contrast to novices, experts were more likely to be engaged in parallel thinking while making diagnostic decisions; this could be the reason they have more nodes and links centered on differential diagnosis and diagnosis.

Nodes most commonly included by experts, rarely present in novice representations, included tests, additional information, and diagnostic processes. One explanation for this occurrence is that novices lack the clinical experiences (which provide context and assessment of expected frequencies of illness and clinical variation) to think about which tests and what additional information is needed to make a proper diagnosis. In other words, novices did not have sufficient knowledge or experience to develop a plan that could rule out many diagnostic possibilities and arrive at reasonable treatment conclusions.

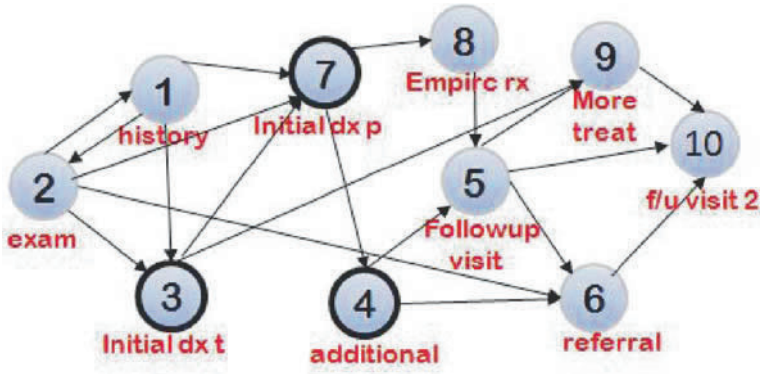


Fig. 16.7 Sample clusters for experts on medical scenario 1. *Highlighted nodes* are the most frequently chosen by respondents

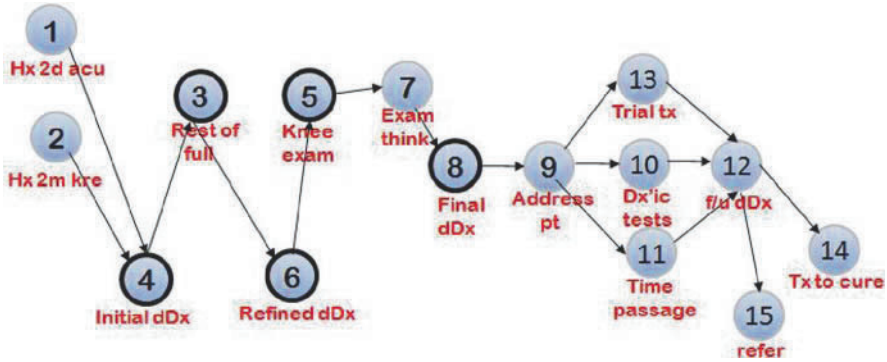


Fig. 16.8 Sample clusters for experts on medical scenario 2. *Highlighted nodes* are the most frequently chosen by respondents

Experts, on the other hand, tended to regard problem solving as an iterative process, in which they seek information, make judgments, seek further information to confirm their judgments, and make further judgments (see Fig. 16.8). These are additional indications of differences between expert and novice problem conceptualizations.

It is worthy to note that in scenario 2 hypothesis testing was a common node in which there was prevalent clustering by experts as well as novices. This might be due to the fact that many respondents felt that scenario 2 was easier than scenario 1, and more detailed information was provided in scenario 2.

Therefore, novice and experts were able to make certain hypotheses based on the comparatively rich information. Further analyses should be conducted in future studies to compare the hypotheses made by the experts and the novices to identify specific differences.

Another difference between experts and novices, especially in scenario 2, is that many experts noted the use of referrals, follow-up, and contingency management,

while the novices tended to neglect these factors. This is another indication that experts have similar patterns in conceptualizing diagnosis situations while the differences noted in the novice representations could be attributed to their predictable lack of experience in providing complete management of patients, which is emphasized mostly in their postgraduate training.

Table 16.10 summarize major differences seen in the analysis and interpretation of expert and novice conceptualizations of the scenarios.

Table 16.10 Summary of differences among expert and novice for three levels of analysis

Surface analysis (level 1)		Structural/semantic analysis (levels 2 and 3)	
Novice	Expert	Novice	Expert
Slightly more nodes (Sc1) Same number of nodes (Sc2) Less dense node description Same number of one-way links Many two-way links	Slightly fewer nodes (Sc1) Same number of nodes (Sc2) More dense node description Same number of one-way links Few two-way links	Equal percent of cause/effect and correlation links Higher percent correlation links than experts Top Clusters (Sc1): history of present illness, hypothesis testing, chief complaint Top Clusters (Sc2): history of present illness, past med history, hypothesis testing	Equal percent of cause/effect and process links Higher percent process links than novices Top Clusters (Sc1): tests, differential diagnosis, diagnosis Top Clusters (Sc2): differential diagnosis, tests, referral

Notes: Sc1 = Scenario 1; Sc2 = Scenario 2.

16.4 Conclusion

Based on the study findings, expert participants exhibited recognizable patterns of problem conceptualizations by identifying similar key factors and describing relationships among those factors in similar ways. However, novice participants conceptualized the problem space in many different ways, all of which were recognizably different from the expert pattern. This research study demonstrated the potential of the DEEP methodology as a learning assessment tool for educational and practice contexts. Comparing the concept maps (problem conceptualizations) of expert and novice subjects provided the basis for inferences about the structural complexity of the problem-solvers' knowledge base as well as a basis for understanding how the problem-solver viewed the problem situation (Spector & Koszalka, 2004; Suzuki & Harnisch, 1995).

The DEEP findings generally showed that the concept maps of expert respondents were also more integrated and contained a greater number of cross-links between and among concepts than the maps of novice respondents (Spector & Koszalka, 2004). It also provided evidence that experts think in similar patterns when presented with the same ill-defined problem in their domain, even though they may develop entirely different solutions. Furthermore, the expert participants' nodes tended to include both information specifically from the scenario and nonstated information considered important to the problem. Novice participants tended to include mostly information that was stated in the scenario by summarizing the given information. Expert participants made inferences from problem situations based on their experience level and a deeper domain and application knowledge than less knowledgeable and novice participants (Spector & Koszalka, 2004).

The expert participants, with the same instructions provided to novice participants, also tended to use denser descriptions of nodes and links than novice participants to describe their representations. This could be because the expert participants had deeper knowledge of the key factors conceptualized in the scenarios and how these factors related to each other (Anderson, 1982; Ericsson & Smith, 1991; Ganesan & Spector, 2000). Furthermore, the expert participants tended to represent more causal relationships that narrowly clustered among key concepts central to their problem representation while novice participants tended to have wider clusters, potentially indicating uncertainty about key areas of focus in problem scenarios.

As a result, the development of DEEP methodology represents important steps forward in developing a robust and reliable assessment methodology for complex domains. This methodology can provide data that suggest progress of learning and relative level of expertise in complex domains for which standard solutions to presented ill-structured problems often do not exist. Data generated shows how individuals conceptualize complex problems, a first step in identifying potential solutions. These data can indicate the level of problem understanding, domain knowledge, and level of higher-order problem-solving ability being applied to presented problems. Such measures taken over time and compared to expert responses can provide a way to analyze learning progress toward expertise. Thus, DEEP methodology can be used by educators to assess learning progress (both comprehensiveness and efficiency) and identify, through robust evidence, ways to enhance the structure of learning activities according to learner's needs.

16.4.1 Medical Domain Issues

There are issues specifically in the medical diagnosis domain that need further consideration. An additional area of interest in the use of this methodology may be to better understand the phenomenon of premature closure in medical diagnosis (Sutherland, 2002). Premature closure is the cognitive leap to a solution, primarily

based on clinical experience or pattern recognition, with a failure to consider the alternative presentations of common illness, the typical presentations of more rare illness, or forthcoming information that may alter the diagnosis. This is a particular issue in health care environment which rewards the rapid disposition of patients and is a particular problem in the realm of medical malpractice. It may be that expert performance in these assessment situations can be too efficient – and likely should include some measure of hypothetico-deductive reasoning to avoid the trap of reliance on pattern recognition only. The appropriate dose of and balance between pattern-recognition and hypothetico-deductive reasoning is yet to be worked out. The DEEP methodology and tool may be helpful in further identifying the level and application of pattern-recognition and hypothetico-deductive reasoning that novices should be practicing in their movement toward expertise.

A possible bias in the selection of the medical experts in this study exists in that academic faculty were used who had particular interest and training in "evidence-based medicine (EBM)." Academic physicians could be more likely to have a stepwise, more analytical approach to decision making because of their teaching role than nonacademic physicians. In addition, the effect of EBM on diagnostic reasoning can be said to decrease the physicians' reliance on pattern recognition somewhat in favor of a more iterative, hypothesis testing approach based on statistical likelihoods found in the published research literature. The novice participants had only foundational training in the skills necessary to develop a diagnosis and management plan. Therefore, any differences seen between the novices and experts in their patterns of decision making may be underestimates if novices were compared to a cohort of nonacademic physician experts.

