Modeling and Measuring Competencies in Higher Education

PROFESSIONAL AND VET LEARNING
Volume 1

Scope

"*Professional and VET learning*" is a book series that focuses on professional competencies and identities, but also on conditions and societal frames of job performances. It includes education in economics, medicine, handicraft, ICT, technology, media handling, commerce etc. It includes career development, working life, work- integrated learning and ethical aspects of the professions.

In recent years the learning in the professions and through vocational education has become a central part of educational psychology, educational politics and educational reflections in general. Its theoretical modeling, practical application and measurement standards are central to the field. They are also specific for a new research realm which is until now, especially in the US, minor developed. For Europe the dual system, learning in the professional school and – at the same time - learning in the firm, can be a model for studying how issues of professional belonging, professional life meaning, professional biographies, professional change, but also especially professional competencies and sovereignties respectively securities are generated.

The books in this series will be based on different theoretical paradigms, research methodologies and research backgrounds. Since the series is internationally connected, it will include research from different countries and different cultures. The series shall stimulate a practical discourse and shall produce steering knowledge for political decisions in the field. We invite contributions, which challenge the traditional thinking in the field. Professionals who are accountable, available and certificated shall receive through this series a fundamental support, but also new horizons and broadened perspectives of the domain.

Modeling and Measuring Competencies in Higher Education

Tasks and Challenges

Edited by

Sigrid Blömeke

Olga Zlatkin-Troitschanskaia

Christiane Kuhn

Judith Fege

SENSE PUBLISHERS
ROTTERDAM/BOSTON/TAIPEI

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

ISBN: 978-94-6091-865-0 (paperback)
ISBN: 978-94-6091-866-7 (hardback)
ISBN: 978-94-6091-867-4 (e-book)

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

Printed on acid-free paper

TABLE OF CONTENTS

SIGRID BLÖMEKE, OLGA ZLATKIN-TROITSCHANSKAIA,
CHRISTIANE KUHN AND JUDITH FEGE

MODELING AND MEASURING COMPETENCIES IN HIGHER EDUCATION: TASKS AND CHALLENGES

INTRODUCTION

Measuring competencies acquired in higher education has to be regarded as a widely neglected research field. The progress made in empirical research on the school system since the 1990s – for example, through large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) and through a massive expansion of instructional research in general – has revealed that nothing comparable exists at the higher education level. This deficit can be traced back to the complexity of higher education and academic competencies. Not only is there a variety of institutions, programs, occupational fields and job requirements, but also the outcome is hard to define and even harder to measure. Thus, the existing research deficit is caused in part by the complexity that characterizes the academic competencies of undergraduate, graduate and doctoral students owing to the inter- and intra-national diversity of study models, education structures, teaching performances, etc.

In the context of a differentiated tertiary education system, assessing the development of competencies among students presents a methodological challenge. From this perspective, modeling and measuring academic competencies as well as their preconditions and effects set high thresholds. Another challenge is the question of a suitable criterion (e.g., future job requirements) that will help to evaluate the acquisition of competence. The requirements of possible job areas and of academics change constantly.

POTENTIAL

To review and structure the multi- and interdisciplinary field of higher education research, a comprehensive analysis of the international state of research on modeling and measuring competencies in higher education was conducted (Kuhn & Zlatkin-Troitschanskaia, 2011). The report is based on a broad documentary analysis in the form of a literature review and analyses of data (including secondary analyses), among others, in the form of a systematic keyword- and category-based analysis of the core research databases and publications. In addition, seven interviews were conducted with international experts on relevant topics. These enabled the authors to identify global tendencies and areas of

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 1–10.

innovative research in higher education. Overall, the report reveals research deficiencies in the modeling and measuring of competencies of students and graduates, especially in Europe.

At the same time, however, the report revealed that sustainable approaches to empirical higher education research exist (cf. the OECD feasibility study "Assessment of Higher Education Learning Outcomes," AHELO, or the studies in the context of TEDS-M, cf. Blömeke, Suhl, Kaiser & Döhrmann, 2012; Blömeke & Kaiser, 2012; Blömeke, Suhl & Kaiser, 2011). The "Teacher Education and Development Study: Learning to Teach Mathematics" (TEDS-M), carried out in 2008 under the supervision of the International Association for the Evaluation of Educational Achievement (IEA), was the first effort to measure higher education outcomes on a large scale using nationally- and internationally-representative samples (for more details see Blömeke in this volume). The challenges which had to be met with respect to sampling, response rates, reliability and validity, scaling and reporting at some points seemed unsolvable. Research perspectives had to be adjusted across academic disciplines, borders and locations.

The remarkable results of TEDS-M provided substantive indications of how to meet the challenges of higher education research. We learned for the first time on a large scale and from test data about the interplay of teaching and learning at universities, the interplay of various facets of professional competencies, about culture – or better philosophies of schooling – driving the development of the teacher education curriculum, the mediating influence of university educators, and so on (see, e.g., Blömeke et al., 2012; Blömeke & Kaiser, 2012; Blömeke et al., 2011). In addition, the study provided the first concept of benchmarks: what could be possible in higher education if specific circumstances, for example, in terms of entry selection, opportunities to learn or quality control mechanisms, were set in place. Such evidence did not exist prior to the study.

AN INTERNATIONAL CONFERENCE IN BERLIN – EXCHANGE AND INSPIRATION

Much research has to be done to reveal the structure of academic competencies and to make them accessible to assessment. A comprehensive understanding of higher education should include the assessment of domain-specific competencies as well as of generic academic competencies. With respect to the development and generalization of meaningful theories, it is important to focus on individual universities and their programs, and to include research on sometimes idiosyncratic features. The lesson learned from prior attempts in higher education research is that there is a need to create research communities among universities and disciplines and to take advantage of expertise gained in other countries.

The conference "Modeling and measurement of competencies in higher education" (www.competence-in-higher-education.com) hosted by the Humboldt University of Berlin and the Johannes Gutenberg University Mainz, provided an opportunity to do just that. The state of the research in this field was summarized from an international perspective and across academic disciplines. Speakers and

participants took part in an interdisciplinary discourse on various theoretical and methodological approaches to modeling competencies acquired in higher education and also reflected on the strengths and weaknesses of these approaches. They offered insight into the most important research projects currently being conducted and they identified state-of-the-art developments as well as future tasks.

Several controversies and challenges became apparent during the conference. Whereas most of the participants agreed on a definition of competencies as context-specific dispositions which are acquired and which are needed to cope successfully with domain-specific situations and tasks, there was an issue about the range of these dispositions. Should the term "competencies" include cognitive facets only or is it important to include attitudes as well? Insufficient response rates and panel mortality were mentioned as the main challenges, but the limitations of paper-and-pencil approaches to the complex issues surrounding the measurement of higher education outcomes were also of concern. Furthermore, only those competencies which can be measured with regard to psychometric criteria typically are regarded as relevant. Would this limit developments in higher education?

All in all, the conference served as an excellent platform for the exchange of research experiences and perspectives and, thus, provided incentive for a new funding initiative (see below). The conference results documented in this volume may instigate improvements in the higher education system. Such improvements can be implemented on the macro level, the institutional level and on the level of individual teaching processes.

A NEW FUNDING INITIATIVE – REASON AND GOALS

To close the research gap and encourage higher education in Germany to become internationally competitive, the funding initiative "Modeling and Measuring Competencies in Higher Education" (KoKoHs) was launched by the German Federal Ministry of Education and Research (BMBF) at the end of 2010. Apart from the development of competence models, KoKoHs focuses on generating appropriate measurement models and instruments. The funding initiative is intended to provide incentives for basic competence research in the tertiary education sector. It has the following goals:

- To increase the performance of the German tertiary education system
- To keep up with international competence research in higher education
- To develop a foundation for the evaluation of competence development in higher education so that evidence-based policy decisions can be made.

In particular, the initiative is intended to support innovative research projects striving for cooperation among universities. The announcement of the funding initiative elicited 97 high-quality proposals for modeling and measuring competencies: in engineering; economics; education and psychology; teacher education in science, technology, engineering and mathematics (the STEM subjects); and social sciences, as well as generic academic competencies. These fields were selected as priority areas where research needs to start for synergetic

effects to be optimized. After an evaluation conducted according to the criteria of the German Research Foundation (DFG), about 20 research projects were selected. They will receive funding from the end of 2011 or beginning of 2012 until the end of 2014. Experts from various disciplines will work together and network nationally as well as internationally in joint multi- and interdisciplinary research projects while integrating diverse methods. The projects are expected to pay attention to certain – almost quasi natural – areas of conflict, for example, the tension between curricular validity, job requirements and the dynamics of changing labor markets in a globalized world.

Proactive funding initiatives based on deficit analyses and aimed at developing a new field of research often face the problem – if one insists on funding according to quality assessments – that only a small number of submissions can be reviewed positively (cf. Nießen, 2011). In the context of earlier initiatives, the federal ministry noticed to its chagrin that financial constraints were not the limiting factor in terms of funding: the quality of proposals was simply not high enough. However, with the new funding initiative on higher education, the picture has started to change and even high-quality applications had to be rejected. This can be regarded as an important signal of the increasing competitiveness of higher education research.

Coordination offices were opened on May 1, 2011 in Berlin (under the direction of Sigrid Blömeke, Humboldt University of Berlin) and Mainz (under the direction of Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz) to administer the projects and the research program. The coordination offices strive to create a systematic framework for the individual projects and a structured approach, aiming to reach the ultimate goals of the program by developing a superordinate concept. The main tasks of the coordination offices are to cultivate exchange and networking opportunities among the projects being promoted, to use synergies, and to foster the systematic and sustainable promotion of young scientists. A special concern is to maintain international cooperation and use it for exchanging communication within the national funding initiative. The coordination offices are expected to remain open for four years so KoKoHs can be supervised during the complete funding period.

OVERVIEW: THE PAPERS IN THIS VOLUME

The conference and the funding initiative will contribute significantly to the advancement of higher education research. Few other factors are as important to sustainable human progress, social justice and economic prosperity as the quality of education – and it is the responsibility of researchers to contribute by conducting high-quality studies, the results of which will lead to improved understanding of the processes and outcomes of teaching and learning. Laying the foundation for this outcome in the field of higher education was the core aim of the conference. Each talk and poster focused on a pressing issue in this field and the conference – with 230 prestigious participants – was an excellent two-day learning experience and, thus, the conference achieved its aim. The interdisciplinary pool of 16 speakers from the Americas, Australia and Europe reached the conclusion that

there are theoretical and methodological approaches to modeling and measuring competencies in higher education that are worth pursuing.

Part I: Theory and Methodology

Royce Sadler, a Professor at Griffith University in Brisbane, Australia, specializes in formative assessment theory and practice, discusses the term "competence" in his paper "Making competent judgments of competence." He points out that the term "competence" differs only slightly in spelling from "competency" but that there is a conceptual distinction between them which in turn leads to distinct approaches to their measurement. A "competency" is often taken to mean an identifiable skill or practice. "Competence," in contrast, is often understood to consist of a large number of discrete competencies which can be tested independently by objective means. Competence involves being able to select from and then orchestrate a set of competencies to achieve a particular end within a particular context. The competent person makes multi-criteria judgments that consistently are appropriate and situation-sensitive. What is more, the range of situations faced by many professional practitioners is potentially infinite. Dividing competence into manageable components to facilitate judgment has value in certain contexts, but the act of division can obscure how a practitioner would connect the various pieces to form a coherent whole. Sadler makes a plea for more integrative and holistic judgments to arrive at consistent evaluations.

Richard Shavelson, Professor (Emeritus) at Stanford University in the US and a specialist on the measurement of human performance, presents an interesting approach to testing and modeling competency. He describes competency as a complex ability closely related to real-life-situation performance. How to make it amenable to measurement is exemplified by research from the business, military and education sectors. Generalizability, a statistical theory for modeling and evaluating the dependability of competency scores, is applied to several of these examples. In his paper he then puts the pieces together in a general competency measurement model. Shavelson points out, however, that there are limitations to measuring competency on various levels in terms of resources, costs and time.

Fritz Oser, Professor (Emeritus) at Fribourg University in Switzerland and a specialist in developing standards for teacher education, in his paper "Competence Profiles" emphasizes the process of generating criteria against which competence can be evaluated. He claims that basic questions on professionalization background and the identification of standards have to be answered before competence profiles at the university level can be modeled and assessed. Oser demonstrates how the Delphi method can identify vital competencies. He has developed an advocatory approach to measuring competencies based on the assumption that the individual situation defines the competence profiles which, therefore, should be defined from the bottom up. He presents corresponding results from his study.

Mark Wilson, Professor at the University of California, Berkeley (USA), and Karen Draney, specialists in educational measurement and psychometrics, focus on an assessment system which has been developed by the Berkeley Evaluation and

Assessment Research (BEAR) Center. They briefly describe a large-scale assessment context in which they have been developing and applying aspects of the BEAR Assessment System. They describe BEAR in terms of its principles and building blocks and discuss its realization in their large-scale context. Throughout their paper they discuss what their experiences have taught them regarding some of the salient issues regarding assessment.

Michaela Pfadenhauer, Professor at the Karlsruhe Institute of Technology in Germany and a specialist in the sociology of knowledge, in her paper "Competence – more than a buzz phrase and an emotive word?" examines the evolving use of the term *competence* as an indicator of changing educational systems. She points out that in educational policy – at both the national and the supranational level – a "competency-oriented turn" has taken place on such a scale that it is hardly conceivable how it was possible to manage without this phrase. Its rise in popularity was accompanied by a massive replacement of customary concepts: where "qualification," "education" and "educational objectives" previously were discussed, "competency" now seems to be the more accurate, adequate or simply more modern expression. Pfadenhauer takes a perspective on situational problem-solving capacity; on the basis of her phenomenological analysis, she makes a plea for including the social dimension in the definition of competencies.

Part II: Instruments and Studies

Sigrid Blömeke, Professor at the Humboldt University of Berlin, a specialist in the measurement of teacher competence and one of the conference organizers, presents an innovative comparative study carried out under the supervision of the International Association for the Evaluation of Educational Achievement (IEA): the "Teacher Education and Development Study: Learning to Teach Mathematics" (TEDS-M). In her paper she describes the theoretical framework of this large-scale assessment and its design to illustrate how the challenges of higher education research were met. Core results of TEDS-M are documented to illustrate the potential of such studies. Finally, conclusions are drawn with respect to further higher education research.

Karine Tremblay, Senior Survey Manager, Organisation for Economic Co-operation and Development (OECD), France, a specialist in statistics in the areas of student mobility and assessment of learning outcomes in higher education, presents the rationales, challenges and insights derived from OECD's feasibility study "Assessment for Higher Education Learning Outcomes" (AHELO). AHELO is intended to provide evidence of outcomes across cultures and institutions for national and international use in developing policies and practices in higher education. AHELO targets discipline-related competencies and generic skills (critical thinking, analytic reasoning, problem-solving, written communication). In contrast to other OECD studies such as PISA, the unit of analysis is not the country but the institution. Feedback is obtained through performance profiles. Major research questions of the feasibility study are whether instruments are valid in diverse national and institutional contexts, whether the tests meet predefined

psychometric standards and how effective strategies are in encouraging institutions and students to participate.

Roger Benjamin, President of the Council for Aid to Education (CAE) in the USA and a specialist in higher education policy and practice, examines "the principles and logic of competency tests for formative and summative assessment in higher education." He starts his paper with a reminder of the reason why such efforts are made: the future of our highly-industrialized society depends on the realization of human capital. Therefore, a need exists for evidence-based decisions focused, in particular, on the improvement of teaching and learning. Benjamin presents the "Collegiate Learning Assessment" (CLA) as an approach to capturing one key competence to be developed in higher education: critical thinking. In his paper, he presents the lessons learned in developing and adapting this performance assessment instrument for international use. The CLA requires students to use their cognitive abilities to construct responses to realistic problems. Benjamin also addresses an important concern: that taking a test has to be more enjoyable than going to the dentist.

Rafael Vidal Uribe, Director of the National Assessment Center for Higher Education (Ceneval) in Mexico and a specialist in large-scale assessments, presents "The case of Ceneval in Mexico" as an example of measuring learning outcomes in higher education. Two main instruments are used to evaluate college graduates. The EXANI-III evaluates the fundamental skills and competencies of those who have completed college and wish to continue with post-graduate studies. The EGEL examinations are designed to assess the required knowledge expected of scholars on completion of their first degree studies. The EGEL examinations are multiple-choice tests centered on the domain-specific knowledge and skills that are considered essential and common to all higher education institutions' curricula in the specific subject. The objective is to identify whether students have the minimum knowledge, skills and competencies they need to enter professional practice. Results for individual students are reported on one of three levels (outstanding, satisfactory, not yet satisfactory) and described on each subscale. Results for the institutions are reported through the distribution of students on the three levels for each subscale across all subjects.

Hildegard Schaeper, Senior Researcher at the Institute for Research on Higher Education (HIS), is involved in Stage 7 (Higher Education and the Transition to Work) of the German National Educational Panel Study (NEPS) and is responsible for project coordination and management. In her article, she first gives a brief overview of the conception and structure of the NEPS and then describes in more detail its general approach to modeling and measuring competencies and its method of addressing the issue of subject-specific competencies in higher education. The NEPS promises to gain new insights into the acquisition of competencies across the entire lifespan, to describe crucial educational transitions, to study educational careers, to identify the determinants of competence development and educational decisions, and to analyze the impact of education and competencies over the life course.

Olga Zlatkin-Troitschanskaia, Professor at Johannes Gutenberg University Mainz, and Manuel Förster and Christiane Kuhn specialize in the measurement of university students´ competence in the domain of business and economics. As one of the conference organizers, Zlatkin-Troitschanskaia presents the research project ILLEV. It is one of the few projects in the German Federal Ministry of Education and Research's funding program "University Research as a Contribution to Professionalizing Higher Education" that focuses on modeling and measuring subject- and subject-didactical competence, especially among students of business and economics and business and economics education. In the study, the effects of the various courses of study (diploma and bachelor/master) on professionalization and its development over the course of four years are examined. After discussing the study's basic aims and research questions, the research design, and the survey instruments employed, this paper provides a description of the main content and measuring results of the first survey (fall 2008). The paper concludes with a discussion and preview of further approaches in this longitudinal study.