The DEEP methodology has potential in medical education for increasing the reliability and validity of assessment of complex cognitive skills such as medical diagnosis and management and holds the promise of enabling this assessment in an efficient manner. Medical diagnosis and management are currently only imperfectly assessed using multiple-choice examinations and the subjective assessments of supervising clinicians. The work done in this project, with its focus on easy-to-use online concept mapping that can be analyzed by computer can enable rapid, large-scale assessment of progress toward expert thinking in medicine.

16.4.2 Further Work on the DEEP Tools and Analysis Methodology

The DEEP methodology shows a potential to assess learning in complex domains but the methodology is still in its early stages. The methodology involved the use of an online problem conceptualization tool. The tool provides a simple way of diagramming the problem space in terms of a simple nexus of undifferentiated nodes and one-way links. The logic of the design of the tool was to provide a simple concept mapping tool that was easy for respondents to learn, allowing respondents to focus on the problem rather than the tool. A critical next step will be to balance the

examination of cognitive process with the required assurance of content knowledge that is integral to comprehensive assessment in medical education.

It should be noted that problems encountered during coding procedures included technical difficulties such as determining the correct category or subcategory for particular responses and difficulty in determining the type of link involved. Additionally, the DEEP research team was concerned about potential loss of holistic information reflected in the respondent's diagram. Because of the nature of the presentation of data and the relatively simple assessment process, participants may have failed to record their complete thoughts given the provided tools.

The data collection process was designed to be problem-oriented and to involve a minimally invasive tool to capture and represent respondents' higher-order thinking processes such that it could be validated and further developed for widespread use in large-scale educational and performance situations. The review of the literature and discussions with various participants, the open-endedness of the responses, and the actual responses that were collected led the project team to conclude that the researchers did not lose valuable information and did not misrepresent respondent intentions during coding the responses (Spector & Koszalka, 2004). Future research should be conducted to investigate to what extent this methodology produces results similar to those in a think-aloud protocol analysis – a process that does not lend itself to automation and large-scale implementation.

The DEEP has since been integrated with two other tools for assessing learning – SMD (Surface, Matching, Deep; Ifenthaler & Seel, 2005) and T-MITOCAR (Model Inspection Trace of Concepts and Relations; Pirnay-Dummer, 2007) and into a set of assessment tools called HIMATT (Highly Interactive Model-based Assessment Tools and Technology) as part of a research collaboration between the University of Freiburg and Florida State University (Pirnay-Dummer, Ifenthaler, & Spector, 2009; Shute, Jeong, Spector, Seel, & Johnson, 2009). Further studies to explore the development of mental models and expertise can be accomplished using the refined DEEP methodology soon to be available in HIMATT and the associated tools. Hopefully, these tools will help promote our understanding of how to more effectively and efficiently develop expertise in complex problem-solving domains.

References

Allison, D. J., Morfitt, G., & Demaerschalk, D. (1996). *Cognitive complexity and expertise: Relationships between external and internal measures of cognitive complexity and abstraction, and responses to a case problem*. (ERIC Document Reproduction Service No. ED 412604).

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369–406.

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*(2), 153–219.

Baker, E. L., & Schacter, J. (1996). Expert benchmarks for student academic performance: The case for gifted children. *Gifted Child Quarterly*, *40*, 61–65.

Chi, M. T. H., & Glaser, R. (1985). Problem solving ability. In R. J. Sternberg (Ed.), *Human abilities: An information processing approach* (pp. 227–257). San Francisco: W. H. Freeman & Co.

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning and diagnostic success. *Medical Education, 37*(8), 695–703.

Cunningham, A. C., & Stewart, L. M. (2002). *Systems analysis of learning theory through causal influence diagrams*. (ERIC Document Reproduction Service No. ED 475928).

Dabbagh, N. H., Jonassen, D. H., Yueh, H. P., & Samouilova, M. (2000). Assessing a problem-based learning approach to an introductory instructional design course: A case study. *Performance Improvement Quarterly, 13*(3), 60–83.

Dörner, D. (1987). On the difficulties people have in dealing with complexity. In J. Rasmussen, K. Duncker, & J. Leplat (Eds.), *New technology and human error* (pp. 97–109). Chichester, NY: John Wiley & Sons, Inc.

Dörner, D. (1996). *The logic of failure: Why things go wrong and what we can do to make them right* (R. Kimber & R. Kimber, Trans.). New York: Metropolitan Books (Original work published in 1989).

Epstein, R. (2007). Assessment in medical education. *The New England Journal of Medicine*, *356*(4) 387–396.

Ericsson, K. A. (2001). Expertise in interpreting: An expert-performance perspective. *Interpreting, 5*(2), 187–220.

Ericsson, K. A., & Simon. H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits in the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1–38). Cambridge: Cambridge University Press.

Frensch, P. A., & Funke, J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 3–26). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Funke, J. (2001). Dynamic systems as tools for analysing human judgment. *Thinking and Reasoning, 7*, 69–89.

Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). New York: Lawrence Erlbaum.

Gagné, R. M. (1985). *The conditions of learning and theory of instruction* (4th ed.). New York: Holt, Rinehart, and Winston, Inc.

Ganesan, R., & Spector, J. M. (2000). Causal influence diagrams as pedagogical and assessment tools. *Presentation at CILT 2000*, Tyson's Corner, VA, 28 October 2000.

Gogus, A., Koszalka, T. A., & Spector, J. M. (2009). Assessing conceptual representations of ill-structured problems. *Technology Instruction, Cognition and Learning (TICL), 7*(1), 1–20.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*(1), 88–96.

Grotzer, T. A., & Perkins, D. N. (2000). *A taxonomy of causal models: The conceptual leaps between models and students' reflections on them*. Paper presented at the National Association of Research in Science Teaching (NARST), New Orleans, LA.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. L., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15*(3–4), 315–333.

Huber, O. (1995). Complex problem solving as multi stage decision making. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European Perspective* (pp. 151–173). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ifenthaler, D., & Seel, N. M. (2005). The measurement of change: Learning-dependent progression of mental models. *Technology, Instruction, Cognition, and Learning*, *2*(4), 317–336.

Jacobson, M. J. (2000). Problem solving about complex systems: Difference between experts and novices. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the learning sciences* (pp. 14–21). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-defined problem-solving learning outcomes. *Educational Technology: Research and Development, 45*(1), 65–95.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research & Development, 48*(4), 63–85.

Kitchner, K. S. (1983). Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. *Human Development, 26*, 222–232.

Klein, G. A. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Liu, X., & Hichey, M. (1996). The internal consistency of a concept mapping scoring scheme and its effect on prediction validity. *International Journal of Science Education*, *18*(8), 921–937.

Markham, K. M., Mintzes, J. J., & Jones, M. G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching*, *31*(1) 91–101.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berlinert & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: Macmillan.

McClure, J., Sonak, B., & Suen, H. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, *36*, 475–492.

Norman, G. R. (1994). Cognitive differences in clinical reasoning. *Teaching and Learning in Medicine*, *6*, 114–120.

Pirnay-Dummer, P. (2007). *Model inspection trace of concepts and relations. A heuristic approach to language-oriented model assessment*. Paper presentation at the Annual Meeting of the American Education Research Association, April, 2007, Chicago, IL.

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2009). Highly integrated model assessment technology and tools. *Educational Technology Research and Development*, doi: 10.1007/s11423-009-9119-8.

Ruiz-Primo, M. A., & Shavelson, R. J. (1997). *Concept-map based assessment: On possible sources of sampling variability*. (ERIC Document Reproduction Service No. ED422 403).

Seel, N. M. (2003). Model centered learning and instruction. *Technology, Instruction, Cognition and Learning*, *1*(1), 59–85.

Shute, V. J., Jeong, A. C., Spector, J. M., Seel, N. M., & Johnson, T. E. (2009). Model based methods for assessment, learning and instruction. In M. Orey (Ed.), *Educational media and technology yearbook 2009*. Westport, CT: Greenwood Publishing.

Spector, J. M., Christensen, D. L., Siotine, A. V., & McCormack, D. (2001). Models and simulations for learning in complex domains: Using causal loop diagrams for assessment and evaluation. *Computers in Human Behavior*, *17*(5–6), 517–545.

Spector, J. M., Dennen, V. P., & Koszalka, T. A. (2005). Causal maps, mental models and assessing acquisition of expertise. *Technology, Instruction, Cognition and Learning*, *3*(1–2).

Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning in complex domains* (Technical Report No. NSF-03-542). Syracuse, NY: Syracuse University, Instructional Design Development, and Evaluation (IDD&E).

Spiro, R. J. (1987). *Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains*. (ERIC Document Reproduction Service No. ED 287155).

Sterman, J. D. (1994). Learning in and about complex systems. *System Dynamics Review*, *10*(2–3), 291–330.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sutherland, D. C. (2002). Improving medical diagnoses by understanding how they are made. *Internal Medicine Journal*, 32(5–6), 277–280.

Suzuki, K., & Harnisch, D. L. (1995). *Measuring cognitive complexity: An analysis of performance-based assessment in mathematics*. Paper presented at the 1995 Annual Meeting of the American Educational Research Association, San Francisco, April 18–22. (ERIC Document Reproduction Service No. ED 390924)

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open ended concept maps. *Educational Technology, Research, & Development, 54*(1), 61–78.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-defined problems. In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.), *The nature of expertise* (pp. 261–285). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wiener, S. L. (1996). Clinical reasoning in the health professions. *Annals of Internal Medicine, 124*(5), 537.

Wood, P. K. (1983). Inquiring systems and problem structures: Implications for cognitive development. *Human Development, 26*, 249–265.

Chapter 17
Selection of Team Interventions Based on Mental Model Sharedness Levels Measured by the Team Assessment and Diagnostic Instrument (TADI)

Tristan E. Johnson, Eric G. Sikorski, Anne Mendenhall, Mohammed Khalil, and YoungMin Lee

17.1 Introduction

Teams are an important part of workplace performance. They can play a critical role in organizational productivity (Fiore & Salas, 2004). In addition to dealing with high workload tasks, teams are often required to take on mentally complex tasks that are not possible to complete by an individual alone (Cooke, Salas, Kiekel, & Bell, 2004; Eccles & Tenenbaum, 2004; Stout, Cannon-Bowers, Salas, & Milanovich, 1999). Teams have even taken on roles in educational settings to facilitate learning difficult subject matter (Johnson, Khalil, & Spector, 2008; Lee & Johnson, 2008; Michaelsen, Knight, & Fink, 2002). There are a few efforts in educational settings that take advantage of team-focused instructional strategies to bring about learning as well as general team skills development (Sikorski, 2009). Problem-Based Learning (Barrows, 1998; Barrows, 2000; Savery & Duffy, 1995) and Team-Based Learning (Michaelsen et al., 2002) are two strategies that have shown positive effects on achievement and knowledge retention, peer interaction, and critical thinking skills.

Observing work and learning settings reveals that teams are a required component for standard tasks and assignments. While there are several benefits that can come from team efforts, there are plenty examples where teams are simply not reaching their optimal potential. Managers and instructors, as team coaches, are expected to bridge the gap between where teams are currently performing and where they are needed to be performing. However, it is difficult for most instructors or managers to coach teams on how to become a higher performing team. There are a number of issues that made it difficult to effectively and efficiently support team development.

Teams are complex and dynamic systems made up of many components making it necessary to consider the interrelationship of various team characteristics and features when looking to improve team performance. Various consulting groups

T.E. Johnson (✉)
Florida State University, Tallahassee, FL, USA
e-mail: tejohnson@fsu.edu

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0_17,

offer team training where they focus on numerous topics such as team characteristics, communication, listening, roles, trust, shared leadership, team leaders, style diversity, self-assessment, barriers, dealing with issues, untangling a team, common mistakes, managing stress, listening skills, establishing an encouraging environment, team problem solving, social skills, etc. (Parker, 2008; Huszczo, 1996; Hinsberg, 1996; Gregory, 1999; Robbins & Finley, 1995; Walker & Harris, 1995; Hitchcock & Willard, 1995; Leonard & Swaps, 1999; Peterson and Behfar, 2003).

Knowing the exact parts of the team system to target for team development is a difficult task. Since there are many parts to a team, this makes it very difficult to know what and how to help a team improve their performance. Most managers and instructors are experts in their respective fields, but most of them are not likely experts in team dynamics and team improvement strategies. Due to the complex nature of teams, it is unlikely that they are trained to identify specific team needs.

Even if one knows the exact needs and parts of the team system to focus on to help the team make improvements, knowing how to bring about a change is another challenge. However, team coaches have the best chance for affecting team improvements if they have accurately identified the team weaknesses and have ideas on how to think about team intervention strategies. While team intervention can be helpful for a team, without proper diagnostic data, there is a risk of selecting a team intervention that is not aligned or only partially aligned with team needs. This process and the ensuing effects can be rather random unless team coaches have appropriate techniques to capture team thinking. With proper assessment data, one can reasonably determine team needs and select appropriate team interventions, thereby deliberately coaching a team based on the team's actual needs to affect team improvements and have a strong impact on team performance. Team coaches need similarity measures in order to determine how much a team agrees on the amount of team skills. Often overlooked, the measure of team knowledge and perception similarity is critical to team development simply because if a team member does not recognize that improvements are needed, it will be difficult to have that team member meaningfully engage in any team performance improvement activity.

The purpose of this chapter is to describe a method to generally assess teams using the Team Assessment and Diagnostic Instrument (TADI) (Johnson et al., 2007) to provide guidelines to make sense of the assessment measures thereby facilitating the diagnosis of team weaknesses and strengths. While in itself this tool is simple, it has the ability to quickly assess teams and can easily be coupled with other computer-based diagnostic tools (Shute et al., in press) that are more sophisticated in their ability to measure and diagnosis.

This tool has been empirically validated (see Section 17.3) and shown to be a strong indicator of a team's shared cognition. Further, this chapter covers general guidelines on how to interpret the team assessment data in order to know if a team intervention is needed. Team coaches can use the assessment data to either continue to monitor a team, gather more data, or recommend a team intervention that focuses on team similarity building, domain knowledge, and/or skill development. Additionally, several team intervention strategies are described that are mapped to the team assessment measures.

17.2 Team Cognition

There has been substantial research over the past few decades in varying fields that have been studying effective team performance (Guzzo & Shea, 1992; Levine & Moreland, 1990). The main focus of this research has led to improved understanding about team's behaviors. This line of research established indicators and predictors of effective team performance. The key indicator that has emerged from this research is the notion of team cognition, which is the degree that members of a team share similar conceptualizations of problems and approaches to solutions (Salas & Cannon-Bowers, 2000). Team cognition is a strong measure of effective team performance based on the evidence that teams with members who think in similar ways about difficult problems are likely to work effectively together (Cannon-Bowers & Salas, 1998b; Guzzo & Salas, 1995; Hackman, 1990). In short, team cognition can be used as an assessment construct that describes the level of a team's ability to think and act as a unit (Salas & Fiore, 2005).

When measuring team cognition, there are numerous methods and techniques that can be used to capture this construct. The predominate technique to capture a team's cognition is to measure and determine the similarity in team member's individual mental models also known as shared mental models. Shared mental models (SMM) have been highlighted as key factors in encouraging positive team performance (Cannon & Edmondson, 2001). A team's SMM consists of the necessary knowledge that is required for effective performance, the skills and behaviors that are necessary to perform the task effectively, along with the appropriate attitudes that promote effective team performance (Cannon-Bowers, Tannenbaum et al., 1995). SMM are comprised of two types of shared knowledge (e.g., task-related and team-related). Task-related knowledge is specific knowledge that is related to the task that the team is performing or learning. Team-related knowledge is knowledge that is not task specific, but relates to the team in general. While knowing the level of task-related knowledge similarity among teammates is important, there is a great deal of methodological constraints in assessing the similarity of task-related knowledge.

In order to assess task-knowledge, an instrument has to be developed that is related to the specific team task. Development of a task-knowledge assessment measure requires a task analysis as well as running a pilot study and then instrument validation must take place. While this can be done (Lee & Johnson, 2008), there is a significant amount of resources tied up into team assessment instrument development. For research purposes, this is expected, but for team assessments in the workplace and learning environments, this is simply not reasonable.

To meet the constraints of conducing decision-making research in the workplace, Johnson and his colleagues (2007) developed and validated an SMM instrument that focuses on the measurement of team-related knowledge. Unlike the challenges with measuring task-related knowledge, team-related knowledge does not consider a specific task but looks at general team knowledge – knowledge that applies to all teamwork independent of the team task. As such, a team-related SMM instrument is reusable in various domains and, therefore, it is possible to use this instrument in the workplace without the need for redevelopment and re-testing. Using a validated

team-related knowledge instrument that measures SMM is a reasonable and practical technique for team assessment and diagnosis for both performance and learning teams (Johnson et al., 2007).