Detlev Leutner, Professor at the Duisburg-Essen University in Germany, and Karoline Koeppen, Johannes Hartig and Eckhard Klieme present the program "Competence Models for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes." This priority program, which is based on proposals written by individual researchers, was set up by the German Research Foundation (DFG) to operate for six years (2007–2013). It coordinates the research of experts on teaching and learning as well as experts on measurement and assessment from the disciplines of psychology, educational science and domain-specific pedagogics in more than 20 projects across Germany.

Part III: Long-term Outcomes

Christiane Spiel, Professor at the University of Vienna in Austria, together with Barbara Schober and Ralph Reimann, specialists in evaluation and quality management in the educational system, stresses the institutional perspective. She focuses on "The Contribution of Scientific Evaluation" to the measurement of academic competencies. Scientific evaluation is based on established standards and systematically combines qualitative and quantitative approaches to data collection and analysis. Spiel makes the plea that evaluation can and should be conducted in all phases of programs and from a longitudinal perspective. Baseline data collected before the start of a program are used to describe the current situation, for example, the generic and domain-specific competencies of students before beginning their university education. In formative evaluation, interim data are collected after the start of a program but before its conclusion. It is the purpose of formative evaluation to describe the progress of the program and, if necessary, to modify and optimize its design. In the case of higher education, the focus might be on how academic study and specific courses support the development of generic and domain-specific competences. Outcome evaluation deals with the question of whether programs achieve their goals. Here, the generic and domain-specific

competences of graduates and freshmen (baseline data) can be compared. Furthermore, the competences of graduates might be evaluated in relation to their correspondence to defined profiles.

Rolf Van der Velden, Professor at Maastricht University in the Netherlands and a specialist in the long-term effects of education on careers, stresses the ultimate criterion of competence acquired during higher education leading to success in life, especially in the labor market. He makes a plea for including non-cognitive facets in this evaluation. Drawing on this background, he discusses two of the main methods of measuring competencies in large-scale surveys among higher education students or graduates: tests, and self-assessments.

Ulrich Teichler, Professor (Emeritus) at the University of Kassel in Germany, and Harald Schomburg, both specialists in the internationalization of higher education, analyze job requirements and the competencies of graduates. Teichler points out that even though the measurement of competencies can be regarded as the most sophisticated approach to evaluating the quality of higher education, drawbacks may exist. Higher education research has to identify the key actors' notions of job requirements and competencies of graduates, that is, the notions of employers, students and academics. He introduces the term "subversity," albeit as a safeguard against the mostly conventional ideas of employers and university professors. Four areas are most salient if improvement is to be achieved: (a) concepts are needed to overcome the "match-mismatch" paradigm, that is, to take into account the necessary concurrent "over-" and "under"-education, the educational tasks beyond professional preparation, the varied values of graduates, the creative function of presumed "over-education," etc.; (b) methods have to become better at de-mystifying misconceptions between job requirements and competencies; (c) ways have to be found to create a better balance between subject-related competencies (e.g., mathematical reasoning) and general competencies (e.g., leadership); and (d) it is still an open question how one should measure competencies and job requirements in such a way that the varied demands in the employment systems and the varied curricular concepts in higher education are taken into serious consideration.

Part IV: Commentary

Judith Gulikers and Martin Mulder took on the task of summarizing and commenting on what was to be learned at the conference from the participants' point of view. They relate the ideas presented in Berlin, among others, to research work done in the Netherlands and, thus, pave the way for an even broader view of measuring competencies in higher education. In particular, they identify the challenges ahead if we are serious about moving forward in this research field.

As the conference organizers and editors of this volume, we are grateful for the contributions of all our speakers and participants. Special thanks go to the members of our Advisory Board, in particular to its head, Prof. Dr Klaus Beck. The Board members supported us with great ideas and recommendations, and also actively participated in the conference by introducing the speakers and leading the

discussions. We are grateful for the support of Manuel Förster, Sebastian Brückner and Katharina S. Bergsma as well. They worked tirelessly prior to, during and after the conference so that it ran smoothly, guests felt welcome and this volume could be issued on time. All contributions were subject to double-blind reviews; therefore, we would like to thank all colleagues who contributed to the reviewing process. Finally, we gratefully acknowledge the funding provided by the BMBF represented by Martina Diegelmann, Michael Kindt and Hartung Hoffmann, who are also in charge of administering the funding initiative. The conference has revealed how complex the task of measuring academic competencies is and that there is a lot of research work to be done. We anticipate, however, that we will move forward substantially over the next three years and beyond – thanks to the more than 20 research projects in this initiative.

Berlin & Mainz, Germany, January 2013

REFERENCES

Blömeke, S., & Kaiser, G. (2012). *Homogeneity or heterogeneity*: Profiles of opportunities to learn in primary teacher education and their relationship to cultural context and outcomes. ZDM – The International Journal on Mathematics Education. DOI 10.1007/s11858-011-0378-6.

Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education*, *62*(2), 154–171.

Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: what matters in primary teacher education? An international comparison of 15 countries. *Teaching and Teacher Education* 28, 44–55.

Kuhn, C., & Zlatkin-Troitschanskaia, O. (2011). *Assessment of competencies among university students and graduates – Analyzing the state of research and perspectives*. Johannes Gutenberg University Mainz: Arbeitspapiere Wirtschaftspädagogik [Working Paper: Business Education], 59.

Nießen, M. (2011). Building structures by research funding? DFG programmes in the field of empirical research in education. Zeitschrift für Erziehungswissenschaft, *Sonderheft*, *13–2011*, 161–169.

PART 1

THEORY AND METHODOLOGY

D. ROYCE SADLER

MAKING COMPETENT JUDGMENTS OF COMPETENCE

INTRODUCTION

Comprehensive English dictionaries list multiple meanings for the words "competence" and "competency". Although the variety of meanings may not matter in ordinary conversations, in rigorous thinking about the measurement and development of competence or competencies, clarity is indispensable. For the purpose of developing the theme in this chapter, a distinction is made between what may be conceptualized as an integrated and large-scale characteristic, capability or attribute, and smaller-scale identifiable elements that contribute to such an attribute, in particular demonstrable skills in performing a task. The first of these, the envelope term, is referred to as *competence*; a contributing element is referred to as a *skill* or *competency*, the latter two being used more or less interchangeably. (Elsewhere, competencies may be called *competences*, and *skill* may be restricted to physical or psychomotor activity.)

The distinction in principle between competence and a skill/competency is convenient but at least partly a matter of degree. Thus mastery of a sufficiently large or complex "skill" may be referred to as "competence in (a particular field)." The nature of the distinction depends on the context and the communicative purpose to be served, and to that extent is arbitrary. Notwithstanding those differences, a competent professional (such as an engineer, dentist or accountant) is characterized by competence in the corresponding field; when professional competence is put into practice, numerous skills or competencies are ordinarily involved. An underlying question is whether competence can be exhaustively decomposed into identifiable constitutive skills, or whether it involves something more than applying a set of separate skills which have been acquired or mastered.

Higher education is heavily involved in the development of both particular competencies and overall competence. Interest in these concepts has increased dramatically in Western countries over recent decades. Many employers along with academics who teach advanced programs have expressed disquiet (or even dismay) about the perceived shortcomings of new graduates' general competencies. Whereas previously it could have been taken for granted that these competencies were developed during degree studies regardless of discipline, field or profession, it is currently alleged that this is no longer the case. Responses to these concerns by higher education institutions and quality assurance agencies have included: the identification of general attributes and skills that are important in contexts after graduation, being potentially transferable from academic degree studies to

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 13–27.

workplaces, to advanced studies, across career sequences and to life in general; the development of sound ways to assess such "graduate competencies"; and the design of strategies to improve student performance on them.

Taking as a given that the concerns are justified, what changes in higher education over the same time period may account for them? The many factors are no doubt interrelated, but only three are identified here, the third being elaborated later in the chapter. First, access to higher education has been progressively opened up from an academically elite segment of the population to a significant proportion of the population (the so-called massification of higher education). A result of this has been that at the point of entry many students are now regarded as being inadequately prepared for academic study. Second, the costs of higher education have generally risen and public financial support has generally either fallen or not kept pace in real terms, forcing institutions to economize (one way of cutting teaching costs being to rely progressively and more heavily on part-time academic teachers). The third has to do with changes in teaching and assessment, the aspect of specific relevance to this chapter.

Not surprisingly, institutional lists of intended graduate capabilities show significant overlap. Common elements include student proficiency in: analytical and critical analysis; problem-solving; locating, evaluating and using relevant information; originality, initiative and creativity; and effective communication. This particular selection has a strong emphasis on cognitive outcomes and these are the ones focused on in this chapter, but institutional lists are typically more expansive. Although interest in these types of competencies has been international, the broad movement does not share a standard terminology. Most lists have been framed under headings which are either "graduate" or "generic" and paired with one of the following: attributes, competencies, capabilities, outcomes or skills.

That said, some institutions have two lists, one labeled "generic skills" for specific competencies of the type listed above; and those labeled "graduate attributes" for large-scale student characteristics related to professional outlook and orientation such as: interdisciplinarity; collaboration and teamwork; high ethical standards; a globalized or internationalist perspective; cultural and linguistic sensitivity; social and civic responsibility; lifelong learning; and commitment to sustainability.

In recent years, significant support has been given to the principle of modeling and measuring competencies by broad-spectrum testing of all graduates in a given jurisdiction, preferably by standardized means. The collection of competency measurements is intended to represent levels of graduate competence. In some contexts, differentiation in the content of tests has been proposed as a means of achieving a satisfactory fit for various areas of specialization, something more difficult to achieve with a single omnibus test for all students. Despite those initiatives, the broad interest remains in measuring competencies which characterize graduates irrespective of the particular courses, programs or institutions in which students enroll. Mass testing of graduate competencies is proposed as a way of enabling trends in teaching effectiveness to be identified,

comparisons across institutions or systems to be made, and quality assurance procedures to be more objective and driven by results.

An additional line of thinking is that if performances in tests of graduate competencies are publicized in the form of institutional rankings, this could incentivize poorly ranking academic programs or entire institutions to redirect some of their effort and resources towards improving the performance and employability of their graduates and thus improve their relative standing among similar institutions. A further possibility is that if mass testing were carried out early in an academic program and then again after graduation, gain scores could provide a measure of the value added by participation in higher education as part of the social return on investment. All in all, this initiative has been widely advocated as a logical, direct, efficient and politically feasible approach to the open scrutiny of institutional attainments, the discovery of shortfalls, the implementation of remedial strategies, and the accountability of higher education institutions in terms of playing their full part in national growth, development and prosperity.

Although the importance of the types of cognitive competencies in the sample list above is widely recognized, it does not automatically follow that the most appropriate way forward is to spell out what is to comprise each competency and then implement mass testing programs. In this chapter, the outline of an alternative view is presented. It does not pretend to be a fully argued case or to set out a comprehensive plan for action. The development flows from a number of reservations held by a disparate group of researchers and commentators about: the philosophical legitimacy of decomposing competence as a complex concept into constituent skills-competencies; the uncoupling of various competencies properly expected of study in higher education from regular academic programs and courses; and the prospect that mass testing and its flow-on effects could divert attention and resources away from the primary sites at which competencies should be developed, practiced and refined, these sites being normal academic studies.

In this alternative view, the generic competencies would remain firmly situated within the various disciplinary or professional educational contexts. The final step would be the assessment of these competencies. This would be integrated into holistic judgments of the quality of student work against recognized academic achievement standards which are comparable across courses and academic programs (and, where appropriate, across institutions). Both this goal statement and tentative principles for achieving the goal through systematic peer consensus processes are developed in more detail in four of the author's articles (Sadler, 2009a, 2009b, 2010b, 2011).

DECOMPOSITION

Conceptualizing competence as made up of a number of underlying competencies is an example of a general approach to tackling complex problems and phenomena. Decomposition into constituent parts has proved a powerful tool for probing and developing understanding in many areas of thought and practice. If a complex entity is to be put to practical use, decomposition often makes it possible to devise

methods for testing all the parts separately and then checking that they all function together as they are supposed to. This is well exemplified in mass manufacturing processes. It has also played a significant part in the way technology and the physical and biological sciences have advanced. Parts have been identified, relationships and dependencies explored, theorizing and hypothesis testing carried out, and predictive models developed so that theorizations can be tested. Where appropriate, processes have been modeled with a view to monitoring and controlling them so they can serve human needs.

At this point, a short digression shifts the focus to an adjacent field of education. Decomposition has been a common feature in post-compulsory education, particularly in the vocational and training sectors of countries such as Australia and the United Kingdom. Complex outcomes have been broken down into smaller and smaller skills or competencies, which have then been taught, practiced, tested and checked off a master list when "achieved." The competencies themselves are typically identified through consultation with representatives of trades, crafts, arts, industry and labor unions in a bid to insure they are empirically based and the full set is as complete as possible. Under this model, attainment of all competencies leads to accreditation as a qualified tradesperson or practitioner. One of the claimed instructional advantages of describing multiple competencies in detail is that the competency descriptors provide highly visible targets for instructors and students alike, increasing the likelihood they will be reached and then counted towards a qualification when achieved. This system therefore sounds rational and straightforward. Furthermore, it can produce competent practitioners provided it is accompanied by overt attention to the development of strategies for choosing the most appropriate skills, materials or actions in order to achieve the solution to a given problem.

The case of vocational education and training is instructive for two reasons. The first is that, in practice, the decomposition of vocational capability has been applied to a particular class of skills or procedures which are distinctively different from the higher education skill-competencies focused on in this chapter (critical analysis and so on). Many of the skills common in vocational and technical education and training are of the physical, practical, concrete kind. Similar types of skills are often applied in a range of settings, each "skill" being identified both by the "object" to which it is applied and the intrinsic nature of the skill itself. Not uncommonly, skills are described in terms of both criteria, making them distinguishable in concept and in practice. The contexts in which they are learned have a certain degree of routinization or repetitiveness about them, allowing the skills to be rehearsed and mastered separately. For these reasons, it makes sense to treat them, at least in the context of initial training, as distinct skills.

The vocational education context is instructive for another reason. Decomposition into constituent skills can lend itself to seriously deficient implementation, as has become evident in the United Kingdom. (The November 2007 issue of the journal *Assessment in Education: Principles, Policy and Practice* contains a number of reports of research into the UK experience.) The troublesome aspect has been that, for some qualifications, the competencies have been so finely

grained and the assessments so compartmentalized that teachers have moved towards deliberately coaching their students over the pass line for each competency, one by one, in order to enable them to gain a marketable qualification. In extreme instances, this practice has resulted in a particular competency exercise being completed by the student just once, with constant prompting by the instructor. This in turn has been openly defended as both appropriate and necessary for scaffolding student learning. With scaffolding as the rationale, the skill has been checked off and elevated to the status of an acquired competency. No doubt this practice is not what the curriculum developers intended but it does illustrate how component-based assessment practices can undermine progress towards the goal of overall competence. The collection of discrete competencies "passed" does not necessarily translate into a coordinated ability to complete a complex task with proficiency (Sadler, 2007).

Although decomposition of a complex entity may be carried out in order to achieve some gain, this gain is accompanied by loss of a different kind: it becomes more difficult to see the whole as a unified competence. The logic of this phenomenon is obvious. If something is divided into pieces, whatever originally held it together and accounted for its integrity has to be either supplied or satisfactorily substituted if the sense of the whole is to be restored. In the context of higher education competencies, the "whole" is the graduate who can operate competently, intelligently and flexibly, in contexts that are known now and in those that have not yet been faced or even envisaged.

HIGHER EDUCATION COMPETENCIES

Compared with many of the technical and vocational competencies, the cognitive attributes previously listed (critical analysis, problem-solving, information literacy, originality, and effective communication) are not as easily defined in concrete terms. It is difficult to describe exactly what "critical analysis" consists of, and in particular whether an actual analysis contains enough of the right kind of stuff for it to warrant the label "critical." Assuming that this property is not an all or nothing affair, it is difficult to describe in words where the threshold for an acceptable amount should be set. The same sorts of difficulties arise with "effective" communication and others in the list. To address this issue, it is not uncommon for higher education institutions to develop extended descriptions of what is covered by each of the attributes, elaborations sometimes running to many pages.

As a limited example, consider the following expansion for information literacy, which has been adapted and condensed from an actual discipline description.

The graduate should be able to:

- Access archives, libraries, the web and other written, oral and electronic sources of data and information;
- Effectively employ appropriate technologies in searching out such information;
- Apply research principles and methods to gather and scrutinize information;
- Manage, analyze, evaluate and use information efficiently and effectively in a range of contexts; and

– Respect economic, legal, social, ethical and cultural norms and protocols in gathering and using information.

In interpreting each of these sub-competencies, contextualized judgments are necessary. What each of them means in the abstract and implies in practice is therefore open to interpretation and debate. When institutional descriptions differ significantly, as they typically do, which should be taken as definitive, if any? How much does it matter if different interpretations exist and are used? Social and contextual dependence is signaled by the fact that different meanings of the key terms and concepts are obviously satisfactory to the institutions in which the statements have been formulated. A further aspect is that much the same wording for competencies can be found in formal lists of desired educational outcomes for various levels of mainstream school education. That is, the intrinsic content of the competencies is not definitively characteristic of any particular level of education. Some (such as problem-solving) appear across the educational range, from kindergarten upwards, presumably because they form part of what is normally expected of education broadly interpreted – that is, what education as a collective enterprise is all about.

The above sample of higher education competencies also serves to illustrate the essential fuzziness of the relationships among them. Although they may appear in the abstract to be conceptually distinct, those distinctions are not simple to sustain in practice. The attributes fuse into one another. For instance, problem-solving as an intellectual and practical activity is difficult to conceptualize without involving analysis, seeking out relevant information, creative development (of possible solutions), and effective communication of the solution. Where one competency or skill finishes and another starts is a fine line to draw. Furthermore, the attainment of a particular subset of competencies may, when applied, have "covered" the territory normally associated with one or more other competencies and thereby made the separate assessment of the latter redundant. Potential overlap, nesting, and partial or full interdependencies are common. This raises the issue of the extent to which it is feasible, as an exercise in the abstract, to differentiate the competencies at all. Separating and clarifying "competencies" for the express purpose of constructing tests to measure them is at best a partial exercise because separate reporting of the competencies cannot capture typical (and inevitable) in-context entanglements.

On the other hand, it is important for some purposes to be able to embrace and use the concepts as concepts. They have meaning, they have labels, and they provide both the vocabulary and the tools necessary for making systematic, functional progress. Where they most appropriately fit into the scheme of things could well be as retrospective explanatory devices – after particular judgments have been made. This would be consistent with the philosophical position that valuing, or making evaluative judgments, is a primary act of situational recognition, the justification for which necessarily invokes relevant criteria that are extracted as needed from a larger pool of potential criteria (Sadler, 2009a).

By way of concrete example, suppose an assessor composes a rationale for a holistic judgment of the quality of a student's written response to an assessment

task. Suppose, too, that the rationale refers to a lack of critical analysis in the work. The main purpose served by the statement that the work lacks the necessary critical edge is to draw attention to a desired feature which is inadequately expressed in the work. The assessor chooses this quality for explicit mention from the pool of properties that potentially matter. This act connects the particular work with the judgment made about its quality. A person interpreting the rationale in the absence of access to the work in question has no option but to guess what the work was like. Interpreting the rationale in the presence of the work, however, means that the text and its referent combine together in a message. The soundness of the reason for specifically emphasizing critical analysis can then be explored. The dynamic of the way in which critical analysis as a concept is used with and without access to the work makes a difference. Without the work, the temptation is to commodify the concept rather than communicate a judgmental framework.

In practice, only a small number of aspects or characteristics may be worthy of specific mention. What makes these salient to the appraisal is that they provide insights into the evaluative reasoning behind the judgment. In the process of operating in this mode, some properties will turn out to be pre-emptive. For instance, if a written document is so poorly expressed that it is fundamentally incoherent, it is technically unlikely it will be able to provide evidence of originality or critical analysis – or even of whether the work addresses the nominated issue at all. If the end user, whether professor, peer reviewer or employer, attempts to read between the lines of the text to figure out what the author was possibly trying to express, the high-order inferences involved in that process come at the risk of a poor judgment of actual mastery of the competency of interest and the credit that should be given to it. (This observation, of course, is not intended to imply that all text must be held to literal interpretation; linguistic convention may clearly signal other interpretations, such as irony or humor.)

Real contexts are in some ways simpler and in other ways more complex than is implied by thinking about separated competencies. They are simpler in that competent practitioners or producers normally go about whole tasks without much conscious thought as to the cognitive processes or competencies they are using. They switch effortlessly from figure to ground, and back again. Real contexts are more complex in that when producers do actually reflect on their processes and have reason to describe them, their descriptions are not necessarily framed in accord with pre-existing typologies, but adverse consequences that arise from doing this are rare. Those concerns aside, there is no denying the critical importance of a shared vocabulary with which to engage in discourse about quality and qualities, competence and competencies.

Further questions arise in relation to the legitimacy of treating competencies carrying the same label as somehow similar in essence, structure and cognitive demand across a range of disciplines, fields and professions. The similarity of labels is presumably the reason for treating them as "generic," but whether the apparent common ground accords with reality is questionable. Research on this topic has revealed wide differences in interpretation of specified competencies or attributes in different fields, and even within different sub-domains of a single field

(Jones, 2009). Critical analysis expresses itself differently in different content areas, and at different academic levels within the same content area. Locating, evaluating and using information is carried out differently in degrees in music, information technology and construction engineering. Within construction engineering, it may depend on the purpose for which the information is required and the time available for obtaining and processing it. A broad-spectrum test that purports to tap into critical analysis and information literacy as graduate competencies may not produce test results that can be interpreted appropriately, that is, matching the label for that competency as used in various curriculum specialties. If de-situated test tasks signal different nuances of academic competencies from the mainstream courses in which students enroll, and if the test results are to be used for high-stake decision-making (both of which seem to be likely propositions), to what extent would teaching time, resources and energies be diverted from the mainstream studies in order to provide space for explicit coaching to the tests?

ASSESSMENT OF COMPETENCE AND COMPETENCIES

A significant challenge facing policy-makers is finding appropriate paths through the assessment of overall competence or of individual competencies. One approach to the measurement task is to first formulate and define the competencies as psychological constructs and then to apply psychometric methods. An influential contribution to the substantial and growing literature on this approach is the review and analysis by Weinert (2001). Tested graduate competencies may be considered to stand each in its own right; alternatively, the collection may be interpreted as an assessment of graduate competence. Suppose it is accepted that a person is judged competent if they perform well over a variety of relevant contexts and challenges time after time, with little likelihood of getting things wrong. In the latter case, the collection should be examined to ascertain the extent to which the competencies tested comprise a necessary and sufficient set. Such examination would have two branches.

The first branch would be a test for necessity: do people who are already recognized as (holistically) competent in the workplace or in professional practice demonstrate all the tested competencies? The second branch would be a test for sufficiency: if graduates were clearly able to demonstrate achievement of all the tested competencies, would they subsequently function as (holistically) competent in the context of work or professional practice? Quite apart from the workplace, would they demonstrate respect for evidence, rigor, people, problems and society in their thinking, communications, and general approach to knowledge and knowing? Are these not the types of "educated" characteristics one should be able to expect of higher education graduates? These are important questions which are in principle open to empirical investigation.

An alternative to the decomposition and measurement approach is to start with the common notion of competence and seek out responsible ways to make judgments about a student's level of competence directly and holistically, rather

than by building up the judgment from components. The motivation for proceeding in this direction is the premise that the whole (competence) does not necessarily equate to the sum of the parts (the competencies). ("Sum" here is intended to include all methods of compounding or combining, as well as simple addition in the case of measurements.) This view implies that judgments of competence can properly take place only within complex situations, and not componentially. Generally, the perception is that if the whole differs from the sum of the parts, it does so in the direction of being more than – not less than – the sum of the parts, but differences in the opposite direction are not uncommon either. As Ford (1992) explained it:

> Organization exists when various components are combined in such a way that the whole is different than the sum of the parts. This "difference" involves both gains and losses. In one sense, the whole is *greater* than the sum of the parts because new qualities or capabilities emerge from the relationships among the parts that none of the parts could accomplish on their own…In another sense, however, the whole is *less* than the sum of the parts because the functioning of each of the parts has been restricted by virtue of being "locked in" to a particular organizational form (p. 22).

Reviewers of films, computer games and other creative works sometimes remark that a work under review satisfies all of the generally accepted criteria of excellence (that is, all of the components appear to be technically perfect) but the work as a whole nevertheless fails to "come together" in a way that sets it apart as outstanding, or even just satisfactory. In some cases, several reviewers independently remark that they have difficulty "putting their finger" on the residual problem or weakness, although they clearly sense it. Conversely, when the whole is judged to be more than the sum of its parts, the "something more" that makes up competence includes any extra qualities or properties, of whatever kind, that were not initially identified as attributes or competencies, and maybe others that cannot be clearly identified and named at all. It also includes the ability to "'read" a particular complex situation which is not exactly like any seen before, and know how to call on the various competencies (assuming they can be identified) productively, adaptively, confidently, safely and wisely.

Put somewhat differently, competence could be conceptualized as selecting and orchestrating a set of acquired competencies to serve a particular purpose or goal. In Ford's (1992) terms, organization makes a difference. The ability to orchestrate competencies, by definition, lies outside (and at a higher level than) the given or specified set of basic competencies. If higher-level competencies were also included in the model, the question would then arise as to how and when these also should be invoked, and the same would apply at even higher levels. In the other direction, as decomposition progresses downwards potentially to the atomistic level, it typically becomes harder and harder to conceptualize the components working together, partly because the number of possible interactions of all orders among competencies escalates rapidly.

With this interpretation of competence, sound judgments of competence require qualitative appraisals of how well a person can get it all together in a given situation. Such judgments are integrative and holistic, and are commonly made subjectively. The term "subjective" is frequently used to denigrate holistic appraisals as being little more than mere opinion or personal taste, in some cases with one opinion being more or less as satisfactory or legitimate as any other. Equally problematic are the terms "impression" and "gut feeling." That line of thinking does subjective judgments a grave disservice. Many professionals constantly rely on so-called subjective judgments that cannot be verified by independent objective means such as a standard laboratory test. Subjective judgments can be soundly based, consistently trustworthy, and similar to those made by comparably qualified and experienced professionals. They can also be poorly based, erratic and unreliable. Furthermore, in some circumstances quite different judgments may be equally appropriate for different purposes.

The goal to aim for is this: when presented with the same phenomena or objects which cover a diverse range, members operating within a guild of like-purposed professionals would make the same judgments within a tolerable margin of error. The judgments hold (that is, are accepted as proper) beyond each judge's personally constructed decision space (that is, the space available only to a particular judge), and the parameters for that shared decision space are set and accepted collegially. For a given set of phenomena or objects, the meaning and significance of evidence are shared, as is what is deemed to count as evidence. In short, given the same stimuli, the people making the judgments would react or respond similarly and judge similarly. The existing term that is probably closest in meaning to this state of affairs is "intersubjectivity," a term used with appropriately nuanced interpretations in phenomenology, psychology, philosophy and several other fields. Intersubjectivity is distinct from interscorer reliability or consistency in that not only are similar judgments made but the grounds for the judgments are shared as well. Consistency on its own can be potentially achieved without that. It is also distinct from objectivity in the sense that it is an *objective* fact that one water molecule contains two hydrogen atoms and one oxygen atom.

As Scriven (1972) has pointed out, the quality of a judgment made by a single assessor is not automatically suspect and deserving of being dismissed merely because it has been made without collaboration and without the help of instrumentation. The two latter conditions do not make all such judgments worthless. Professionals who consistently arrive at sound judgments are effectively "calibrated" against their competent peers and also, in professional contexts, against any relevant socially constructed external norms. This points to the direction in which the development of an appraiser's ability to make high-quality holistic judgments can conceivably take place – by providing them not only with experience in making multiple judgments for objects or phenomena in a given class in a wide variety of settings but also with experience in verbalizing their reasons and discussing them with appropriate colleagues who at least initially have

access to the same objects or phenomena so that the shared decision space which is crucial to the enterprise can be constructed.

In the context of assessment where judgments are holistic and integrated, the characterization above is suggested as the appropriate goal statement. The starting-point for making progress towards acceptable levels of intersubjectivity is daunting, given the well-established research finding that assessors who make judgments by operating within their personal decision spaces generally exhibit low interscorer reliability. Furthermore, in some higher education contexts, the right of academic teachers to make grading decisions that way (that is, as they individually see fit) is strongly defended. The challenge ahead is to find ways to create and value shared rather than individuated meanings and knowledge as a primary resource for making competent professional judgments. What might that involve?

COMPLEX JUDGMENTS – THE IMPORTANCE OF NOTICING

In his characterization of knowledge types, Ryle (1949) made a distinction between "knowing how," which is being able to do something whenever required, and "knowing that," which is knowing something such as a fact, a theorem or a classification. Know-that knowledge is commonly memorized, and tested by using language-based items or tasks (words, symbols or other material representations). Know-how knowledge is commonly learned through practice, and tested by setting up various skill-based tasks. Largely overlooked in Ryle's dichotomy is another form of knowing: "knowing to," in which an appraiser notices, detects or "senses" as salient-in-the-circumstances some aspect that contributes to or detracts from the overall quality or effectiveness of a work. In knowing-to, high-level judgments are critically important. This type of knowledge cannot necessarily be made explicit, that is, expressed in words. It nevertheless exists and is widely used, even when a person cannot define it in concrete terms or otherwise explain it. Such know-to accounts for part of what chemist-philosopher Polanyi (1962) called "tacit knowing," captured in his remark that one can know more than one can tell. A decade earlier, Wittgenstein (1953) had expressed much the same idea in his observation:

> I contemplate a face, and then suddenly notice its likeness to another. I see that it has not changed; and yet I see it differently. I call this experience 'noticing an aspect' [XI, p. 93].

Similarly, Abercrombie (1969) in her classic work on judgment discussed the intricacies of perception and prior expectations and how they influence what is noticed and deemed to count as data in a particular context. Consistent with the work of Polanyi, Wittgenstein and Abercrombie is that of Dreyfus and Dreyfus who argued that experts regularly use their "intuitive rationality," on occasion engage in "'deliberative rationality" (when time permits and this provides a workable way forward), and much less often employ formal "calculative rationality." In their 1984 article, they put it this way:

> [E]xperience-based similarity recognition produces the deep situational understanding of the proficient performer. No new insight is needed to explain the mental processes of the expert. With enough experience with a variety of situations, all seen from the same perspective or with the same goal in mind, but requiring different tactical decisions, the mind of the proficient performer seems gradually to decompose this class of situation into subclasses, each member of which shares not only the same goal or perspective, but also the same decision, action, or tactic. At this point, a situation, when seen as similar to members of this class, is not only thereby understood but simultaneously the associated decision, action or tactic presents itself. [p. 225].

The substantial literature on the nature of expertise and how it is developed is an important resource for further thinking. A great deal of what experts do has to be learned through extended experience, but not necessarily through experience alone, a particularly important contribution to that aspect being the seminal volume of Bereiter and Scardamalia (1993). As well as the authors listed, the literature includes research relating to the development of competence in appraisal by medical and health practitioners, airline pilots and many other professionals who are involved in complex decision contexts.

DEVELOPING HIGHER EDUCATION COMPETENCIES

In this section, the third possible cause of concerns about current levels of higher education competencies is picked up again. The dual agenda consists of two questions: what current aspects of teaching and assessment inhibit the development of higher education competencies? how might improvement be brought about? The proposal outlined below is based on the notion that the responsibility needs to be shared between academics as educator-assessors and higher education institutions as controllers of the parameters within which academics work. An approach followed by some institutions is to make it mandatory for course designers and directors to embed at least some of the higher education competencies in each course. The hope is that over an entire degree program all competencies would be developed. This assumes, of course, that competencies are conceptually separable, something which goes against the grain of the theme in this chapter, but on the positive side it allows competencies to be expressed in ways relevant to individual courses. A second approach is to focus on developing the assessment competence of higher education teachers, strengthen their resolve to award course grades according to appropriate academic standards, and concurrently reset the system and the institutional parameters to facilitate both of these.

With the second approach in mind, academics would need to develop high-level capability in: designing assessment tasks that are clearly specified and outline the higher-order cognitive outcomes required, including the specification of a particular product type (such as a critique or a design) if appropriate; holding students to the specifications (task compliance); and becoming calibrated in their appraisal practice so that the standards they employ are not peculiar to themselves.

Task compliance (Sadler, 2010a) implies not awarding credit at pass level or higher for works that do not deliver on the specifications. In particular, merely compiling and reproducing written material without serious intellectual engagement with it may not qualify as evidence of academic achievement, nor would purely the effort that may have been put into producing a response.

That might seem an obvious way forward except for the fact that many academics assert that if they were to apply the standards in their heart of hearts they know they ideally should, the result would be failure rates that are unacceptable to their institution or to external higher education authorities. The policy settings of many institutions and systems work against the realization of the goal. To illustrate, consider an academic faced with a high level of difficulty in deciphering a student's work. In theory, almost incoherent or poorly communicated work should disqualify a student from passing, that is, from gaining credit and progressing to the next stage of the academic program. In practice, non-achievement variables can and do influence many grading decisions. One such variable is related to maintaining high student retention rates in contexts where recruitment of students is competitive and institutional income from public or government sources is, by legislation or policy, inversely related to attrition rates. That external constraint is mirrored by an internal dispositional aspect on the part of an academic: the assessor may wish to award a realistic course grade to a student but in reality is faced with the likelihood of adverse consequences (poor student evaluations or an institutional inquiry into too high a failure rate) and so leans towards giving the student the benefit of the doubt.

An additional practice that detracts from the integrity of course grades is the awarding of marks or points for what on the surface may appear to be sound educational reasons (for encouragement, for demonstrating improvement, or as a reward for engagement or participation) but in reality amount to giving false credit for elements that are not strictly achievements at all. Furthermore, if the course grade is intended to represent the level of achievement reached by the end of the course, something usually implied or stated in the intended course learning outcomes, accumulating points throughout a unit of study is unsound. These and related topics are dealt with in detail in Sadler (2010b). Such practices can create a near-pass score by dubious means. The consequence is that students then have to attain relatively few of the most highly valued educational outcomes in order to pass, gain course credit and progress to the next course. Regardless of whether these factors have their origins in overt institutional policy or are simply practices that have been accepted incrementally into the assessment culture, they reduce the likelihood of attaining the desirable higher-order academic outcomes in graduates. Turning things around would not be fast or simple but the payoff might well be worth the effort.

CONCLUSION

The point of view reflected in this paper follows a different line from that of most contemporary developments. The focus is not on large-scale modeling of

competence or competencies and the measurement of their attained levels in higher education institutions. Instead, the focus is on the concept of competence as the capability to orchestrate knowledge and skill independently, in a range of contexts, on demand and to a high level of proficiency. The complementary focus is on competence as it is acquired and developed by students within their regular academic programs, and how that competence might be enhanced and assessed.

Underlying this orientation is the premise that each academic program and each academic course provides the most appropriate site for learning higher-order cognitive and other skills. This defines a key role for academics as educators and an aspiration for higher education as an enterprise which is central to attainment of academic aims and objectives. What are currently being labeled graduate attributes need to revert to being integral elements of academic learning, with performance in them ultimately being reflected in the course grades recorded on academic transcripts. The success of moving in this direction would depend directly on having not only competent academics but also an institutional commitment to sophisticated outcomes and high academic achievement standards.

REFERENCES

Abercrombie, M. L. J. (1969). *The anatomy of judgement: An investigation into the processes of perception and reasoning*. Harmondsworth, UK: Penguin.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise*. Chicago, IL: Open Court.

Dreyfus, H. L., & Dreyfus, S. E. (1984). From Socrates to expert systems: The limits of calculative rationality. *Technology in Society*, *6*, 217–233.

Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: SAGE.

Jones, A. (2009). Redisciplining generic attributes: The disciplinary context in focus. *Studies in Higher Education*, *34*, 85–100.

Polanyi, M. (1962). *Personal knowledge: Towards a post-critical philosophy*. London: Routledge & Kegan Paul.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy & Practice*, *14*, 387–392.

Sadler, D. R. (2009a). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education***,** *34*, 159–179.

Sadler, D. R. (2009b). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*, 807–826.