The mechanism that SMM influences team performance is by decreasing the communication demands during team performance, thereby allowing teams members to allocate mental energy to the task at hand (Langan-Fox, Anglim, & Wilson, 2004). By using an SMM assessment instrument, team coaches can efficiently and effectively capture an indirect measure of team productivity. This measure then can be used to support decision-making related to team performance improvement strategies. We first provide a description of the TADI instrument before explaining the process of determining if an intervention is needed and what intervention strategy should be used.

17.3 Team Assessment and Diagnostic Instrument

The Team Assessment and Diagnostic Instrument (TADI) was specifically created to measure the degree of team-related knowledge in order to determine team-related knowledge sharedness. By knowing the level of SMM in a team, specific decisions can be made regarding the potential productivity of a team. In addition, the SMM can serve as an indicator of potential team success. The TADI was developed specifically to measure general types of knowledge and not specific task-knowledge. The team shared knowledge construct refers to general knowledge but also includes general task (not specific task) knowledge such as "My team likes to do various team tasks." As such, this instrument can be used in a wide variety of team settings because its focus is not domain dependent.

As part of the factor analysis (Johnson et al., 2007), seven sets of team knowledge frameworks were identified and used as the base set of items that were included in the development of an instrument to measure team-related knowledge, knowledge that can be shared among team members (Fiore, Salas, & Cannon-Bowers, 2001; Cannon-Bowers, Salas, & Converse, 1993; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers,, 2000; Klimoski & Mohammed, 1994; Cooke, Salas, Cannon-Bowers, & Stout, 2000; Fiore & Salas, 2004; Mohammed, Klimoski, & Rentsch, 2000). The result of the exploratory factor analysis provided the factorial structure of the team-related knowledge construct. The structure revealed the following five factors.

Factor 1 – General Task and Team Knowledge relates to general types of knowledge, including team knowledge and task general knowledge. This factor does not measure task-specific knowledge (specific procedures, sequences, task actions and strategies).

Factor 2 – General Communication Skills refers to activities that are needed to share information with teammates such as exchanging information, providing summaries, and seeking information. This can involve teams sharing their plans, actions,

and responsibilities, as well as sharing of information to allow them to act in concert without the need for overt communication.

Factor 3 – Attitude toward Team and Task includes various shared values – what teammates believe in and what they are willing to work for.

Factor 4 – Team Dynamics and Interactions consist of team processes and procedures such as task actions, interactions, supporting behavior, and guidance. This factor includes meta-level processes that teams engage in while performing team tasks including problem solving, decision making, and exchanging information.

Factor 5 – Team Resources and Working Environment include technology, organization, synchrony, and geographic dispersion as well as organizational factors (e.g., culture, structure, standard procedures) that affect team interactions.

Reliability of TADI was measured in the factor analysis with Cronbach's alpha values as follows: Factor 1, 0.76; Factor 2, 0.89; Factor 3, 0.75; Factor 4, 0.81; and Factor 5, 0.85. The level of Cronbach's alpha for the items was shown to have adequate reliability (Johnson et al., 2007).

Considering validity, the instrument's content is based on the link of the items to several key theoretical frameworks of team shared knowledge (Johnson et al., 2007, see Table 1). In addition, the instruments construct validity was demonstrated via the six-phase process of the exploratory factor analysis (EFA). In Phase 6, confirmatory factor analysis (CFA) was "used to investigate whether a posteriori postulated factor structure of EFA provided an adequate model for explaining the correlation among the observed variables in a set of data and for determining item redundancies. Specifically, the CFA was used to establish a rigorous model with the closest fit to the model of exploratory factor analysis" (p. 447).

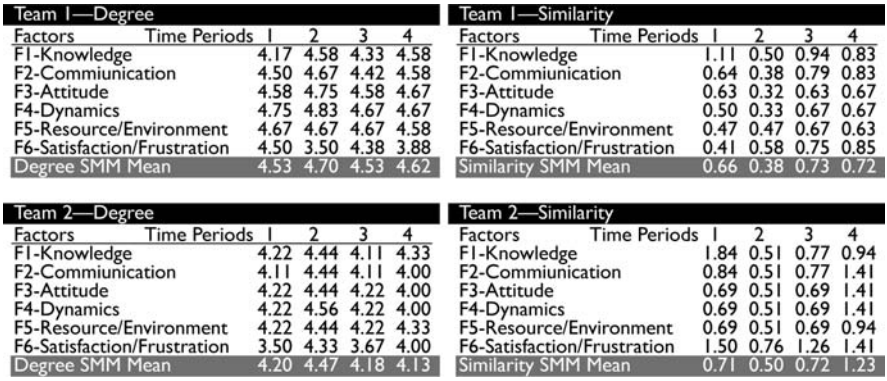
The following results supported of goodness of fit of the team mental model factor structure to the data. The following was reported: chi-square statistic ($\chi^2/\mathrm{df} = 2.186, p = 0.189$) failed to reject the null hypothesis (fit of the initial model was correct); RMSEA of 0.071 was an acceptable fit; AGFI was 0.943 indicating acceptable fit; CFI was 0.989 representing acceptable fit; NFI was 0.981 indicating acceptable fit; and NFI was 0.990 indicating acceptable fit.

17.4 Data Collection and Analysis

The TADI is typically implemented after some level of team interactions. It is repeated several times (typically biweekly) over the duration of the team's task until a task completion is reached or time has run out. Depending on the length and progress of SMM development, the data collection can be terminated based on SMM stabilization. Data is collected by eliciting individual perceptions for the five factors (see Fig. 17.1). If repeated measures are used, each teammate will complete the same instrument again. Due to the amount of individual participants as well as the number of data points, a web-based measure has been used to collect the TADI data.

In an educational setting, the TADI is used typically in conjunction with a team activity. The activity can be related to a lecture or course assignment, or formal project. As a team submits a deliverable, they individually submit their responses to the TADI, typically using a web form.

Once that data is collected, it is exported into a spreadsheet application that is set up to automatically calculate the mean and standard deviation for each team (often times we are studying several teams completing the same task) separated out for each factor for each time period (see Fig. 17.1).



Team 1—Degree

Factors	Time Periods 1	2	3	4
F1-Knowledge	4.17	4.58	4.33	4.58
F2-Commiunication	4.50	4.67	4.42	4.58
F3-Attitude	4.58	4.75	4.58	4.67
F4-Dynamics	4.75	4.83	4.67	4.67
F5-Resource/Environment	4.67	4.67	4.67	4.58
F6-Satisfaction/Frustration	4.50	3.50	4.38	3.88
Degree SMM Mean	4.53	4.70	4.53	4.62

Team 1—Similarity

Factors	Time Periods 1	2	3	4
F1-Knowledge	1.11	0.50	0.94	0.83
F2-Commiunication	0.64	0.38	0.79	0.83
F3-Attitude	0.63	0.32	0.63	0.67
F4-Dynamics	0.50	0.33	0.67	0.67
F5-Resource/Environment	0.47	0.47	0.67	0.63
F6-Satisfaction/Frustration	0.41	0.58	0.75	0.85
Similarity SMM Mean	0.66	0.38	0.73	0.72

Team 2—Degree

Factors	Time Periods 1	2	3	4
F1-Knowledge	4.22	4.44	4.11	4.33
F2-Commiunication	4.11	4.44	4.11	4.00
F3-Attitude	4.22	4.44	4.22	4.00
F4-Dynamics	4.22	4.56	4.22	4.00
F5-Resource/Environment	4.22	4.44	4.22	4.33
F6-Satisfaction/Frustration	3.50	4.33	3.67	4.00
Degree SMM Mean	4.20	4.47	4.18	4.13

Team 2—Similarity

Factors	Time Periods 1	2	3	4
F1-Knowledge	1.84	0.51	0.77	0.94
F2-Commiunication	0.84	0.51	0.77	1.41
F3-Attitude	0.69	0.51	0.69	1.41
F4-Dynamics	0.69	0.51	0.69	1.41
F5-Resource/Environment	0.69	0.51	0.69	0.94
F6-Satisfaction/Frustration	1.50	0.76	1.26	1.41
Similarity SMM Mean	0.71	0.50	0.72	1.23

Fig. 17.1 Data output report of the mean and standard deviation for three teams over four time periods summarized by factor and average of all five factors

In order to calculate sharedness, the average rating for each instrument item would be computed for a given team. This average, the degree of agreement measure or degree measure, represents the degree of knowledge sharedness agreement among the team members. Of similar importance, the standard deviation (SD) of the average score, the similarity of agreement measure or similarity measure, represents how closely aligned the team is on the agreement of a particular item. Depending on the research focus, averages of the item means and standard deviations for each factor could be calculated to determine overall degree and similarity of a specific factor or an overall score could be calculated that combines all the factors together.