Sadler, D. R. (2010a). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*, 535–550.

Sadler, D. R. (2010b). Fidelity as a precondition for integrity in grading academic achievement. *Assessment and Evaluation in Higher Education*, *35*, 727–743.

Sadler, D. R. (2011). Academic freedom, achievement standards and professional identity. *Quality in Higher Education, 17*, 103–118.

Scriven, M. (1972). Objectivity and subjectivity in educational research. In L. G. Thomas (Ed.), *Philosophical redirection of educational research* (71st NSSE Yearbook, pp. 94–142). Chicago, IL: National Society for the Study of Education.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies: Theoretical and conceptual foundations* (pp. 45–65). Seattle: Hogrefe & Huber.

Wittgenstein, L. (1953). *Philosophical investigations* (Anscombe, G. E. M., trans.). Oxford: Basil Blackwell.

D. Royce Sadler
Teaching and Educational Development Institute
The University of Queensland

RICHARD J. SHAVELSON[1]

AN APPROACH TO TESTING & MODELING COMPETENCE

This paper presents an approach to measuring competence, and to statistically modeling the reliability and validity of the scores produced. To be sure, there are many possible approaches. By presenting this model, I hope to stimulate debate and data. My goal is to illustrate how the field of competency testing might develop the best possible working model of competence measurement through improving the model and measurements over time.

In my approach, competence is defined by a set of six facets. These facets carve out the domain in which measures of competence – their tasks, response formats and scoring – might be developed. Assuming an indefinitely large number of possible forms of a competence measure, a particular competence test may be viewed as a sample of tasks and responses from this large domain. Under certain reasonable assumptions, the assessment tasks/responses[2] and the raters who score test-takers' performance can be considered as randomly sampled. In such cases, a statistical theory for modeling the reliability and validity of competence scores, generalizability (G) theory, can be used to evaluate the quality of the competency measurement.

In sketching the model, I follow the now well-known assessment triangle (National Research Council, 2001): **cognition** or, more generally, the construct to be measured; **observation** of behavior; and the **interpretation** of observed behavior with inference back to cognition. First, then, I attend to the definition of the *construct*, *competence.* Then I turn to *observation*. This entails sampling tasks from the domain bounded by the construct definition. In this way, a test of competence is built. The intent is to produce an *observable* performance from which to infer competence. Finally, consideration is given to the *interpretation* of performance scores. To what extent are competence test scores reliable? To what extent are interpretations of competence test scores supported by logical and empirical evidence? These questions call for a combination of quantitative and qualitative studies. Throughout this paper, I provide concrete examples drawn from the fields of business, the military and education. I conclude with a summary of the model.

THE CONSTRUCT: COMPETENCE

The term c*onstruct* refers to an attribute of a person that is to be measured. In this case, the construct is *competence*. Competence, therefore, is an idea, a construction created by societies; it is not directly observable. Instead, it is inferred from

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 29–43.

observable performance on a set of tasks sampled from a domain of interest, such as a job or an educational discipline.

In broad terms, competence is a "… complex ability… that … [is] closely related to performance in real-life situations" (Hartig, Klieme, & Leutner, 2008, p. v; see also McClelland, 1973 and Weinert, 2001). More specifically, I (Shavelson, 2010a) identified six facets of competence from the literature: (1) *complexity* – a complex physical and/or intellectual ability or skill; (2) *performance* – a capacity not just to "know" but also to be able to do or perform; (3) *standardization* – tasks, responses, scoring-rubric and testing conditions (etc.) are the same for all individuals; (4) *fidelity* – tasks provide a high fidelity representation of situations in which competence needs to be demonstrated in the real world; (5) *level* – the performance must be at a "good enough" level to show competence; and (6) *improvement* – the abilities and skills measured can be improved over time by education, training and deliberative practice (see Shavelson, 2010a for details).

Tasks and responses that are included in competence measurement, therefore, should meet the following criteria:

1. Tap into complex physical and/or intellectual skills and…
2. Produce an observable performance using a common…
3. Standardized set of tasks with…
4. High fidelity to the performances observed in "real-world" "criterion" situations from which inferences of competence can be drawn, with scores reflecting…
5. The level of performance (mastery or continuous) on tasks in which…
6. Improvement can be made through deliberate practice.

Ideally, competence assessments would satisfy all six criteria. Practically, competence assessments will most likely tap into a subset of these criteria. Criterion 2, combined with the other criteria, emphasizes *constructed responses*, for example, an extended written response, a physical performance or a product. However, this criterion does not preclude the possibility that some portion of the assessment may include selected responses such as multiple-choice questions that will probably focus on the declarative knowledge that underpins competence. Criterion 4 is an ideal and the level of fidelity (low to high) may vary due to cost, time and logistical constraints. It seems that criteria 1, 3, 5 and 6 should be satisfied on any competence assessment.

In this chapter, I focus on assessments that meet, as closely as possible, all six criteria to a greater or lesser extent. Examples are drawn from several fields. As a first example of how the construct definition circumscribes an assessment (tasks, responses and scoring system), consider a measure of job performance, albeit an unusual one – a measurement of an astronaut's performance on general maintenance tasks on Earth and in lunar and zero gravity. Qualified astronaut-like participants performed tasks in three clothing conditions – shirtsleeves, deflated space suit and inflated space suit (Shavelson & Seminara, 1968). This study, the first of its kind, found a considerable performance decrement, measured by error rate and time, as participants went from Earth's 1 gravity to the moon's 1/6 gravity

to the zero gravity of space, and from shirtsleeves to deflated spacesuit to inflated spacesuit.

This performance assessment was not built as a measure of competence at the time, and the one attribute in the definition of competence that is missing is a criterion for the *level of performance*. With this exception, this assessment tapped into largely physical skills, produced observable performances in three gravity conditions and three clothing conditions, was a reasonably high-fidelity simulation of tasks which need to be performed in space and on the moon as defined by the National Aeronautics and Space Agency (NASA) for a lunar mission and performance could be improved by practice.

Thus far, I have picked the low-hanging fruit from the assessment tree. The parallel between job-performance measurement in high-fidelity simulations and the construct of competence as defined here seems obvious. How might competence be measured in less-well defined domains, such as education? I draw on two examples: (1) assessing performance in middle-school science; and (2) assessing 21st century skills in college students.

Ever since the Soviet Union put Uri Gagarin into orbit around the Earth on 12 April 1961, the U.S. has been in an endless loop of science education reform. One prevalent notion which emerged from the 1960s curriculum reform was inquiry-based science – students should be taught through inquiry, just as scientists inquire into the natural world in order to understand it. With a combination of textual materials and hands-on science investigations as the argument went, science can be learned better than by simply memorizing facts. This meant that in order to measure the outcomes of science education, something in addition to and more than multiple-choice tests of declarative and procedural knowledge was needed.

This curricular reform eventually led to the development of "performance assessments", in which students were provided with a problem and lab equipment, and asked to carry out an investigation. In the investigation, they would design a study, collect data and draw inferences in order to reach a conclusion. That is, students would do what scientists do, albeit in a much more limited way. To this end, performance assessments were designed for such topics as electric circuits, the control of variables, the identification of substances, the Earth-Sun relationship, forces and motion.

An Electric Mysteries assessment, for example, asked students to build electric circuits outside of "mystery boxes" in order to determine their contents – for example, the box might contain a battery, a wire, a battery and bulb or nothing (Shavelson, Baxter, & Pine, 1991). A Paper Towels assessment asked students to determine which of three different brands of paper towels held, soaked up or absorbed the most (and least) water (Shavelson et al., 1991).

Once again, with the exception of setting a criterion for the level of competent performance, these science performance assessments tapped into a combination of cognitive and physical skills and produced an observable performance. This performance was evaluated by raters who observed students' performance directly or from their responses in written science notebooks. The assessments were a reasonably high-fidelity simulation of the tasks and responses found in science

classroom inquiry activities; performance on these tasks could be improved by instruction and practice.

The final example of how the definition of competence constrains measurement is drawn from an ongoing project aimed at measuring college students' capacity to think critically, reason analytically, solve problems and communicate clearly (e.g., Shavelson, 2010b). The Collegiate Learning Assessment (CLA) samples real-world situations – for example, from education, work, civic engagement and newspapers – and asks students to solve a problem, recommend a course of action under conditions of uncertainty, and so on. For example, the "DynaTech" task asks students to evaluate the cause of an aircraft accident and to recommend a course of action for dealing with possible causes and negative press and perceptions. Students are provided with an in-basket of information regarding the aircraft and the accident to help them to reach and justify with evidence a decision on a course of action. Some of the information is reliable and some is not, some is related to the problem and some is not, some invites judgmental errors and some does not.

The CLA appears to satisfy the definition of a competence measurement. It taps into "21st century" cognitive skills; produces an observable performance that is evaluated by raters, either human or machine; it is a reasonably high-fidelity simulation of tasks found in everyday life, for example reading and evaluating a newspaper article; and performance on these tasks can be improved by instruction and practice. Finally, the CLA program provides information to colleges and universities that can be used to help set a standard for "competent" performance.

OBSERVATION OF PERFORMANCE

Observation refers to an individual's overt response to a sample of tasks from a certain domain (e.g., the job of an astronaut) on a measure of competence. The aim of the task sample is to elicit an observable performance. From the observed performance on the sample of tasks, an individual's level of competence can be inferred with greater or lesser accuracy. The construct definition, "competence in a domain," sets boundaries for which tasks fall within the domain and therefore which tasks fall outside of the domain. The universe of possible tasks and responses for observing performance, therefore, logically follows on from the definition of the construct. For the purpose of building an assessment, a sample of tasks is drawn from this universe in order to form the competence measurement.

Examples of Assessment Tasks

Consider the assessment of astronauts' performance (Shavelson & Seminara, 1968). The following steps were taken in order to build the assessment: (a) the performance domain – tasks and corresponding responses – was identified and enumerated from a lunar mission set by NASA; (b) "generic occupational tasks" were then enumerated. These were tasks that were required across a number of mission activities; (c) tasks were purposively sampled from the universe of generic tasks. That is, specific common tasks were systematically (not randomly) sampled

because they appeared in multiple activities and seemed to characterize these activities; (d) performance was observed on all tasks in all gravity conditions and all clothing conditions; (e) accuracy and time were measured; and (f) inferences were drawn from the task sample regarding performance in the domain of generic tasks.

We used a similar procedure to specify the universe of tasks and responses for science performance assessments. We began by: (a) identifying a domain of science investigations. To this end, we examined hands-on science materials, textbooks, teacher and student workbooks and so on. We then (b) sampled tasks from this domain. We drew purposive samples so as to produce an assessment that was highly representative of the kinds of activities which students carry out in inquiry-based science. Next, we (c) created a performance assessment from the tasks that fit within classroom space and safety restrictions. We (d) scored performance using trained raters. Finally, we (e) interpreted a student's score over tasks, raters, occasions and measurement methods as a reflection of the student's capacity to inquire.

As a third example, consider an assessment of military job performance (Shavelson, 1991; see Wigdor & Green, 1991). Assessment developers once again enumerated a universe of job tasks and took a sample from that universe. Specifically, the developers: (a) identified the universe of job tasks as specified in the military "doctrine" for a particular military occupational specialty (MOS), such as infantrymen; (b) sampled tasks from that domain—one question was whether the sample should be drawn purposively or randomly; (c) formed the task sample into a job performance test; and (d) scored performance using either objective evidence – such as an infantryman's accuracy in shooting targets in simulated combat situations – or expert judges' performance ratings. Finally, (e) they interpreted infantrymen's performance scores as representative of their performance over all areas of their job.

The Task Sampling Issue

The issue of how to sample tasks for an assessment of competence is important because it has critical implications for interpreting competence scores. The goal of competence measurement is to draw inferences from a person's performance on a sample of tasks regarding the person's performance on the entire universe of tasks in that domain (e.g., job, educational discipline). Scientifically speaking, statistical sampling is the preferred method. Sampling theory provides a method for sampling – simple random and more complex procedures – that ensures representativeness and provides a numerical estimate of the margin of sampling error.

However, performance measurements typically employ a small sample of tasks, and leaving the composition of the assessment to chance may, as many argue, often produce an unrepresentative test. The alternative is purposive sampling. With purposive sampling, complete control, rather than chance, is exercised over task selection. Based on the judgment of experts, for example, a sample of tasks can be selected that "looks" representative of the job.

The issue remains, however, of the representativeness of a purposive sample and how this representativeness can be measured. This issue is important because inferences about an individual's competence in a domain depend on the representativeness of the tasks that he or she performed in the assessment.

In the course of the military job-performance measurement project, I developed a method for evaluating the representativeness of a purposive sample against various forms of random sampling (Green & Shavelson, 1987). Consider the job of a Navy radioman. For each task in the job, incumbents rated the task with regard to: (1) whether they had performed it (PCTPERF); (2) how frequently they had performed it (FREQ); and (3) how complicated it was to perform (COMP). In addition, for each task, supervisors indicated (4) whether they had supervised the performance of the task (PCTSUP) and rated the task for: (5) its importance for the success of the mission (IMPORT) and (6) how often it was performed incorrectly (ERROR). From these data, the "universe" mean (μ) and standard deviation (σ) over all 124 job tasks could be calculated for each of the six ratings. This information provided the basis for specifying a sampling distribution.

Then, job experts drew a purposive sample of 22 tasks performed by radiomen. For these tasks, the sample means (m) and standard deviations (s) were calculated and compared to the universe parameters. Moreover, three random sampling schemes were identified for drawing 22 tasks: simple random sampling from an infinite universe; simple random sampling from a finite universe of 124 tasks; and stratified random sampling from a finite universe. Using the central limit theorem and sampling ratios, for each rating (e.g., ERROR), I calculated the distance (in σ units) between the purposive sample mean based on the selected 22 tasks and what would be expected from each of the random sampling methods.

It emerged that the purposive sample tended to include tasks that "looked like" the job (PCTSUP), and were performed frequently (PCTPERF). That is, the purposive sample included a disproportionate number of tasks that were performed frequently on the job. The purposive sample also included tasks that job incumbents rated as less complicated to perform than the average task (COMP). For this and other scientific reasons, the U.S. National Academy of Sciences urged the use of some form of random sampling for selecting tasks for job performance measurement. These sampling methods include stratified random sampling, whereby the most important tasks can be sampled with a probability of 1.00.

Task Sampling from Fuzzy Universes

There are many cases in which the task universe is not immediately evident, and so sampling from that universe is difficult, or perhaps impossible. Of course, a task can be created that looks like it belongs. However, without rough boundary conditions as to what constitutes a legitimate task with which to observe competence, it is difficult to infer conclusions with regard to the task universe. Instead, inferences can be made only as regards the universe of "convenient" tasks. Inferring competence in such a domain, therefore, becomes problematic and the validity of interpretations is suspect.

The CLA provides a case in point. It is not immediately obvious how to define the universe of 21st century tasks from which sampling might occur. However, upon reflection, the CLA specifies boundary conditions that are useful for defining this universe. The tasks selected should reflect *everyday situations* that arise when reading a newspaper or other informative text, engaging in civic activities, working at a job, working on personal finances, deciding which college to attend, visiting a museum, and so on. The student might be given a *document library* that provides the background and evidentiary basis for carrying out and responding to the task which has been set out. A major *constraint* in the CLA's universe definition is that the information provided in the document library must be comprehensible to any college student; the CLA measures generic critical thinking skills. The *utility of the documents* in the libray vary as to their: (a) validity – relevant or irrelevant to the task at hand; (b) reliability – some information is based on reliable sources and some not; and (c) susceptibility to error – some material may lead to the use of judgmental heuristics – mental shortcuts – that produce errors in judgment such as interpreting correlation as causation. Finally, the documents serve as the basis for the *product* of students' *deliberations*, such as solving a problem, deciding upon and recommending a course of action or characterizing sets of events along a series of dimensions.

Tasks generated by these constraints would be said to fall within the domain. These tasks could also be considered to have been randomly sampled from the vast universe of tasks that fall within the generic critical thinking and communication domain.

INTERPRETATION OF PERFORMANCE

Interpretation refers to the inferences drawn from an individual's behavior during a sample of tasks regarding what his or her behavior would be, on average, if he or she performed all of the tasks in the universe of possible tasks. That is, can one reliably and validly *interpret* (that is, infer from a person's performance on *a small sample of tasks*) the presence or absence of competence, or the level of competence *in the full domain*?

The question of reliability and validity is critical for several reasons. First, proposed interpretations of test scores have to be specified in some detail. Specifically, proposed interpretations of test scores need to be laid out in what Kane (2006) calls an *interpretive argument* – a chain of reasoning that leads from scores to claims of competence and decisions based on those scores. As Kane points out, interpretations can be complex and underspecified, making interpretative challenges difficult.

Second, following Kane, once an interpretative argument is laid out, the question arises as to what empirical evidence is needed in order to confirm or disconfirm the interpretative argument. He calls this the validity argument. There is also good reason for concern here, because competence measures will typically contain a small sample of tasks from a very large domain, and so a substantial sampling error might be expected. Any interpretation of performance as

measuring competence, therefore, will be accompanied by some degree of uncertainty. A statistical model is needed in order to evaluate the degree of uncertainty and error.

Reliability and Validity

Therefore, having created an assessment and observed and scored the person's performance on the sample of tasks, the question remains: do the scores actually (reliably and validly) measure competence?

Statistical models such as generalizability theory can estimate the degree of uncertainty and suggest how to reduce it. In what follows, I provide examples of the application of this theory to performance measurement. There are, of course, many other quantitative models for evaluating validity claims, both experimental and correlational, but their discussion goes beyond the purview of this paper.

Statistical models, however, are insufficient in themselves. Evidence is needed to support the claim in the interpretative argument that a person's observed performance involves the cognitive and physical skills and abilities which are believed to underlie competent performance. Hence, evidence of what I call *cognitive validity* is also required; evidence that the tasks evoke the kinds of thinking and reasoning that are part of the inference on which a judgment of competence is made (Ruiz-Primo, Shavelson, Li, & Schultz, 2001). Such evidence can be gathered through the "think aloud" method, whereby students verbalize their thoughts as they proceed through a task (Ericsson & Simon, 1993; Leighton, 2004). While the think aloud method provides important information on cognitive validity, its treatment is beyond the scope of this chapter (Ericsson & Simon, 1993).

STATISTICAL MODEL FOR COMPETENCE ASSESSMENT

The approach I have espoused for constructing a competence assessment can now be formalized statistically. A competence assessment contains a random sample of tasks. A person's performance on each task is observed on several occasions and scored by a set of randomly selected, well-trained raters. With this formulation, we are in a position to evaluate the dependability or reliability of the competence measurement statistically.

In addition, it might be necessary to include different methods for observing performance on a competence assessment. For example, in evaluating the performance of jet engine mechanics in the military job performance project, some tasks were carried out exactly as they are on the job. When it came to working specifically on a very expensive jet engine, a mistake would be very costly. Therefore, a "walk-through" method was used and the enlistees explained how they would carry out the task instead of doing the task.

By incorporating a methodological facet into the definition of the complex universe for observing performance, this formulation moves beyond reliability into

a sampling theory of validity. Specifically, the methodological facet represents all possible methods – for example, short answer, computer simulation, hands-on, walk-through, multiple-choice, video – that a decision-maker would be equally willing to interpret as reflecting a person's competence.

Once a person's performance has been conceived as a *sample* of performance from a complex universe the statistical framework of generalizability theory can be brought to bear on the technical quality of the competence assessment (Cronbach, Gleser, Rajaratnam, & Nanda, 1971; see also Brennan, 2001; Cardinet, Johnson, & Pini, 2009; Shavelson & Webb, 1991).

In concrete terms, consider the study of Navy machinist mates' job performance (Webb, Shavelson, Kim, & Chen, 1989). We examined the consistency of expert raters' real-time judgments of machinist's mates' performance on the assessment. In this case, *expert examiners* observed a machinist's mate (*person*) as he performed a sample of 11 job *tasks*. Two examiners scored each machinist's mate's performance on each of the 11 tasks. The total variability among these scores could be attributed to several sources. Scores may vary because of differences in the machinist mates' performances (person) – the variance the assessment was designed to measure. Alternately, scores may vary due to rater disagreement, task difficulty or a combination of the two. A random-effects model of the analysis of variance can then be used to partition and estimate the variance components statistically (Table 1).

Table 1. Generalizability of Machinist Mates' Scores
(Webb, Shavelson, Kim, & Chan, 1991, p. 137)

Source of Variance	*Estimated Variance Component (×1000)*	*Percent of Total Variation Due to Each Source**
Person (P)	6.26	**14.45**
Examiner (E)	0.00	0.00
Task (T)	9.70	**22.40**
P × E	0.00	0.00
P × T	25.85	**60.00**
E × T	0.03	0.00
P × E × T, error	1.46	3.37

*Over 100 percent due to rounding

The partitioning of the total variability in the scores can be found in the "Source of variance" column in the table. The magnitude of the variability in the scores contributed by each source in the assessment is shown in the "Estimated variance" column. The proportion of the total variability in the scores contributed by each source of variability is shown in the last column. This column provides a brief impression of the major sources of variability – desired or expected variability between persons – and error variability among the other sources of variability ("facets") of the measurement and in interaction with person.

The variability due to the person performing the task (14.45% of the total variability) was expected. Machinist mates vary in the level of their performance. Some are more competent performers than others.

The variability due to the examiner and the interaction between the examiner and the machinist mate was zero, contrary to expectations at the time. Raters did not introduce error into the measurement.

However, the variability due to the task was large (22.40%). This indicates that the sample of tasks in the assessment differed substantially in terms of the difficulty experienced by machinist mates in performing them.

Most importantly, the person x task interaction accounted for an enormous 60% of the total variability in the scores, also contrary to expectations at the time. The level of a person's performance depended on the particular task being performed.

The reliability of the scores using one examiner and 11 tasks was 0.72 on a scale from 0 (chance) to 1.00 (perfect reliability). Adding another examiner had no influence on reliability, as the examiners scored the performances consistently. However, by adding another six tasks, reliability was raised to 0.80.

The results of this study exemplify what has been found in job performance measurement and other domains such as education (e.g., Shavelson, Baxter, & Gao, 1993). At the time, these results and others on military performance measurement were surprising. Contrary to expectations, for example, examiners were able to rate Navy machinist's mates' performances reliably; they closely agreed in their scoring of complex performances in real time. Heretofore, examiner disagreement was expected to be a major source of measurement error in performance assessment.

Moreover, contrary to expectations, a very large degree of task sampling variability was observed. That is, the level of an incumbent's performance varied from one task to the next, and some tasks which were easier for certain machinist mates were more difficult for others. Generalized job expertise, therefore, may exist more in the eyes of the observer than in the observable performance itself. In addition, task sampling variability, not examiner sampling variability, was (and continues to be) a major concern in terms of cost, time and logistics.

A second example shows how measures that are intended to assess the performance of institutions rather than individuals can be modeled. The CLA provides measures of college performance and learning at an aggregate level – program, college or university. Students are sampled and respond to performance tasks and critical writing tasks. Random samples of tasks are given to random groups of students at each campus. Consequently, while scores are reported back to the students, the focus is on estimates of the generalizability of institutional scores. Interestingly, in this case, the students become a source of measurement error – when there are fewer students in the assessment sample, the mean estimate for that particular campus becomes less reliable.

The results of a G study of CLA performance tasks are shown in Table 2. Once again, we can see that task sampling variability gives rise to measurement error.

Raters do not make an important contribution to measurement error. The large final term reflects variability that has not been captured in the *school x task x judge* design. With six tasks and two judges, the level of reliability is 0.71. For the critical writing tasks, the level of reliability is well above 0.80.

Table 2. Generalizability of CLA performance task scores (Webb, Shavelson, & Steedle, in press)

Source of Variability	*Variance Component*	*Estimate*	*% Total*
School (*s*)	σ^2_s	817.47	20.90
Task (*t*)	σ^2_t	0[a]	0
Judge (*j*) 1.60	σ^2_j	62.50	
s × *t*	σ^2_{st}	671.42	17.10
s × *j* *1.60*	σ^2_{sj}	62.18	
t × *j*	σ^2_{tj}	0[a]	0
s × *t* × *j*, *e*	$\sigma^2_{stj,e}$	2305.77	58.80

Standard Setting

What distinguishes a competence measurement is that there needs to be a standard of performance above which a person is judged to be competent. The question, therefore, is "how much performance is good enough" to be judged *competent* in a particular domain? In my opinion, judgmental methods for standard setting are all problematic for a variety of reasons, not least because they are inconsistent, dependent on the method used and can be manipulated (e.g., Cizek, 2001; Haertel & Lorie, 2004; Rekase, 2000). While most competence measures employ some version of judgmental standard setting, in the long run, a concerted effort needs to be made to provide a more objective way of setting standards.

ACT's study of college readiness provides an example of objective standard setting (e.g., www.act.org/research/policymakers/pdf/benchmarks.pdf). In its study of U.S. students' college readiness, ACT located the score range on its college admissions test above which 50% or more of students earned a B or better grade point average at college in courses relevant to the competence domain of interest, for example mathematics.

SUMMARY: A MODEL OF COMPETENCE MEASUREMENT

This paper sketches a model for measuring competence and evaluating the quality of competence measurements. This is but one possible model. My intent with this model is to initiate collaboration among competence measurement researchers. The goal is to build one or more initial working models from which a more elaborate model can be built. In this way, I hope that research on competence measurement

builds on itself rather than taking diverse, divisive and non-comparable paths. The goal is to build better and better measurement methods and models over time.

The Model

In the model, competence is defined broadly as "… complex ability… that… [is] closely related to performance in real-life situations" (Hartig, Klieme, & Leutner, 2008, p. v). Competence is characterized by a set of six facets: (1) a complex physical and/or intellectual ability or skill that evinces itself in (2) overt performance on tasks that are (3) standardized across individuals, and may be conceived as (4) samples of real-life "criterion" situations (McClelland, 1973), for which a (5) level or standard of performance is identified to indicate competent performance, and (6) this competence is malleable and can be improved by education, experience and deliberative practice.

These facets define the domain in which measures of competence (tasks, responses and scoring) might be developed in order to form a competence measure. This chapter focused on constructed response tasks, but the construct definition does not preclude some selected response tasks, such as multiple-choice questions focused on declarative knowledge.

Assuming an indefinitely large number of possible tasks in the universe which can be used to define competence in a domain, a competence test is viewed as a sample of tasks from this large domain. Under some reasonable assumptions the tasks, responses and raters who score the resulting performances may be considered to have been sampled at random. With this assumption, G theory can be used to evaluate the quality of the competency measurement.

As most responses will be constructed by the test-takers, scoring will have to be done initially by humans and then perhaps subsequently by computers (Klein, 2007). This calls for the development of scoring rubrics to capture performance; these rubrics should also create a common framework for scoring performance across tasks in the universe.

The sampling framework which underpins this model of competence measurement leads to the statistical evaluation of the quality of the measures – their generalizability and interpretability – within the framework of G theory. This theory statistically evaluates the dependability of scores and can be used to determine the number of samples of tasks, human judges or occasions needed in an operational assessment in order to attain a reliable measurement of competence.

However, as noted, quantitative modeling can only go so far. It can address only parts of the "interpretative" argument for measuring competence. Questions as to whether an assessment taps into the kind of thinking that employs the hypothesized abilities and skills which underpin a competent performance, for example, demand additional evidence. Such questions need to be addressed with evidence of cognitive and consequential validity, and such methods are largely qualitative.

Limitations of the Model

The performance-oriented model sketched here certainly has limitations. Perhaps paramount among them is that, in order to obtain a reliable measurement, the assessment will need to contain multiple tasks (Ruiz-Primo & Shavelson, 1996). Performance tasks take longer than traditional multiple-choice and short-answer tasks. They are typically more expensive to build. They entail greater logistical demands than traditional tasks, and they are more expensive to score. Performance tasks of varying lengths as well as selected response tasks will be needed in order to address this limitation. However, care must be taken to ensure that selected-response tasks do not dominate the scores and ultimately the assessment.

There is a possibility that the model may under-represent the competence construct. Not only are cognition and performance involved in our notion of competence, but motivation and emotion ("dispositions") are also involved (Weinert, 2001). That is, competent performance requires motivated individuals to perform well, if not as well as possible. It also involves individuals whose identities are tied up in the tasks that they are competent in doing. Finally, our notion of competence implies the capacity to work with others and to take into account their perspectives.

To some degree, the use of high-fidelity simulated tasks incorporates these conative and affective characteristics of competence. Successful performance requires motivation and affect. However, an assessment is a simulation of real life; it is not real life. To what degree do the motivations and affect in a successful assessment performance overlap with real-world performances? To what degree should so-called "non-cognitive" measures be incorporated into the model? This said, can such measures be incorporated and yet at the same time *not* be susceptible to deception and social desirability?

Performance tasks are difficult to build, as considerable "know-how" goes into them. However, only a few people or organizations can build high-quality performance tasks. An infrastructure for building such tasks will need to be created.

Human scorers can be trained, as we have seen, to judge performance reliably. However, they are expensive and take considerable time to produce scores on a large scale. Computer technology is now at a stage at which it can score performance as reliably as human scorers in certain contexts. This technology may help to reduce the scoring limitation.

Concluding Comment

My hope, therefore, is that the model presented here, or another preferable model, will be adopted across research and development groups involved in measuring competence. By beginning with a "straw model", I hope to create a center of gravity so that new advances in one competence domain will inform measurement in another. The ultimate goal is to foster continuous improvement in both measurement methods and theories of competence.

NOTES

[1] I would like to thank Fritz Oser for his kind introduction, and the conference organizers, Sigrid Blömeke and Olga Zlatkin-Troitschanskaia, for inviting me to participate. I would also like to thank Christiane Kuhn for keeping me well informed about the conference and answering my various questions.

[2] "Task" refers to a situation, problem or decision to be made that is presented to the person taking the test. "Response" refers to action taken by the test taker as demanded by the task. The distinction is important, because tasks and responses indicate what behaviour is required in the criterion situation, and both tasks and responses can be fairly far removed from reality. For example, in multiple-choice questions, the stem (the material presented typically presented in a multiple-choice item before the alternatives are enumerated) typically presents a very short, synoptic task and the response is to choose from among four or five alternatives. Neither is a high-fidelity representation of most criterion situations in life. As referring to a task/response is awkward, I use the term "task" throughout the paper, but in doing so, I refer to both tasks and responses.

REFERENCES

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279–298.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying generalizability theory using EduG*. New York: Routledge/Psychology Press.

Cizek, G. (2001). *Setting performance standards: Concepts, methods, and perspectives*. New Jersey: Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

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

Green, B., & Shavelson, R. J. (1987). Distinguished panel discussion on issues in the joint-service JPM program. In H. G. Baker & G. J. Laabs (Eds.), *Proceedings of the Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies* (pp. 11–13). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel), 20301–4000.

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement, 2*(2), 61–103.

Hartig, J., Klieme, E., & Leutner, D. (2008). *Assessment of competencies in educational contexts: State of the art and future prospects*. Göttingen: Hogrefe & Huber.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Connecticut: Praeger, 17–64.

Klein, S. (2007). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In D. Nolan & T. Speed (Eds.), *Probability and statistics: Essays in honor of David A. Freedman*. IMS Collections (Vol. 2, pp. 76–89). Beachwood, OH: Institute for Mathematical Statistics.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6–15.

McClelland, D. C. (1973). Testing for competence rather than testing for "intelligence". *American Psychologist, 28*(1), 1–14.

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

Rekase, M. D. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements*. Washington, DC: National Assessment Governing Board, 41–70. Retrieved from http://www.nagb.org/publications/studentperfstandard.pdf.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*(10), 1045–1063.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99–141.

Shavelson, R. J. (1991). Generalizability of military performance measurements: I. Individual performance. In A. K. Wigdor & B. F. Green Jr. (Eds.), *Performance assessment for the workplace: Technical issues* (Vol. II, pp. 207–257). Washington, DC: National Academy Press.

Shavelson, R. J. (2010a). On the measurement of competency. *Empirical Research in Vocational Education and Training*, *2*(1), 43–65.

Shavelson, R. J. (2010b). *Measuring college learning responsibly: Accountability in a new era*. Stanford, CA: Stanford University Press.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–362.

Shavelson, R. J., Klein, S., & Benjamin, R. (2009). The limitations of portfolios. *Inside Higher Education*. Retrieved from http://www.insidehighered.com/views/2009/10/16/shavelson.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 61–71.

Shavelson, R. J., & Seminara, J. L. (1968). Effect of lunar gravity on man's performance of basic maintenance tasks. *Journal of Applied Psychology, 52*, 177–183.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133–166.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

Shavelson, R. J., & Seminara, J. (1968). Effect of lunar gravity on man's performance of basic maintenance tasks. *Journal of Applied Psychology, 52*(3), 177–183.

Webb, N. M., Shavelson, R. J., Kim, K.-S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinists mates. *Military Psychology*, *1*(2), 91–110.

Webb, N. M., Shavelson, R. J., & Steedle, J. T. (in press). Generalizability theory in assessment contexts. In C. Secolski & D. B. Denision (Eds.), *Handbook on measurement, assessment and evaluation in higher education*. New York: Routledge.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Seattle: Hogrefe & Huber.

Wigdor, A. K., & Green Jr., B. F.. (1991). *Performance assessment for the workplace* (Vol. I). Washington, DC: National Academy Press.

Richard J. Shavelson
Stanford University and SK Partners, LLC, USA

FRITZ OSER

"I KNOW HOW TO DO IT, BUT I CAN'T DO IT"

Modeling Competence Profiles for Future Teachers and Trainers[1]

ISSUES OF DEFINITION

There is an ongoing debate in higher education about the extent to which both a knowledge base and a competence profile are needed, and how these two elements can be successfully combined. In this context, it is necessary to be aware that a) competence is not the same as academic knowledge and b) academic competence is not the same as professional competence. With regard to b), solving a mathematical task for examination purposes is not the same as a) an engineer using mathematics to calculate the weight of a bridge to ensure pressure security. However, both a) and b) can be combined, in that a calculation of the bridge's weight could itself become an examination task. To have competence means to know how things work, whereas to perform successfully means to be able to demonstrate that competence. Both factors substantially depend on each other. However, b) alone can often create a situation in which professionals feel that they know how to do something but cannot actually do it.

Nevertheless a) is more complicated than b). It includes questions like "What constitutes the difference between knowledge and competence?" and "Why do we speak about competence profiles?" In order to be able to clarify this difference I use the case of a student in developmental psychology. This student needs to be aware of key factors of cognitive development, language development, moral development, physical development, motivational development, and perceptual development. Developmentalists require knowledge of stages (critical), phases, styles, developmental transformation models, and research techniques for longitudinal and cross sectional measurement, including developmental modeling techniques like discontinuity/continuity progresses, and special issues related to contingent life phases like childhood, midlife, and old age. Such knowledge can be acquired by studying the relevant textbooks. However, the development of a competence profile implies the ability to perform tasks like analyzing and identifying the language gaps of a first grader who is experiencing difficulties in expressing feelings, or diagnosing the social deficits of students in adolescence under situational peer group pressure, or applying the concept of the "unhappy moralist syndrome" (Oser & Reichenbach, 2000) to different age groups by using varying forms of testing. Some of these competence profiles relate to educational psychology and others to psychological counseling. Another good example would be treating people from three different age groups who are experiencing motivational difficulties in terms of their academic

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 45–60.

self-concept, with particular reference to the "big fish/little pond" effect. Such complex competence profiles are based on situations in which professionals need more than just knowledge. They also require a capacity for situational analysis, combining different forms of knowledge, creating action blueprints and finding effective ways of changing a situation.

A RESOURCE MODEL OF COMPETENCE PROFILES

This leads us to the question what a competence profile should include. Besides ethical, motivational and emotional aspects a competence profile encompasses a number of specific competences, for example as named above familiarity with language deficit correction programs and the ability to test adaptation capacities, as well as observational and perceptive skills which have been developed in various areas, like in child-care. Whereas the acquisition of academic knowledge includes only the process of learning material from canonical textbooks on developmental psychology, a competence profile is more complex. Even if a vast amount of scientific knowledge material exists, its effective actualization depends on how well those in the field can apply it. That is why the notion of competence profile can only be used if it can be applied to situations in which a professional who already has the relevant background knowledge is able to act. That means that the knowledge does not merely exist, but it is applied for solving concrete professional problems. A professional needs to be aware of the nature of the specific situation and be able to take relevant action. This involves being able to choose the appropriate action from a range of potential forms of such actions. In the case of developmental psychology this may include an awareness of the relevant developmental framework and the consideration of similar cases and may involve the selective application of existing rules, the formulation of diagnostic statements, and the planning of a practical program which takes these factors carefully into account. We speak about diagnostic and counseling competences. If knowledge is not applied in an appropriate manner, it leads to the problem that people who possess the knowledge do not know how to deal with it. And even if they know how to do it they cannot really do it.