While calculating the mean and standard deviation is not mathematically difficult, the time involved to compute each of these by hand is very time consuming for just one team let alone for repeated measures and for multiple teams. With the TADI computer spreadsheet tool, we can quickly view the team results for multiple teams over multiple measures quickly (see Fig. 17.1). Further with the raw data in a spreadsheet, one can access raw data by clicking on the calculated mean and standard deviations.

Using the data from the valid TADI, two measures are calculated – the mean that represents the perceived amount of a given factor that a team has (degree measure) and the SD that represents the level of variation in the individual ratings (similarity measure). Specifically, these two measures comprise the team's SMM which focuses

on how much a team believes they have for each factor (degree) as well as how closely (similarity) they believe they have for each factor.

The TADI SD measure is a measure of team members' mental model similarity with regard to team and task-related elements corresponding to five SMM factors. A low TADI SD measure is a good indication that team members tend to be thinking alike while a high TADI SD measure indicates there is some discrepancy in thinking among the team members. The TADI mean measure, by contrast, represents the level of agreement with a series of statements that correspond to the same five SMM factors. Though the agreement score provides an indication that team members view their team with high regard on SMM factors, the score does not provide an indication of the similarity of team member mental models. A high TADI mean measure is a good indication that most or all team members view the team in a positive light while a low TADI measure indicates that team members generally have a less than positive perception of the team.

What is known about the construct of team factors has been established with an exploratory factor analysis (EFA). Based on the EFA, Johnson et al. (2007) have empirical evidence on how many factors underlie the instrument set. Specific items factor to specific grouping thus providing a data measure about the underlying latent variables represented by the specific factor groups. With the use of this quantitative data, it makes it possible "to assess the accuracy of the knowledge, to aggregate individual results to generate a representation of team knowledge, or to compare individual results within a team to assess knowledge similarity" (Cooke et al., 2000).

With data collected using the TADI, there is evidence to show how teams are sharing their mental models and how these SMMs are changing over time. This data provides key information to help team coaches, managers, instructors, and team members make decision about what to do to improve team performance. The TADI data can be used to determine the strengths and weaknesses of a team, thereby facilitating the selection of team interventions that will support team shared mental model development and ultimately support team performance improvements.

17.5 Interventions

SMMs offer a good value in terms of explaining team processes and predicting team performance. SMMs allow team members to coordinate their actions and adapt their behavior, given task and team member demands (Cannon-Bowers et al., 1993). Team member shared knowledge is crucial for team effectiveness because it allows team members to tailor their behavior in accordance with what they expect teammates to do. Practically, SMM "research can help establish an understanding of the elements of effective teamwork, which can in turn lead to better interventions for improving team performance" (Cannon-Bowers & Salas, 2001, p. 196).

There are a number of team performance improvement interventions that aim to maximize team effectiveness through team-building activities that are domain independent. Such interventions can be found on various websites,

through such organizations as the American Society for Training and Development (Silberman & Phillips, 2005), and by browsing the shelves of any local bookstore. Despite the popularity of such "team-help" books, there exists only a handful of theoretically based and empirically tested strategies for improving team interaction and performance that appear in various psychology, business, and communications journals (Groesbeck & Van Aken, 2001; Guzzo & Dickson, 1996; Kipp & Kipp, 2000; Proehl, 1997).

In a recent meta-analysis, Klein et al. (2009) found team-building activities to be generally effective for improving team cognitive, affective, process, and performance outcomes. Given the link between team SMM and performance, it is no surprise that the team knowledge sharing intervention strategies have been found to enhance both SMM and performance. Before selecting a specific intervention based on the TADI data, however, there are a number of things to consider. This chapter details several considerations to help in the selection of appropriate team development and performance strategies based on team needs as indicated by the findings of the TADI data. Using the TADI mean and standard deviation data for each factor and for the factors combined, a manager or instructor can make practical decisions about selecting team interventions. Strategic decisions can be made regarding the specific nature of the intervention depending on various combinations of the TADI standard deviations and TADI means.

17.5.1 Intervention Decision Making

There are several types of interventions that could be employed to facilitate team improvements. Before specific interventions are considered, the general heuristic presented in this chapter considers the team's level of mental model similarity as represented by the *similarity measure* (TADI standard deviation). What decision makers would want to know first is if the team is thinking alike or not. If the team's thinking is similar, then you would proceed to look at the degree of team development as represented by the *degree measure* (TADI mean). So the overall team improvement strategy looks at (1) seeking to validate or build team consensus, (2) focusing on team improvement planning, and then (3) developing team and task skills.

Because there exist numerous team and task skill development interventions, the emphasis in this section is on presenting the process for determining if a consensus building intervention is needed and then discussing types of consensus building interventions. Furthermore, the emphasis is on determining if a team improvement planning intervention is needed and then discussing the various types of interventions. Although there are a number of interventions out there for building team and task skills, the chapter focuses primarily on suggestions related to team improvement planning interventions whereby teams (and coaches) can plan interventions on how to help correct themselves and become more self-sustaining.

To make the appropriate selection of team interventions, a decision-maker needs to consider the similarity measure and then the degree measure (see Figs. 17.3).

Each component of the TADI has different implications. Each measure is considered separately starting with the similarity measure in phase 1 to determine if consensus building is needed and then the degree measure in phase 2 to determine if skill development is needed.

Phase 1: Determining the Need for Consensus-Building (CB) Interventions

Before significant team interventions begin, teachers and coaches need to be aware of the level of the team's shared mental model. Is the team thinking alike? Do they perceive things in a similar manner? Consider the similarity measure first. If a team has low sharedness, then it is critical to develop shared understanding first (i.e., it is important to get them to closely agree on the levels of the team factors).

When determining whether to use a CB intervention, there are three levels of analysis to consider with the similarity measure: (1) criteria level, (2) time level, and (3) task level (see Fig. 17.2). Each level is described below followed by an intervention decision-making tree that provides a set of heuristics for when to consider introducing a CB intervention.

Level I: Criteria Referenced – The first thing to look at is the level of the similarity measure. The standard deviation ranges (see Fig. 17.2, Level 1) provide a useful criterion for assessing team similarity. This is a logical place to start when determining if a CB intervention should be used. With these criteria in place, it is possible to isolate a particular teams' similarity measures at a given measurement point to see how they compare to these criteria. As stated above, a similarity measure of 0.0–0.33 is considered low and indicates a high level of sharedness. A similarity measure of 0.33–0.66 is considered moderate and indicates a moderate level of

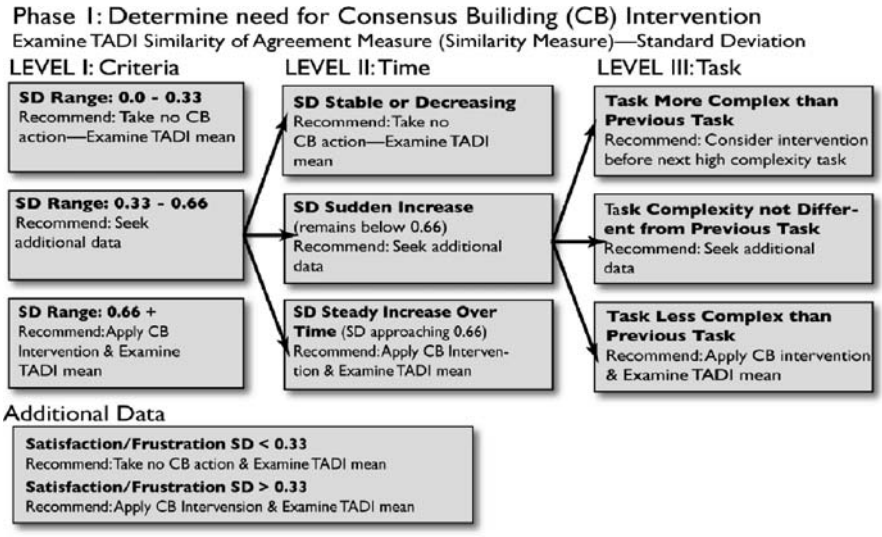


Fig. 17.2 Intervention decision-making heuristic phase 1

sharedness. A similarity measure of 0.66 or greater is considered high and indicates a low level of sharedness.

Using the above criteria alone to determine whether or not to use an intervention is suggested when (1) the similarity measure is lower than 0.33 and the degree measure is greater than 4.0 in which case an intervention is not necessary or, (2) when the similarity measure is higher than 0.66 in which case you may not want to wait for another performance episode before prescribing an intervention. Otherwise, it is suggested that you proceed to Level II and consider similarity measure progression over time.

Level II: Time Referenced – It is often the case that teams work together for several performance episodes, such as on various missions, projects, or assignments. For each of these performance episodes, the TADI should be administered to get SMM data at that particular measurement point corresponding to the performance episode. In the case of several successive performance episodes, it is necessary to consider the teams SMM development over time.