Thus, in order to exercise a professional competence profiles successfully the professional will require:

a) more knowledge than he/she will actually need
b) additional situational, social and applicative abilities (such as learning climate adaptation, the ability to plan therapy for a child, and the ability to analyze systemic influences for a specific handicap)

A similar analysis of professional requirements can be applied to teacher training programs. As Shulman (1987) states, a teacher needs to acquire a range of different types of knowledge, including content knowledge (CK), pedagogical content knowledge (PCK), pedagogical knowledge (PK), management knowledge (MK) and developmental knowledge (DK). The teacher needs to acquire all these forms of knowledge, but must be aware that such knowledge is not in itself sufficient for

successful teaching. This is why, about 10 years later, this author created the notion of "signature pedagogy" to refer to typical professional situations which require special professional performative competences such as "bedside teaching" for medical doctors, "weight testing" for engineers, and "defending games" for lawyers. Such signature situations rely on specific competence profiles for particular performative occasions. There are certain situations that teachers need to be able to overcome by developing a relevant competence profile with a similar basic structure. In order to do this, at least two sources are needed, namely a source of academic knowledge bases and a source of practical field necessities. Both are complex and action bound. Thus, the term *competence profiles* is used to include many single actions and complex capacities.

One examples of a teacher competence profile based on these reflections would be: The teacher is able to organize different forms of group work which all students participate in and profit from and the result is integrated into the next phase of the teaching-learning process. Another example would be: The teacher is able to solve group conflicts between students in concrete daily classroom situations by forming roundtables and setting the criteria for realistic discourses (see Oser & Oelkers, 2001).

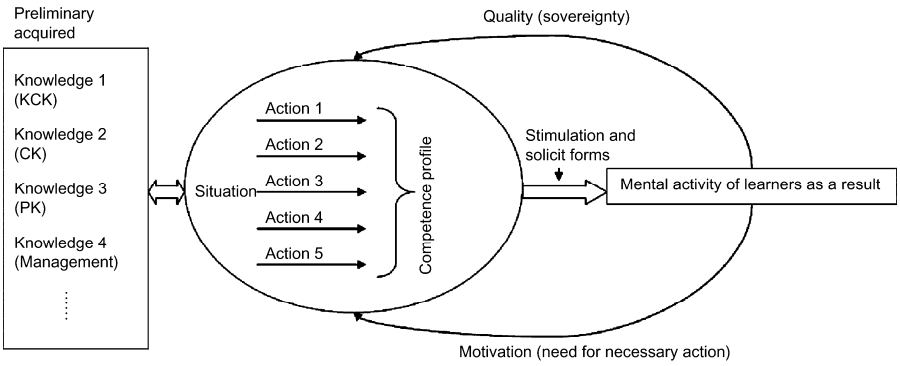


Figure 1. Resource-model of competence profiles (Oser, in prep).

Each competence profile mentioned is based on a resource-model of competence (see figure 1) which includes different single competencies that are connected through the situation in which these actions must take place, including also a sense of the necessity to act, as well as a measure of quality that is based on respective activity, namely the concept of reference. Reference means directedness to the stimulated inner activity of the learners, which consists of the so-called operations they perform as a result of that teaching. For each competence profile the same resource-model can be used. However, each model must be applied differently, particularly in terms of situation-specific knowledge, with situation-specific practical actions and a situation-specific sense of necessities. (This notion, which is specific for teaching, assumes that professionals must sense what the best thing to do is in any particular moment. They need to be able to make judgments

about such factors as presenting content, supporting learning, and providing such elements as scaffolding and reconciliation (Oser & Heinzer, 2010).

ON THE GENESIS OF PROFESSIONAL COMPETENCE PROFILES

In the belief that competence profiles applied by vocational education teachers (VET) can provide useful models for the formulation of relevant competencies, we collaborated with experienced vocational teachers. With the help of a Delphi-study (Häder & Häder, 2000; Brosi, Krekel, Ulrich, 1999) 45 competence profiles were produced that these teachers applied in daily teaching situations and were validated in a representative survey. The strategy was directed by a "bottom-up" process and included asking these experienced VET teachers to name central teaching situations that required the activation of such competence profiles. In total we used four rounds, with the first and the second rounds consisting of panel discussions to identify complex situations in the professional teachers' daily work. The third round consisted of a condensation of these situations into 45 competence profiles which we grouped according to plausibility statements into four main classes and nine subgroups (see Table 1).

Table1. Main classes of competence profiles for VET teaching (see also Heinzer et al., 2009)

Main-groups	*Sub-groups*
A Competence profiles of the teaching act itself	**A1** Preparation skills
	A2 Methods and styles of teaching
B Competence profiles of the learning environment	**B1** Social conditions for learning (social climate)
	B2 Value and conflict management, classroom organization
C Competence profiles for supporting learning	**C1** Diagnostic capacities
	C2 Monitoring skills
	C3 Evaluation abilities
D Accomplishment of vocational requirements and cooperation	**D1** Cooperation within the school and with the firms
	D2 Teacher's coping strategies

Three specific examples will be focused on here. Firstly, when "the teacher is able to organize learning situations, he/she gives clear and friendly directives for engaging in tasks, being able to keep each single student and each of his/her learning states in view". This competence profile falls under category C2 as "monitoring capacities of the teacher". Secondly, a situation in which "the teacher is able to provide supporting feedback – in critical situations when students give incorrect answers or have chosen an inappropriate strategy". This would indicate a C3 competence profile. Thirdly, a situation in which "the teacher can connect his teaching with what happens at the work place of the apprentice". This competence profile formulation falls under group D1. All these formulations include a group of

teaching actions that are guided by the situation. In the first example, the situation has to do with tasks that the teacher arranges. For instance, we can imagine that he/she sets the task of solving a given mathematical problem, interpreting a technical figure or analyzing a complex text that discusses forms of participatory democracies.

With regard to the competence of these five groups, we checked if they were a) related to concrete situations in the classroom or the firm, in which an apprentice acted, b) if they were part of a concrete learning chain (tailored to be relevant to this part of the lesson), c) if they were part of a cluster of professional actions, d) if there was a benchmark with respect to quality, and e) if there would be a possibility of chaining with respect to adjacent competence profiles.

The fourth round consisted of a validation ($N = 793$) with respect to the following criteria: 1) importance, 2) frequency of application, 3) difficulty of application, and 4) implications for teacher training in general. The sample consisted of 470 professional teachers (59%), including 204 teachers without a diploma (26%) and 115 non-teachers (15%). For the presentation here we chose only two examples (for others see Heinzer et al., 2009). Tables 2 and 3 elicit some surprising results. Preparing instructions and learning conditions were seen as the most important competence profile groups, although they were seen as being the least difficult. On the other hand, collaboration with colleagues and managing conflicts were seen as the least important, but the most difficult. However, these are only examples selected from a comprehensive study of teacher competences (Oser & Bauder, in prep.)

Table 2. Estimations of the importance of competence profile groups (see Heinzer et al. 2009)

Under groups of Competence Profiles	Mean	Standard Deviation
B1: learning conditions	2.684	.147
A1: lesson preparation	2.618	.068
D2: coping of the teacher	2.505	.065
A2: mediation forms	2.457	.118
C1: diagnosis	2.456	.138
C2: accompaniment	2.477	.621
C3: evaluation	2.317	.186
B2: value- and conflict management	2.315	.147
D1: cooperation with the College	2.18	.179

(0 not important at all, 1 rather not important, 2 rather important, 3 very important, m = 2.44, sd = 0.195)

Table 3. Estimation of realization difficulty (see Heinzer et al. 2009)

Under groups of Competence Profiles	Mean	Standard Deviation
D1: cooperation with the College	1.994	.164
C1: diagnosis	1.643	.055
B2: value- and conflict management	1.527	.155
D2: coping of the teacher	1.373	.189
C2: accompaniment	1.363	.231
C3: evaluation	1.353	.076
A2: mediation forms	1.33	.204
B1: learning conditions	1.274	.188
A1: lesson preparation	1.243	.105

(0 not important at all, 1 rather not important, 2 rather important, 3 very important, m=1.38, sd=0.21)

In using these examples, it is our intention to illustrate how we generated and tested the 45 competence profiles of professional teachers, with reference to their reasons for going into the field of teaching, as well as to collect ethnographically what teachers do, then to model these actions – in cooperation with the teachers themselves – into competence profiles. Finally we strategically grouped them into four, and later nine, competence groups. The most important step we took after completing this process was to go back and validate these competences by asking a comprehensive sample of teachers, non-teachers and special technical instructors about the necessity, the application frequency, the quality structure and the importance of these competences in the setting of teacher training.

A BOTTOM UP APPROACH

To summarize what has been said so far, behind the genesis of such competences there is a principle that is connected to the relationship between what we know theoretically about teaching and what actually happens in the field. We, the researchers, proceeded – as figure 2 suggests – from the "bottom up", connecting the realities in the field to theoretical reflections and then validating them from the top down. This entire procedure was repeated in terms of the Delphi study mentioned earlier, and with regard to such factors as observation studies and expert questioning. The basic idea was to focus on the elements, which professionals actually consider in their practical daily world.

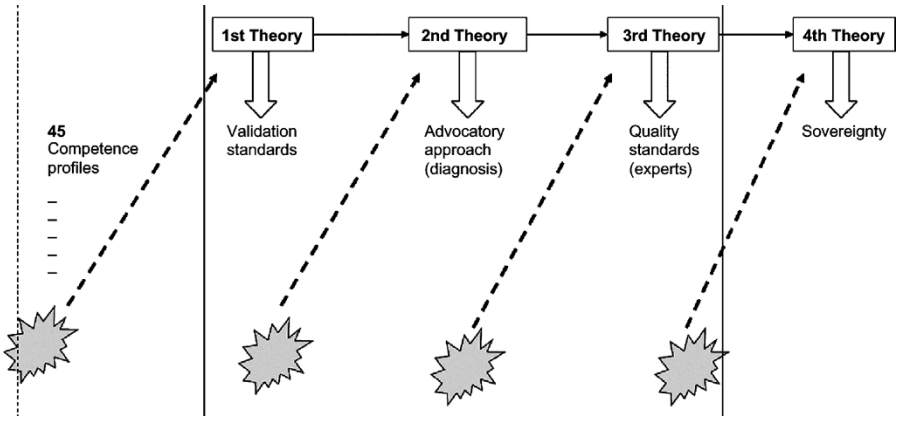


Figure 2. Theoretical elements of the bottom up approach: Delphi studies, advocatory approach, expert studies.

As shown in figure 2, a bottom up approach to the development of teacher competencies and a top down falsification through different modeling procedures were always used. This resulted in different theoretical elements generated in each phase. Firstly, there was the simple creation of competence profile formulations through the mentioned Delphi study and the respective validation questioning of a representative sample of VET teachers (including non-teachers and new teachers). Step two included filming of example situations in which competence profiles were required along with quality judgments of teachers carried out using an advocatory approach (see below). The third step consisted of the validation of the film vignettes by experts. These experts began their work by looking at the concrete teaching situations and then structured them by using classical quality criteria. A fourth step was the development of a "sovereignty measure" which was conducted by first looking into the field of modeling teacher competencies. For this we used what we call "daily simple action clusters" (rather than best practice, exceptional or extraordinary behavior).

THE AVOCATORY APPROACH: A VALIDATION OF TEACHING QUALITY

The avocatory approach is a method in which teachers (professionals) judge the competence profile of a colleague by means of a film vignette. This depicts a unit of a lesson which has a relatively closed form, and which can be said to be clearly distinguishable from other units. As figure 3 suggests, the judgment of the person responsible for rating the work gives hints about their capacity to judge others. The way that a person uses words in making this judgment shows their sensitivity to the professional issues in the situation (competence profile) depicted in the film.

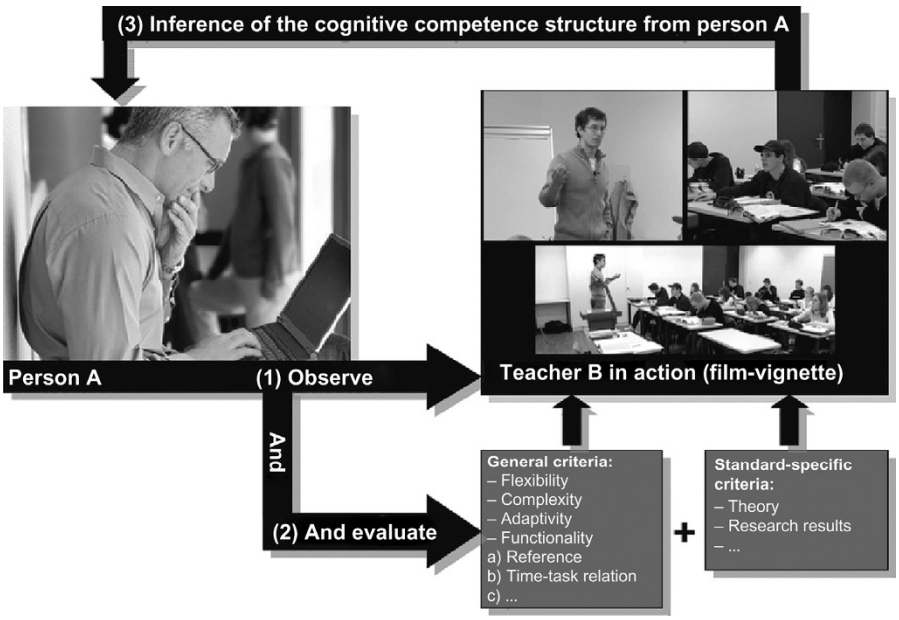


Figure 3. Scheme of the advocatory approach: Teachers judge teachers and thus indirectly elicit their own competences (see Oser et al. 2009).

As researchers, we then assessed the judging teacher, comparing him/her with a representative sample of colleagues and then stating what we think he/she recognized and what he/she did not observe. We then formed a quality judgment of the work of that evaluating teacher. In addition, since the teacher would be making judgments with respect to clear cut general and/or standard specific criteria (see table 4), we could compare this judgment with the judgment of other professionals, such as new or experienced teachers, non-teachers, teachers without diplomas, or with other experts. This would enable us to develop a sensitivity measure for professionals with respect to creating one single competence profile. As seen in figure 3, the teachers received an online questionnaire which presented the task of evaluating what they saw according to their best knowledge and experience, and to respond according to the specifics of the situation (a) and to general instructional criteria (b) (see table 4).
Table 5 represents an example of a comparison between teachers' and non-teachers' estimations. It became clear that all non-teachers estimated levels of quality at a significantly higher rate than teachers. Thus, in general they believe that what the teacher does is appropriate with regard to quality. This indicates that non-teachers demonstrated a weaker level of evaluating teaching issues. This may be because they remembered their time at school, but had no criteria to judge the teaching professionally. We did not identify differences between new and experienced teachers, but between teachers and non-teachers.

Table 4. Quality dimensions: specific and general (cross standard) criteria for evaluating a film vignette according to the advocatory approach

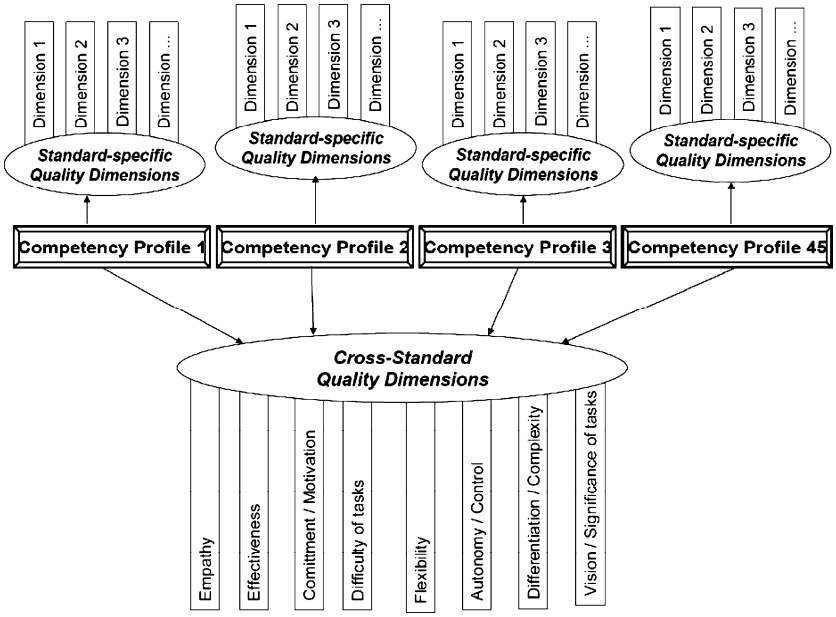


Table 5. Differences between teachers (LP) and non-teachers (N-LP) judging film vignettes according to general, cross-situational criteria

		Mean	Standard Deviation	*t*
Empathy	N-LP	4.4907	0.85761	2.93**.
	LP	3.9250	0.87004	
Effectiveness, division of time constraint	N-LP	4.1358	0.76939	2.32*
	LP	3.7375	0.77167	
Commitment, motivation (teacher)	N-LP	4.6543	0.98485	3.98**
	LP	3.7500	1.03184	
Commitment, motivation (trainees)	N-LP	3.8519	0.91793	2.58*
	LP	3.3750	0.79556	
Difficulty of tasks, Adequacy of taks	N-LP	4.2130	0.71288	0.53
	LP	4.1188	0.82579	
Flexibility	N-LP	4.3889	0.87217	2.37*
	LP	3.8979	0.94634	
Autonomy, control	N-LP	4.1259	0.76891	2.51*
	LP	3.6275	0.92872	
Differentiation, complexity	N-LP	4.0648	0.76458	4.06**
	LP	3.2813	0.89688	
Vision, importance of tasks	N-LP	4.7269	0.62600	2.66**
	LP	4.3063	0.73344	

* = p≤.05; ** = p≤.01

BENCHMARK SETTING AND EXPERT JUDGMENT

The last part of our program consisted of discussing benchmarks for each given criterion. This involved investigating whether there was a way of finding out which evaluation would be "right". This question is extremely important because the advocatory approach measures competence sensibility but does not measure performance accuracy.