There are several possible scenarios that involve the similarity measure rising or falling over time. Many times deciding to prescribe an intervention will involve a judgment call in comparing the most recent TADI similarity and degree measure with past measures. In our set of heuristics, we refer to the established criteria to determine whether an intervention should be considered or prescribed based on similarity measure over a given time interval. For a steady increase approaching the critical 0.66 level, we suggest applying an intervention. Any sudden increase in the similarity measure or decrease in the degree measure should also be considered in the context of the tasks the teams are performing which is Level III described below.

Level III: Task Referenced – When a team is working together over several performance episodes, they are likely to encounter different tasks that vary with regard to such factors as complexity, intensity, constraints, and role/interaction requirements. Considering these factors can help to better understand the TADI measures and consequently which intervention strategies are likely to work best. For instance, if the similarity measure for a particular team suddenly rises beyond the 0.66 level it may be useful to consider the nature of the task the team was working on for that measurement point. If there is a steady increase in the team's similarity measures, it may be the case that tasks are becoming inherently more complex or that a deadline is approaching. We would expect that the TADI similarity and degree measures are most reliable when the team is facing its greatest challenges and really having to pull together to achieve its objectives – the team required to interact at a higher level.

Seeking Additional Data – In certain cases, the interpretation of the TADI similarity and degree measures does not suggest a definitive recommendation. In these cases, additional data is suggested before making a recommendation. Two useful items relate to team satisfaction and frustration (see Figs. 17.1 and 17.2). Based on the additional data, a team coach can quickly determine if the team is generally similar in their thinking or not and ultimately if a team could use a CB intervention. Once a team has reached a certain level of common ground, it makes sense to focus on other team improvement issues.

For an example, consider the data from team 1 in Fig. 17.2. First consider Phase 1, Level 1. If you are focusing only on the time period 1 data, look at the similarity measure first for the SMM Mean (0.66). This is very close to the cut point so you would most likely want to see additional information. For Phase 1, Level 2, there is no prior TADI elicitation so you will proceed to check the levels of satisfaction and frustration (0.41). This is below the recommended level of 0.66 so we can conclude that a consensus building (CB) intervention is not needed at this point in time. You would next proceed to phase 2 analyses.

If you were focusing on time period 2, your would look at level 1 and see a similarity measure of the SMM Mean of 0.38 and conclude that no consensus building intervention is needed at this point in time. If looking at time period 3, the similarity measure is 0.73 thus indicating that a CB intervention is merited. Looking at time period 4, the 0.74 similarity measure indicates that a CB intervention is appropriate.

There is a tendency for the similarity measure to be less stable in the beginning stages of team development. Individuals are more susceptible to inaccurate perceptions if the team has relatively little experience working together. As a team gains more experience and the task demands increases, individual perceptions tend to stabilize thus providing a more accurate SMM measure.

Phase 2: Determining the Need for Team Improvement Planning (TIP) Interventions

As teams develop toward a high level of SMM, it is important to consider developing the team in other dimensions. As mentioned earlier, team and task skills development is critical, but there are existing interventions that can be used for this. However, one of the most important skills that teams need to develop is the skill of planning and strategizing how to improve the team (see Fig. 17.3).

To determine if a team needs to plan and strategize, consider the degree measure. This measure represents the degree that a team believes that they have a specific factor. A typical range for the mean is 4.0–5.0. Normal teams typically have degree measures in the range of 4.0–5.0 for the overall degree measure or for separate factor degree measures. If a team is in this range, the recommendation is to not give the team a TIP intervention. Teams in the 1.0–2.5 range would definitely see a benefit from TIP interventions.

While not as complex as in Phase 1, there is a degree measure range where consideration of other data is necessary. If the degree measure range is between 2.5 and 4.0, consider the similarity measure. If the measure is greater than 0.5 this would indicate that the degree measure is not very reliable and it is recommended that the team focuses on consensus building. If the standard deviation is less than 0.25, then the degree measure would appear to be reliable and the team should focus on a TIP intervention. If the standard deviation is between 0.25 and 0.5, then seek additional data related to satisfaction and frustration. If satisfaction and frustration levels are high (4.0+) then the team would appear to be getting along.

If the satisfaction and frustration levels are below 4.0, then a TIP intervention would be appropriate.

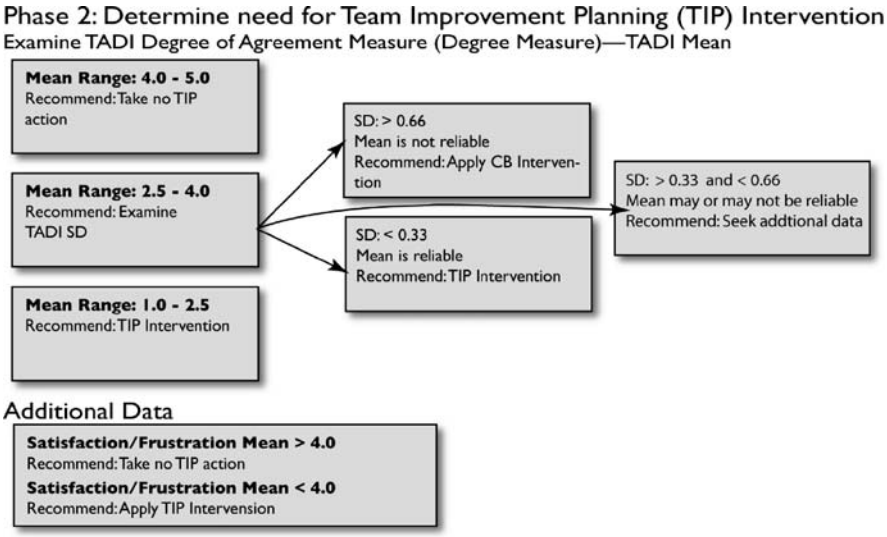


Fig. 17.3 Intervention decision-making heuristic phase 2

17.5.2 Intervention Types

There are a number of interventions to choose from to help a team develop their SMM and consequently perform better. In this chapter, we focus on seeking to validate or build team consensus and on team improvement planning interventions that relate to team needs as indicated by the similarity measure and the degree measure specifically. These interventions are based on empirical evidence to suggest they have been effective for enhancing SMM and/or improving team performance. These interventions can be tailored to suit the team's specific needs (perhaps indicated by problems with a certain TADI factor) or you can develop your own intervention. An advantage of the interventions described below, however, is that they have been empirically tested.

Consensus Building (CB) Interventions

A consensus building intervention may be most appropriate if the team data reveals relatively high similarity measures (considering the criteria, time, and task) or a high frustration and low satisfaction level. The logic behind choosing a consensus building intervention is depicted in the TADI similarity measure decision tree (see Fig. 17.3). Consensus building interventions, such as those described below, are designed to have all team members realize and communicate about the actual state of the team before moving on to strategizing for an optimal state.

Building interpersonal relations is a type of CB intervention that emphasizes increasing teamwork skills such as mutual supportiveness, communication, and

sharing of feelings. The goal is to have team members develop trust in one another and confidence in the team (Klein et al., 2009). Another type of CB intervention is role clarification designed to increase communication among team members regarding their respective roles within the team. The goal is to have team members improve understanding of their own and others' respective roles and duties within the team. Klein and his colleague (2009) found team-building activities, including interpersonal relations and role clarification, to be generally effective for improving team cognitive, affective, process, and performance outcomes.

Team cross-training is yet another CB intervention defined as "an instructional strategy in which each team member is trained in the duties of his or her teammates" (Volpe, Cannon-Bowers, Salas, & Spector, 1996, p. 87). More specifically, cross-training involves team members rotating roles or positions in order to gain the knowledge required for other team members to perform their tasks (Cannon-Bowers & Salas, 1998). Marks, Sabella, Burke, and Zaccaro (2002) examined the efficacy of cross-training by administering the strategy to nearly 100 three-person undergraduate student teams. Results of two experiments described in the paper revealed that cross-training facilitated the development of team interaction mental models.

Team Improvement Planning (TIP) Interventions

One of the most important steps in managing performance improvement is to plan the key activities that will result in overall increase in performance. Based on the TADI similarity and degree measures, team coaches and team members have data to show strengths and weaknesses related to the five-team factors. Due to the complexity and challenges associate with many workplace teams, team intervention planning has the potential to strengthen teams by supporting their development to become self-sustaining and self-leading. Teams need to know how to improve without much supervision and team expertise requires team members to work on gradual improvements over a long period of time.

A TIP intervention may be most appropriate if the team data reveals relatively low degree measures or moderate degree measures with low similarity measures. The logic behind choosing a team improvement planning intervention is depicted in the TADI degree measure decision-making heuristic (see Fig. 17.4). TIP interventions described below involve all team members discussing strategies for how to achieve optimal team SMM (i.e., TADI similarity and degree measure) as well as means for implementing those strategies. The team must then adopt these strategies in order to realize a benefit from this class of interventions.

Goal setting is a type of TIP intervention that emphasizes setting objectives and developing individual and team goals. Team members become involved in action planning to identify ways to achieve goals and become motivated to achieve those goals and objectives. By identifying specific outcome levels, teams can determine what future resources are needed (Klein et al., 2009). The problem-solving TIP intervention involves investigating major task-related problems within the team. Team members become involved in action planning, implement solutions to identify problems, and evaluate those solutions (Klein et al., 2009). In a recent meta-analysis,

Klein et al. (2009) found team-building activities, including goal setting and problem solving, to be generally effective for improving team cognitive, affective, process, and performance outcomes.

A popular TIP intervention for self-directed teams is team self-correction training and is based on the idea that intra-team feedback can foster greater SMM and improve team performance. Team members giving each other feedback about task performance relative to established expectations can help foster similar and correct expectations about team and task work (i.e., shared and accurate mental models) among team members (Blickensderfer, Cannon-Bowers, & Salas, 1997).

A study conducted by Smith-Jentsch, Zeisig, Acton, and McPherson (1998) found that a specific form of team self-correction training called team dimension training (TDT) led to increased similarity of teamwork mental models among shipboard instructor teams. Instructor performance ratings also revealed that TDT aided teams in diagnosing team problems, focusing their practice on specific goals, and generalizing lessons learned (Smith-Jentsch et al., 1998).

17.5.3 Intervention Focusing on Consensus Building and Team Improvement Planning

While we have presented CB and TIP interventions separately, there are a number of interventions that address both of these together. The team knowledge sharing (TKS) intervention is designed to promote team member knowledge sharing relative to five-team and task-related knowledge factors. Team members engage in (1) individual reflection regarding team- and task-related knowledge, (2) discussion of team- and task-related knowledge, and (3) discussion and documentation of ways to improve on specific team- and task-related knowledge areas (Sikorski, 2009). In a study with undergraduate Meteorology student teams, Sikorski (2009) found that the TKS intervention generally enhanced Team-SMM and lead to greater team performance.

Guided team reflexivity is designed to induce reflection in groups. This strategy asks teams to reflect on their performance so far, to consider potential improvements, and to develop plans for how the improvement strategies should be implemented (Gurtner, Tschan, Semmer, & Nägele, 2007). The new strategies are then implemented through action or adaptation. Gurtner et al. (2007) found that guided team reflexivity enhanced the similarity of the team interaction mental models, generated more effective team communication processes, and ultimately improved team performance on a military air-surveillance simulation.

Team coordination and adaptation training is aimed at improving teamwork during periods of high-stress by teaching team members to alter their coordination strategy. The goal is to reduce the amount of extraneous communication associated with performing a given task. A study conducted by Entin and Serfaty (1999) found that Navy officer teams that were given adaptation and coordination training exhibited significantly better teamwork and performance on a flight simulation mission when compared to the control group that received general training.

Another main CB and TIP combination intervention is team-interaction training which is defined as "the training of task information embedded in the necessary teamwork skills for effective team task execution. The content of this training refers to how to work better as a team, not how to perform the task requirements better per se" (Marks, Zaccaro, & Mathieu, 2000, p. 974; Sweet & Michaelsen, 2007). To test the effectiveness of team interaction training, Marks et al. (2000) trained three-person undergraduate teams how to coordinate their actions while engaged in a war game simulated mission. Teams that received interaction training showed greater SMM and mental model accuracy compared to the control group. Performance was also greater among teams that received the training (Marks et al., 2000).

17.6 Extension of TADI

While the Team Assessment and Diagnostic Instrument is rather simple in design, it is open and unconstrained so that it can assess more "natural" team models about knowledge, attitudes, communication, and dynamics. If one desires to reliably and quickly "see" where a team is at, the TADI can provide that information in team settings independent of the team goal. TADI measures, being a simple computer-based diagnostic tool, can be used with other diagnostic tools such as Discussion Analysis Tool (DAT) and jMap, and those found in HIMATT (Shute et al., in press).

DAT calculates transitional probabilities of team discourse. The tool carries out a sophisticated computation, but in order to do so, the data needs to be coded before calculations can take place. The resultant data provides a transitional state diagram that can easily be compared to see potential differences between diagrams. With similar levels of benefit and constraints, jMap is able to elicit and automatically code the resultant models, and has the constraint associated with data interpretation. These computer-based diagnostic tools have varying abilities and constraints.

The TADI is extensible and can be coupled with other mental model measures such as DAT and jMap to carry out systematic diagnostics. For example, one strategy would be to quickly assess a team and then based on their similarity and degree measure, then a focused inquiry such as DAT could be implemented to narrow down the scope of diagnostics there by effectively and more timely reaching the goal of problem identification by examination of the general symptoms as presented by the TADI.

17.7 Application of the TADI Measures for Selection of Team Interventions

Until team coaches and teams become accustomed to using the TADI as a diagnostic tool, follow the guidelines below regarding analysis at each level. Eventually, you

will get a good feel for using TADI similarity and degree measures for diagnostic purposes and to make decisions about intervention selection.

17.7.1 Consider the TADI Similarity Measure

When selecting interventions based on the TADI measures whether at the general level or for each factor, we suggest first looking at the similarity measure to determine whether there is any discrepancy in team member thinking. A high or increasing similarity measure likely calls for a more communication-based intervention before taking corrective action. Taking corrective action when not everyone on the team is aware of problems can be detrimental and lead to further team discrepancy. As your proficiency develops, you will begin seeing the relationship between the similarity measure and degree measure as teams work together in various situations over extended time periods. Understanding this relationship can allow you to combine various intervention components to address both the TADI similarity and degree measure adequately.

17.7.2 Look at the Range of the Similarity and Degree Measures

Making intervention decisions-based TADI similarity and degree measures can become highly complex if you begin by looking at patterns over time or nature of tasks. For instance, a team's similarity measure may be consistently increasing but it is difficult to determine if this is a problem unless you have established criteria to reference. The team's similarity measures are increasing but they are far below the 0.66 critical level. You will likely want to continue to track this team's measures but an intervention is only likely necessary when the measures are steadily increasing and getting close to the 0.66 critical point. As your proficiency develops, you will come to appreciate that the levels are not mutually exclusive as it will often be the case that all three levels or a combination of each level must be considered. For example, teams are often working on various tasks over a given time period. As this point, you may have a good feel for your teams and the TADI instrument that will allow you to make decisions about interventions even when critical levels are not reached.

17.7.3 Examine All of the TADI Factors

There are five factors making up the assessment and diagnostic instrument and six factors if you include satisfaction and frustration. As a novice user, it can become complicated to consider each of the factors in the three-level analysis structure described in Fig. 17.2. When selecting an intervention, first consider the overall similarity measure relative to the criteria then at time and task level. As your proficiency develops, you will be able to pick out specific factors on which the team

is having trouble and select an intervention accordingly. For example, if the similarity measure for attitudes toward teammates and task is high, you will select an intervention that promotes team communication and correction regarding their attitudes.

17.7.4 Focus on One Team at a Time

Often several teams are working independently on the same task during the same time period. This is likely in a classroom environment where student teams are all given the same assignment with the same due date. Until you gain experience using the TADI as a diagnostic tool in several situations, we suggest focusing on each team independently based on the criteria, time, and task rather than making decisions bases on comparisons of the TADI measures across teams. As your proficiency develops, you can compare several teams TADI measures to determine if similar patterns exist across several teams that may be due to some environmental factor.

In summary, there are various benefits for using the TADI. It can be used as a tool to measure the current state of the shared mental model in learning and performance teams. More importantly, the TADI can be used to identify levels of a team's SMM and also a team's degree of perception for five key team factors including: general task and team knowledge, general task and communication skills, attitude toward teammates and task, team dynamics and interactions, and team resources and working environment. Based on the TADI similarity and degree measures, team coaches, instructors, and teams themselves can make strategic decisions whether or not to implement one of the three types of team interventions by looking to (1) validate or build team consensus, (2) focus on team improvement planning, and then (3) develop team- and task-specific skills. While this chapter does not discuss the development of team- and task-specific skills, it does make a case for consensus building along with team improvement planning.

Within each of these two types of interventions, there are specific interventions strategies that could play out in the long run. However, before specific interventions are considered, the general strategy considers the team's level of mental model similarity as represented by the similarity measure. It is most reasonable to determine how much the team is thinking alike. If the team's thinking is similar based on the heuristic rules (see Figs. 17.3 and 17.4) and the degree of team development, as represented by the degree measure, this information supports the selection of appropriate team interventions that provide an effective and efficient mechanism to assess and diagnose team interactions.