It may be necessary to explain what we mean by "benchmark setting". This notion relies on the assumption that varying forms of competence realization exist. When observing a teacher, many people (even experts) believe they know how to assess them. However, they tend to disagree about what criteria to use to make such a judgment. It is possible to set certain benchmarks by calculating a mean average of the quality estimation of 600 or more teachers. It is also true that famous pedagogues, or certain charismatic teachers, have defined what they consider to be "good teaching". However, such definitions remain fundamentally unsatisfying, because they all somehow include a blind matrix, a random quality or unjustifiable positions.

We attempted to use a more quantifiable method of assessment. For each of the filmed vignettes, we invited three different experts (see advocatory approach) to participate. The first expert was one whose competence was mainly in the area of content knowledge. The second was a specialist in pedagogical content knowledge and the third was a teacher trainer who was also responsible for practical issues. They watched the film together, discussed each quality indicator exhaustively and were then asked to bring their evaluations of the quality to a consensus. The results are shown in figures 4 and 5, representing some examples of the differences.

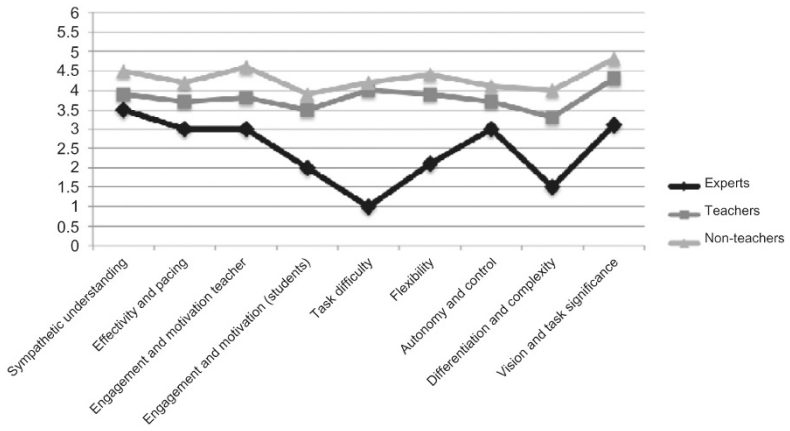


Figure 4. Example of quality estimations of teachers, non-teachers and experts regarding different cross-standard dimensions, targeting the competence profile "organizing powerful group work" (see Oser & Heinzer, 2010).

Figure 4 indicates that the judgment of experts tends – with respect to cross-standard dimensions – to be more severe than that of teachers and non-teachers. The teachers tended towards a more positive mean average and the experts towards a more negatively framed extreme.

In figure 5 we again present the mean values of teachers from different schools and the experts' judgments on "giving supportive feedback" with respect to standard-specific dimensions. The figure, astonishingly, yields a different result. The experts judge the indices in a more extreme manner, as being either substantially better or substantially worse than the teachers.

These two results made us aware that experts are either stricter (see figure 4) or that their judgment is more extreme in a more positive or negative way (see figure 5). This indicates that the benchmark setting of experts for quality judgments is substantially different from that of the professionals themselves. This can be seen as either a normative guideline, or as evidence of the need for change. – Competence profiles thus must be validated by different groups of users. This is in addition at least one way to begin to understand what kind of knowledge each competence profile contains.

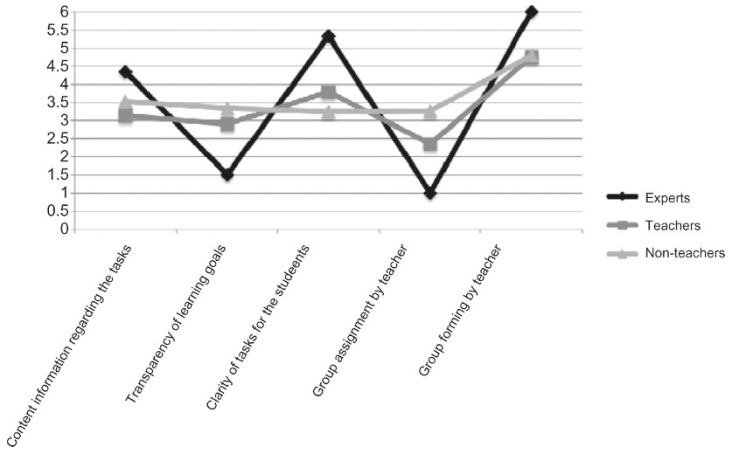


Figure 5. Results of the extreme experts' estimations compared to teachers' and non-teachers' judgments on standard specific dimensions: the case of giving feedback.

PROFESSIONAL FOUNDATIONS

Finally, it is necessary to ask ourselves about the legitimation criteria for each given competence profile. Apart from the above discussed Delphi study, in Fig.6, central elements are presented by which we can judge the competence profile of a teacher (VET teacher) as valid and reliable, namely a) entitlement, b) accountability, c) availability and d) professional status. All four are necessary for the inclusion of a competence profile into a new curriculum of the professional

competences that need to be demonstrated within each teacher training setting. How do we apply these four criteria to a concrete competence profile? As an example, let us look into the Pharmacy curriculum we helped to develop with the support of pharmacy professionals. The competence profile is:

> The vocational trainer can measure the trainee's level of responsibility in comparison to their year of learning. Thus the trainer can help trainees to estimate the quality of their own part-competencies and can help them expand them, so that the learner can progress slowly from controlled to autonomous actions.

Here are two examples that show the necessity of applying the four criteria:

> A customer had ordered a pharmaceutical product on the previous day. She had requested that the product be prepared early in the morning because she had to go to work. The apprentice forgot to get the product ready. Because of this the customer missed the bus, which caused her considerable annoyance.

Or:

> The pharmacy has a little online candy shop. The ordering and buying procedures are similar to those used for merchandise management in the pharmacy itself. The learner is given direct responsibility for the online shop.

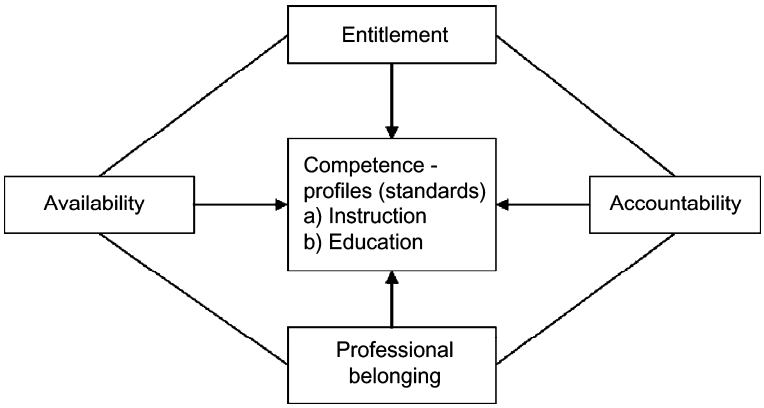


Figure 6. Professional criteria for declaring a competence profile as valid.

First, only the chief pharmacist has the right and the knowledge to distribute such responsibilities. She/he is entitled to do so (criteria a). It is her/his duty to estimate what a learner can do and what he/she supports. No outsider can ask for the same right. No one has even the right to tell her that she must distribute responsibilities. Because of her certification, he/she is the only one who publicly and officially has the right and the duty to judge what autonomy can be given to the apprentice and what must be taken under his/her control. So he/she in the first example demands

hard consequences from the apprentice, in the second he/she must have an eye on what the apprentice is doing. Both situations refer to complex competence profiles.

The accountability criteria b) go beyond those of a). If something happens to the apprentice, the training supervisor is fully responsible. The learner is not responsible but the trainer (teacher) has to take all the consequences and must be able to justify publicly why he/she acted in this way. We can argue that, because he/she is qualified, he/she is accountable. This issue is extremely important for measuring the competence of primary or secondary teachers who often think that the school they work in is accountable for their actions, or who may think that they are only accountable for their competence as instructors and nothing else. In identifying competence profiles for assessing professional competence, this needs to fall under the rubric of being responsible if it is applied. "To assign the learner adequate responsibility" belongs to this group because the teacher must be responsible for all the possible consequences.

Availability c) is the third criterion for choosing a competence profile as being absolutely necessary professionally. Teachers or training supervisors must be available for the student during the time he/she is in charge. Their role is a more cognitive presence, a form of participation in the other's existence. In this way, indifference is avoided. If a teacher or a training supervisor assigns responsibilities to a student, as formulated in our example, he/she clearly cares about the student's development. In caring he/she is available in a sense of always having an eye on what happens (see Watson et al., 1997; Noddings, 2002). Availability means not only "I am here if you need me" but "I am here as a part of your professional development".

The fourth criterion is professional belonging. If we choose a competence profile as being valid for the teacher training or the supervisor's training, we must recognize that the whole group of professionals in the same field accepts it as being necessary, including for instance the teachers' and pharmacists' unions. Medical doctors are strongly organized in professional groups, mechanics are strongly organized and teachers have their professional community. All these groups must accept the basic competence clusters of their own profession, and professional belonging means that members must also accept the respective standards which are being applied.

Thus, competence orientation – since it is more than knowledge orientation – is based on situations in which a cluster of professional acts must be adapted so as to change the respective situation precisely. It is necessary to discover, formulate and develop these competence profiles through a bottom-up process with the help of the respective professions. The justification of these competence clusters relies on what we may call the quadruplet transparency, namely the process of legitimation through entitlement, accountability, availability and professional belonging.

CONCLUSION

In this study, our intention was to show that competence profiles should be developed from the bottom up and theoretically modeled from the top town. They

must be formulated and validated with the help of professionals. As a further step it was useful to make these profiles visible by using film or story vignettes. Finally, it is only appropriate to judge them according to indices that are chosen in advance and are differentiated according to quality dimensions. This whole program was realized by the leading house "Professional Minds", and with it by an impressive group of young researchers (see footnote 1).

Our hypothesis is that this approach may help to overcome the fact that many of our university students leaving university say "I know how to do it but I can't do it". They refer to knowledge and maybe to imagined actions. In addition, the entire bottom up approach can lead to a better consensus with respect to a new competence oriented curriculum. It would be enrooted in both, in university knowledge and in professional situations. And the advocatory approach can be a reminder of what a practical job actually consists of. In this way professional knowledge becomes embedded in the professional field in which the real problems are generated.

What university professors do is transmitting knowledge. This is important, and it is necessary for structuring a knowledge field. But competence profiles and the necessity to develop them first give this knowledge a different meaning and give the student teachers a higher motivational framing.

We can surmise that similar approaches are necessary for all tertiary academic competence formulations. People must come together and must be urged to look at how they operate in their own practical fields to ensure that they are really competent. As stated earlier, each competence profile is constructed by combining many different single competences (fig. 1), and each competence requires specific knowledge. Thus when using our approach, knowledge is presumed, or with other words knowledge and competences come together. If the bottom up approach reveals it as being hidden, we will then know that in most cases textbook learning was the only way to lead the individuals towards certification. This would be a great pity. Coming back to the title at the beginning of this paper we must state that the entrance into the competence area opens up a huge application field for discovering basic acting.

NOTES

[1] Co-researchers in this project were S. Heinzer, T. Bauder, P. Salzmann, C. Joho, S. Grueter.

REFERENCES

Brosi, W., Krekel, W. M., and Ulrich, J. G. (1999). Delphi als ein Planungsinstrument zur Berufsbildungsforschung? Erste Ergebnisse einer BIBB-Studie In: *Berufsbildung in Wissenschaft und Praxis*, *6*, 11–16.

Häder, M., and Häder, S. (2000). Die Delphi-Methode als Gegenstand methodischer Forschung. In: M. Häder & S. Häder (Hrg.) *Die Delphi-Technik in den Sozialwissenschaften*. (11–31). Wiesbaden: Westdeutscher Verlag.

Heinzer, S., Oser, F., and Salzmann, P. (2009). Zur Genese von Kompetenzprofilen. In: *Lehrerbildung auf dem Prüfstand, 2*(1), 28–56.

Oser, F., and Oelkers, J. (2001). *Die Wirksamkeit der Lehrrbildungssysteme*. Zürich: Rüegger.

Oser, F., and Reichenbach, R. (2000). Moral resilience. What makes a moral person so unhappy? German version in: W. Edelstein & G. Nummer-Winkler (eds.). *Moral im sozialen Kontext*. Frankfurt: Suhrkamp, 203–233.

Oser, F., Bauder, T., Heinzer, S., and Salzmann, P. (2012). *Ohne Kompetenzen keine Qualität.* Klinkhardt (in print).

Oser, F., Heinzer, S., and Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. In: *Unterrichtswissenschaft, 38*(1), 5–28.

Oser, F., Salzmann, P., and Heinzer, S. (2009). Measuring competence-quality of vocational teachers: An advocatory approach. In: *Empirical Research in Vocational Education*, *1*(1), 65–83.

Noddings, N. (2002). *Educating moral people. A caring alternative to character education*. New York: Teachers College Press.

Shulman, L., (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*, 1–21.

Watson, M., Battistich, V., and Solomon, D. (1997). Enhancing students' social and ethical development in schools: An intervention program and its effects. *International Journal of Educational Research, 27*(7), 571–586.

Fritz Oser
Institute of Education,
Université de Fribourg, Switzerland

MARK WILSON AND KAREN DRANEY

A STRATEGY FOR THE ASSESSMENT OF COMPETENCIES IN HIGHER EDUCATION

The BEAR Assessment System

The Berkeley Evaluation and Assessment Research (BEAR) Center has developed an assessment system called the BEAR Assessment System (BAS), which is based on four principles of sound assessment (Wilson, 2005). In turn, each principle is associated with a practical "building block" that embodies the way in which the principle is used in an assessment context, and the whole system is brought together by an integrative activity that can take on different aspects under different circumstances (e.g., assessment moderation, cut score setting, etc.). Its original deployment was as a curriculum-embedded system in science education (Wilson & Sloane, 2000), but it has clear and logical extensions to other contexts, such as in higher education (Wilson & Scalise, 2006), large-scale assessment (Wilson, 2005) and disciplinary areas such as mathematics (Wilson & Carstensen, 2007) and chemistry in higher education (Claesgens, Scalise, Wilson, & Stacy, 2009).

The principles have been described in detail by Wilson and Sloane (2000) and Wilson (2005). The first of these principles is that assessment should be based on a developmental perspective. Second, there should be a match between classroom instruction and assessment. Third, instructors are the managers of the system in their classrooms. Finally, all forms of assessment should be based on high-quality evidence, in terms of both reliability and validity.

In the next section, we will provide a brief description of the large-scale assessment context in which we have been developing and applying aspects of the BAS. After that, we will describe the BAS in the large-scale context. Throughout this paper, we will discuss what our experiences have taught us regarding some of the salient issues concerning assessment.

THE GOLDEN STATE EXAMINATION PROGRAM

The Golden State Examination (GSE) program in the state of California consisted of a set of high school honors examinations.[1] GSEs were end-of-course examinations in a number of subjects, including mathematics (algebra, geometry, high school mathematics), language (reading and literature, written composition, Spanish

[1] Note that the GSEs were absorbed into the NCLB Tests for California.

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 61–79.

language), science (physics, chemistry, biology, coordinated science) and the social sciences (U.S. history, government & civics, economics). Each examination consisted of a set of multiple-choice items and several performance items.

Based on their scores on a particular GSE, examinees were categorized into one of six hierarchically-ordered performance levels – descriptive categories of student performance in each subject area. Figure 1 contains these categories for algebra as an example. The top three levels (4, 5 and 6) were considered "honors" levels (School Recognition, Honors and High Honors, respectively). If a student achieved one of these honors levels on six exams (including U.S. history, reading and literature or written composition, a mathematics exam and a science exam), the student was also eligible for a state honors diploma.

Although the origins of the GSE program predate much of our work on the BAS, we worked over several years to infuse the principles of our assessment system into elements of the GSE program, as evidenced by our work in cut score setting, scaling, written response item scoring and item development.

THE BEAR ASSESSMENT SYSTEM

Building Block 1: Progress Variables

Progress variables embody the first of the four principles: that of a developmental perspective on the assessment of student achievement and growth. The term "variable" is derived from the measurement concept of focusing on one salient characteristic to be measured at a time. A progress variable is a well-thought-out and researched hierarchy of qualitatively different levels of performance. Thus, a variable defines what is to be measured or assessed in terms which are general enough to be interpretable across a curriculum or state testing program, but specific enough to guide the development of other components. When instructional objectives are linked to the variable, it also defines what is to be taught. Variables are one model of how assessments can be integrated with instruction and accountability. Variables provide a way for large-scale assessments to be linked in a principled way to what students are learning in classrooms, while largely remaining independent of the content of a specific curriculum.

This approach assumes that, within a given curriculum, student performance on curricular variables can be traced over the course of the year, facilitating a more developmental perspective on student learning. Assessing the growth of students' understanding of particular concepts and skills requires a model of how student learning develops over a set period of (instructional) time. A growth perspective helps to move away from "one-shot" testing situations and cross-sectional approaches to defining student performance, toward an approach that focuses on the process of learning and on an individual's progress through that process. Clear definitions of what students are expected to learn and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material are necessary in order to establish the construct validity of an assessment system.