While there are a number of interventions out there for building team and task skills, this chapter focused on team consensus building and improvement planning interventions. According to these interventions, teams (and coaches) can recognize where they have weaknesses and plan for how to help correct themselves and become more self-sustaining.

With the key assessment and diagnostic measures, team members, leaders, and coaches can better anticipate team problems thereby guiding the selection of team performance interventions ultimately mitigating these problems and improving team learning and performance.

References

Barrows, H. S. (1998). The essentials of problem-based learning. *Journal of Dental Education*, *62*(9), 630–633.

Barrows, H. S. (2000). *Problem-based learning applied to education*. Springfield, IL: Southern Illinois University School of Medicine.

Blickensderfer, E., Cannon-Bowers, J. A., & Salas, E. (1997). Fostering shared mental models through team self-correction: Theoretical bases and propositions. In M. Beyerlein, D. Johnson, & S. Beyerlein (Eds.), *Advances in interdisciplinary studies in work teams series* (Vol. 4). Greenwich, CT: JAI Press.

Cannon-Bowers, J. A., & Salas, E. (1998a). *Making decisions under stress: Implications for individual and team training*. Washington, DC: APA.

Cannon-Bowers, J. A., & Salas, E. (1998b). Team performance and training in complex environments: Recent findings from applied research. *Current Directions in Psychological Science, 7*(3), 83–87.

Cannon-Bowers, J. A., & Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior, 22*, 195–202.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan (Ed.), *Current issues in individual and group decision making* (pp. 221–246). Mahwah/Hillsdale, NJ: Lawrence Erlbaum Associates.

Cannon, M. D., & Edmondson, A. C. (2001). Confronting failure: Antecedents and consequences of shared beliefs about failure in organizational work groups. *Journal of Organizational Behavior, 22*, 161–177.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. J. (2000). Measuring team knowledge. *Human Factors, 42*(1), 151–173.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. xi, 268). Washington, DC: American Psychological Association.

Eccles, D. W., & Tenenbaum, G. (2004). Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport. *Journal of Sports and Exercise Psychology, 26*, 542–560.

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors, 41*(2), 312–325.

Fiore, S. M., & Salas, E. (2004). Why we need team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 235–248). Washington, DC: American Psychological Association.

Fiore, S. M., Salas, E., & Cannon-Bowers, J. A. (2001). Group dynamics and shared mental model development. In M. London (Ed.), *How people evaluate others in organizations* (pp. 309–336). Mahwah, NJ: LEA.

Gregory, H. (1999). *Public speaking*. New York: McGraw-Hill.

Groesbeck, R., & Van Aken, E. M. (2001). Enabling team wellness: Monitoring and maintaining teams after start-up. *Team Performance Management, 7*(1/2), 11–20.

Gurtner, A., Tschan, F., Semmer, N. K., & Nägele, C. (2007). Getting groups to develop good strategies: Effects of reflexivity interventions on team process, team performance, and shared mental models. *Organizational Behavior and Human Decision Processes, 102*(2), 127–142.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology, 47*, 307–338.

Guzzo, R. A., & Salas, E. (1995). *Team effectiveness and decision-making in organizations*. San Francisco: Jossey-Bass, Inc.

Guzzo, R. A., & Shea, G. P. (1992). Group performance and intergroup relations in organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 3, pp. 269–313). Palo Alto, CA: Consulting Psychologists Press.

Hackman, R. A. (1990). *Groups that work (and those that don't): Creating conditions for effective team work*. San Francisco: Jossey-Bass.

Hinsberg, C. (1996). *Stress rehearsal*. Farmington Hills, MI: Corporate Detroit.

Hitchcock, D. E., & Willard, M. L. (1995). *Why teams can fail*. Chicago: Irwin.

Huszczo, G. E. (1996). *Tools for team excellence*. Palo Alto, CA: Davies-Black.

Johnson, T. E., Khalil, M. K., & Spector, J. M. (2008). *The role of acquired shared mental models in improving the process of team-based learning*. Englewood Cliffs, NJ: Educational Technology Publications.

Johnson, T. E., Lee, Y., Lee, M., O'Connor, D. L., Khalil, M. K., & Huang, X. (2007). Measuring sharedness of team-related knowledge: Design and validation of a shared mental model instrument. *Human Resource Development International, 10*(4), 437–454.

Kipp, M. F., & Kipp, M. A. (2000). Of teams and teambuilding. *Team Performance Management, 6*(7/8), 138–139.

Klein, C., Diaz Granados, D., Salas, E., Le, H., Burke, C. S., Lyons, R., et al. (2009). Does team building work. *Small Group Research, 40*, 181–221.

Klimoski, R., & Mohammed, S. (1994). Team Mental Model – Construct or Metaphor. *Journal of Management, 20*(2), 403–437.

Langan-Fox, J., Anglim, J., & Wilson, J. R. (2004). Mental models, team mental models, and performance: Process, development, and future directions. *Human Factors and Ergonomics in Manufacturing, 14*(4), 331–352.

Lee, M., & Johnson, T. E. (2008) Understanding the Effects of Team Cognition Associated with Complex Engineering Tasks: Dynamics of Shared Mental Models, Task-SMM, and Team-SMM, Performance Improvement Quarterly, 21 (3) pp. 73–95.

Leonard, D., & Swaps, W. (1999). *When sparks fly: Igniting creativity in groups*. Boston: Harvard Business School Press.

Levine, J. M., & Moreland, R. L. (1990). Progress in Small Group Research. *Annual Reviews Psychology, 41*, 585–634.

Marks, M. A., Sabella, M. J., Burke, C. S., & Zaccaro, S. J. (2002). The impact of cross-training on team effectiveness. *Journal of Applied Psychology, 87*(1), 3–13.

Marks, M. A., Zaccaro, S. J., & Mathieu, J. E. (2000). Performance implications of leader briefings and team-interaction training for team adaptation to novel environments. *Journal of Applied Psychology, 85*(6), 971–986.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*(2), 273–283.

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (2002). *Team-based learning: A transformative use of small groups in college teaching*. Sterling, VA: Stylus Publishing, LLC.

Mohammed, S., Klimoski, R., & Rentsch, J. (2000). The measurement of team mental models: We have no shared schema. *Organizational Research Methods, 3*(2), 123–165.

Parker, G. M. (2008). *Team players and teamwork: New strategies for developing successful collaboration*. San Francisco: Jossey-Bass.

Peterson, R., & Behfar, K. (2003). The dynamic relationship between performance feed-back, trust, and conflict in groups: A longitudinal study. *Organizational Behavior and Human Decision Processes, 92*(1/2), 102.

Proehl, R. A. (1997). Enhancing the effectiveness of cross-functional team. *Team Performance Management, 3*(3), 137–149.

Robbins, H., & Finley, M. (1995). *Why teams don't work*. Lawrenceville, NJ: Peterson's.

Salas, E., & Cannon-Bowers, J. A. (2000). The anatomy of team training. In S. Tobias & J. D. Fletcher (Eds.), *Training & retraining: A handbook for business, industry, government, and the military* (pp. 312–335). New York: Macmillan Reference.

Salas, E., & Fiore, S. M. (2005). *Team cognition: Process and performance at the inter-and intra-individual level*. Washington, DC: American Psychological Association.

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. *Educational Technology, 35*(5), 31–38.

Shute, V. J., Jeong, A. C., Spector, J. M., Seel, N. M., & Johnson, T. E. (2010). Model-based methods for assessment, learning, and instruction: Innovative educational technology at Florida State University. In M. Orey (Ed.), *2009 Educational media and technology yearbook*. Westport, CT: Greenwood Publishing Group.

Sikorski, E. (2009). *Team knowledge sharing intervention effects on team shared mental models and team performance in an undergraduate meteorology course*. Unpublished Doctoral Dissertation, Florida State University.

Silberman, M., & Phillips, P. (2005). *The 2005 ASTD team & organization development sourcebook*. Alexandria, VA: ASTD Press.

Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., & McPherson, J. A. (1998). Team dimensional training: A strategy for guided team self-correction training. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 271–298). Washington, DC: American Psychological Association

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors, 4*(1), 61–67.

Sweet, M., & Michaelsen, L. K. (2007). How group dynamics research can inform the theory and practice of postsecondary small group learning. *Educational Psychology Review, 19*(1), 31–47.

Volpe, C. E., Cannon-Bowers, J. A., Salas, E., & Spector, P. E. (1996). The impact of cross-training on team functioning: An empirical investigation. *Human Factors, 38*, 87–100.

Walker, M. A., & Harris, G. L. (1995). *Negotiations: Six steps to success*. Upper Saddle River, NJ: Prentice Hall.

Author Index

D. Ifenthaler et al. (eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, DOI 10.1007/978-1-4419-5662-0,

Subject Index

T