Level 6	Student work demonstrates evidence of rigorous and in-depth understanding of mathematical ideas; the work: • Is consistently correct and complete, and shows thorought understanding of mathematical content and concepts • Communicates clear and logical explanations of solutions to problems that are fully supported by mathematical evidence • Shows problem-solving skills that include appropriate generalizations, connections, and extensions of mathematical concepts • Includes effective use of mathematical language, diagrams, graphs, and/or pictures • Shows skillful and accurate use of mathematical tools and procedures, often with multiple and/or unique approaches
Level 5	Student work demonstrates evidence of solid and full understanding of mathematical ideas; the work: • Is essentially correct and complete, although it may contain minor flaws • Communicates explanations of solutions that are supported by mathematical evidence • Shows problem-solving skills that include connections and extensions of mathematical concepts • Shows appropriate use of mathematical language, diagrams, graphs, and/or pictures • Includes accurate use of mathematical tools and procedures
Level 4	Student work demonstrates evidence of substantial understanding of mathematical ideas; the work: • Is usually correct and complete, although it may contain flaws • Communicates explanations of solutions that are supported by mathematical evidence for most tasks • May contain evidence of problem solving without connecting or extending mathematical concepts • Includes frequent use of mathematical language, diagrams, graphs, and/or pictures • Usually shows evidence of appropriate use of mathematical tools and procedures
Level 3	Student work demonstrates evidence of a basic understanding of mathematical ideas; the work: • Is sometimes correct; however, it may lack either depth across the mathematical content areas or may show gaps in understanding of some concepts • Communicates explanations of solutions that are supported by mathematical evidence for some tasks, but explanations are very weak or missing for other tasks • May show ineffective or inconsistent problem solving • Shows some evidence of use of mathematical language, diagrams, graphs, and/or pictures • Shows some appropriate use of mathematical tools and/or procedures for some taske
Level 2	Student work demonstrates evidence of limited understanding of mathematical ideas; the work: • Shows little evidence of correct solutions and is incomplete • Provides limited explanations of solutions that are not supported by mathematical evidence • Shows limited evidence of problem-solving, arithmetic computations may be correct but unrelated to the problem • Shows limited evidence of use of appropriate mathematical language, diagrams, graphs, and/or pictures • Includes limited or inappropriate use of mathematical tools and procedures
Level 1	Student work demonstrates little or no evidence of understanding of mathematical ideas; the work: • Is rarely correct and has major mathematical errors • Provides little or no explanations of solutions • Shows little or no evidence of problem solving • Shows little or no evidence of the use of appropriate mathematical language, diagrams, graphs, and/or pictures • Includes little correct or appropriate use of mathematical tools and/or procedures

Figure 1. GSE performance level descriptions for algebra.

The use of progress variables also offers the possibility of significantly increasing the efficiency of an assessment: although each new curriculum prides itself on bringing something new to the subject matter, in truth, most curricula are

composed of a shared stock of content. As the influence of common national and state standards increases, this will become more true, and also easier to assemble into progress variables. Thus, we might expect even innovative curricula to have one or two progress variables that do not overlap with typical curricula, but the remainder will form a fairly stable set of variables that will be common across many curricula. For instance, examples of science progress variables included in a common focus might be: the ability to design a scientific investigation, gather data, and draw legitimate conclusions from the analysis of those data, as well as more standard science content variables. A common set including such variables could reasonably be the focus of a large-scale assessment program.

By building assessments related to these common progress variables, work done on one curriculum development may be much more easily transferable to others, leading to the efficiency mentioned above. This means that training for curriculum developers who will carry out this development can be done in a fairly generalized way that will allow them to adapt more quickly to new curricula. Of equal importance, schema for assessments, and even assessments themselves, that are developed for one curriculum, may be transferable or more easily adaptable to others with the same variables. This will also be helpful to instructors, who will not necessarily have to adapt to the same extent to new curricula when the curriculum changes.

Given a common "progress variables" basis for defining curricula, in a large-scale context, it would be possible (perhaps even fairly easy) to tailor a test to the curricula used in a particular system, by matching sets of items relating to the progress variables represented by those curricula. It should be possible to predict which progress variables would be affected by specific parts of a curriculum, and which should not, leading to the possibility of more highly-focused measurement in program evaluations.

Rather than focusing on an item-by-item or standard-by-standard content match, progress variables allow the matching of sets of tasks to over-arching frameworks. For example, if a variable such as "Designing and conducting investigations" is well represented in a state- or district-level assessment, one can be confident that content standards relating to that aspect of science inquiry are being measured by the assessment.

The idea of a progress variable need not be confined to a single curriculum, a single large-scale assessment system, or even a particular level of education. Various projects with which we have worked have developed progress variables in order to track the progress of the typical student from secondary school through various levels of postsecondary education. For example, the ChemQuery project (Claesgens et al., 2009) designed a set of progress variables in order to describe students' understanding of chemistry for chemistry students at both high school and university. The Carbon Cycle project (Mohan, Chen, & Anderson, 2009) is using a progress variable approach to investigate students' understanding of complex issues in global warming for students in upper elementary school through high school, and has recently developed a set of assessments for students at university level.

Another example, which is similar in nature to a progress variable, is the Common European Framework (CEF) for second language learning. This system has been developed in order to describe the proficiency of students and the difficulty of assessment tasks throughout the whole of the second language learning process, from secondary school through higher education. See Draney and Kennedy (2010) for an account of our work with a CEF-based English language assessment system in Germany.

Building Block 2: The Item Design Process

Assessment tasks create a match between classroom instruction and the various types of assessment. The critical element for ensuring this match in the BAS is that each assessment task is matched to at least one variable.

Explicitly aligning the instruction and the assessment also addresses the issue of the content validity of the assessment system. Traditional testing practices – in standardized tests as well as in tests created by instructors – have long been criticized for oversampling items that assess only basic levels of content knowledge and ignore more complex levels of understanding. Relying on progress variables to determine what skills are to be assessed means that assessments focus on what is important, not what is easy to assess. Once again, this reinforces the central instructional objectives of a course. According to Resnick and Resnick (1992), "assessments must be designed so that when teachers do the natural thing – that is, prepare their students to perform well – they will exercise the kinds of abilities and develop the kinds of skill and knowledge that are the real goals of educational reform" (p. 59). Variables that embody the aims of instruction (e.g., "standards") can guide assessment to do just what the Resnicks were demanding. In a large-scale assessment, the notion of a progress variable will be more useful to the parties involved than simple number-correct scores or standings relative to a norming population, providing the possibility of genuine diagnostic information, as is so often requested.

A variety of different task types may be used in this assessment system, based on the requirements of the situation at hand. There has always been tension in assessment situations between the use of multiple-choice items, which are perceived to contribute to more reliable assessments, and other, alternative forms of assessment, which are perceived to contribute to the validity of a testing situation. This tension can be summed up in a two-way diagram like the one in Figure 2 (which is based on one in Wilson & Adams, 1996). In this diagram, we can see that there are at least two kinds of possible forms of control over a testing situation: control over task specification (with the extremes being externally prescribed tasks vs. the "ad hoc" tasks that an instructor develops in order to meet the needs of students) and control over judgment (with the extremes being machine scorable vs. an instructor giving an overall rating or "grade" based on his or her perceptions). The point is not that testing situations with high or low levels of control are better, but that various tasks with varying levels of control must be designed in order to meet the range of assessment needs in classrooms, schools and districts.

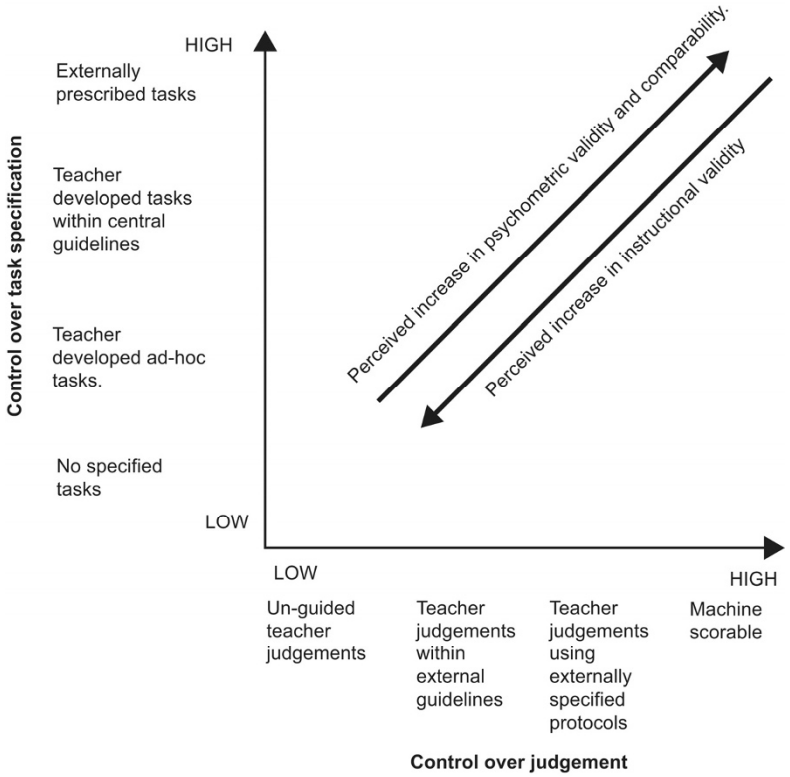


Figure 2. Perceived relationship between control, reliability and validity.

When using this assessment system within a curriculum, a particularly effective mode of assessment is what we call "embedded assessment". By this we mean that opportunities to assess student progress and performance are integrated into the instructional materials and are virtually indistinguishable from day-to-day instructional activities. We found it useful to think of the metaphor of a stream of instructional activity and student learning, with the instructor dipping into the stream from time to time in order to evaluate student progress and performance. In this model or metaphor, assessment then becomes part of the teaching and learning process, and we can think of it as being "assessment for learning" (Black, Harrison, Lee, Marshall, & Wiliam, 2003). If assessment is also a learning event, then it does not take unnecessary time away from instruction, and the number of assessment tasks can be increased more efficiently in order to improve the reliability of the results (Linn & Baker, 1996). However, in order for an assessment to become fully and meaningfully embedded in the teaching and learning process, the assessment must be linked to a specific curriculum, i.e., it must be curriculum-dependent, not curriculum-independent, as must be the case in many high-stakes testing situations (Wolf & Reardon, 1996).

In embedded assessments in classrooms, there will be a variety of different types of assessment tasks, exactly as there is a variety of instructional tasks. These may include individual and group "challenges", data-processing questions, questions following student readings and even instruction/assessment events such as "town meetings". Such tasks may be open-ended, requiring students to explain their responses fully in order to achieve a high score, or they may be multiple-choice, freeing instructors from having to laboriously hand-score all of the students' work (e.g., see Briggs, Alonzo, Schwab, & Wilson, 2006).

There are many variations in the way in which variables can be realized in practice, from using different assessment modes (multiple choice, performance assessment, mixed modes, etc.), to variations in the frequency of assessment (once a week, once a month, etc.), to variations in the nature of the embedding of the assessments (all assessments embedded, some assessments in a more traditional testing format, etc.).

In large-scale testing situations, the basis on which the mix of task types is decided may be somewhat different from that in embedded assessment contexts. Many large-scale tests are subject to tight constraints, both in terms of the time available for testing, and in terms of the financial resources available for scoring. Thus, although performance assessments are valued because of their perceived high validity, it may not be possible to collect enough information through performance assessments alone to estimate each examinee's proficiency level accurately; multiple-choice items, which require less time to answer and which may be scored by machine rather than by human raters, may be used to increase the reliability of the large-scale test.

For example, the GSEs each contained a set of multiple-choice items and at least one open-ended item. Most of the exams contained two written response items, each designed to take approximately 20 minutes to answer, and each was scored with a scoring guide containing 4 to 6 points (the number of score points varies by subject area). In addition, for the science exams, the GSE program used an analytic scoring system in which each written response is assigned multiple component scores, as opposed to a single holistic score.

Building Block 3: The Outcome Space

The outcome space is the set of ordered categorical outcomes into which student responses are categorized for the items associated with a particular progress variable. In practice, these are presented as scoring guides for student responses to assessment tasks, which are the primary means by which the essential element of expert professional judgment is implemented in the BAS. These are supplemented by exemplars: examples of student work at every scoring level for every task and variable combination, and blueprints, which provide the instructors with a layout showing opportune points in the curriculum at which to assess the students on the different variables.

In order for the information from assessment opportunities to be useful to instructors, it must be couched in terms that are directly interpretable with regard

to the instructional goals of the variables. Moreover, this must be done in a way that is intellectually and practically efficient. Scoring guides have been designed to meet these two criteria. A scoring guide serves as a practical definition of a variable by describing the performance criteria which must be met in order to achieve each scoring level of the variable.

Scoring guides may be structured in a variety of ways. One way we have approached is to start with scoring guides generated from a common underlying structure, i.e., the Structure of the Learning Outcome (SOLO) taxonomy (Biggs & Collis, 1982). Figure 3 shows the SOLO taxonomy as it has been used for the initial development of several applications of the BAS. These levels are then adapted to the specific needs of the curriculum.

An *extended abstract* response is one that not only includes all relevant pieces of information, but extends the response to integrate relevant pieces of information not in the stimulus.
A *relational* response integrates all relevant pieces of information from the stimulus.
A *multistructural* response is one that responds to several relevant pieces of information from the stimulus.
A *unistructural* response is one that responds to only one relevant piece of information from the stimulus.
A *pre-structural* response is one that consists only of irrelevant information.

Figure 3. The SOLO taxonomy.

The scoring guides are meant to help make the performance criteria for the assessments clear and explicit (or "transparent and open", to use Glaser's (1990) terms) – not only to the instructors, but also to the students, administrators, and/or other consumers of assessment results. In fact, we strongly recommend to instructors that they share their scoring guides with their students, as a way of teaching students what types of cognitive performance are expected and to model the desired processes. While a little uncomfortable with this at first ("Isn't that 'teaching to the test' or 'giving students the answers'?" some ask), many instructors have found that explicit discussions of what they expected and of how students could improve their performance can be a useful pedagogical tool. In some classrooms, instructors have taught students to score their own (or their partners') work using modified scoring guides.

In addition, students appreciate the use of scoring guides in the classroom. In a series of interviews with students in a Kentucky middle school that was using the BAS (reported in Roberts & Sipusic, 1999), the students spontaneously expressed to us their feeling that, sometimes for the first time, they understood what their teachers expected of them, and felt that they knew what they were expected to learn:

> She gave us a chance to see what we did and what we did wrong. You really can understand the work you're doing.

They also found out what other students were thinking:

> You learn how different students can have different scores even though they're from the same classroom and have the same teacher. You can see what their understanding and knowledge is and you can compare it to your own understanding and knowledge.

The teachers of these students found that the students were often willing to redo their work in order to merit a higher score.

As there will be inevitable questions of interpretation when applying a scoring guide to a particular task, especially for instructors who are new to using the assessment system, we recommend supplementing scoring guides with what we call exemplars: pieces of student work at each possible level of performance on individual assessments. Exemplars provide concrete examples of students' work for the various scoring levels. These are actual samples of student work, selected by experienced instructors in order to illustrate typical responses for each scoring level for specific assessment activities, and accompanied by brief explanations of what to note.

The idea of scoring guides is not new in large-scale testing; however, "rubrics" are often written to be item-specific, rather than based on a more general underlying structure. In addition, a form of exemplar (referred to in the GSE program as an "anchor paper") is fairly often provided for the raters of written response items in large-scale testing contexts.

The existence of scoring guides can be an advantage even when there is no explicit need for them. Multiple-choice items do not need a scoring guide for scoring, but something very similar to a scoring guide is important when developing multiple-choice items, for both the question itself and the distractors. For example, it may be helpful to develop the distractors in such a way that they represent various levels of incomplete or incorrect understanding, in order to appeal to examinees whose understanding is at that level. The items may even be scored this way, as in ordered multiple-choice items (Briggs et al., 2006). Of course, the development of a scoring guide should be an essential step in developing open-ended prompts. In addition, the use of scoring guides by instructors can lead to the internalization of the progress variable, which enables an instructor to use it in informal as well as formal settings.

Building Block 4: Wright Maps

Wright maps[1] represent the principle of high-quality evidence. A *Wright* map is a graphical and empirical representation of a progress variable, showing how it unfolds or evolves over instructional time in terms of student performance. It is derived from empirical analyses of student data on sets of assessment tasks. It is based on the empirical ordering of these assessment tasks from relatively easy tasks to more difficult and complex ones. A key feature of such a map is that both students and tasks can be located on the same scale, thereby facilitating the substantive interpretation of student proficiency, in terms of what the student knows and can do and where the student is having difficulty.

We typically use a multi-dimensional Rasch modeling approach to calibrate maps for use in the BAS (see Adams, Wilson, & Wang, 1997 for the specifics of this model). A Wright map has several advantages over the traditional method of reporting student performance as total scores or percentages: first, it allows instructors to interpret a student's proficiency in terms of average or typical performance on representative assessment activities; and second, it takes into consideration the relative difficulty of the tasks involved in assessing student proficiency.

Once constructed, a Wright map can be used to record and track student progress and to illustrate the skills a student has mastered and those that the student is working on. By placing students' performance on the continuum defined by the map, instructors can demonstrate the students' progress with regard to the standards that are inherent in the progress variables. Such maps, therefore, are one tool with which to provide feedback on how the class as a whole is progressing. They are also a source of information to use in providing feedback to individual students on their own performances.

Wright maps come in many forms, and have many uses in classrooms and other educational contexts. In order to make the maps flexible and convenient enough for use by instructors and administrators, we have also developed software for instructors to use in order to generate their own maps. This software, which we call ConstructMap (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2008), allows instructors and other users to enter the students' assessment scores, and then maps the performance of individual students, either at a particular time or over a period of time. In addition, one can map the performance of a class or larger group on a given set of assessments.

Wright maps have other uses beyond the classroom. The maps can be very useful in large-scale assessments, providing information that is not readily available through numerical score averages and other traditional summary information. An excellent example of the type of information available through progress maps can be found in a report on Australia's National School English Literacy Survey (Masters & Forster, 1997). This report uses maps to display levels of student achievement in writing, reading, viewing, speaking and listening skills. The level definitions are based on the analysis of empirical data from portfolios of written work from a nationally representative sample of students in grades 3 and 5 in Australia. Two maps from this study are shown in Figures 4(a) and 4(b). Each of the levels in the map in Figure 4(a) is described by skills that are typical of a student performing at this level, and that range from the easiest to the hardest for a child to master. For example, of the language indicators on the writing scale, the easiest skills include: "Uses some correct initial letters and other sounds" and "Can be read back by the child at the time of writing". The most difficult skills include "Experiments with rearranging sentences" and "Revises writing to be consistent in content and style".

Such a map can be used for a variety of purposes, including summarizing the average and range of student performance at each grade level, and investigating the differences between subgroups. As the numerical averages and ranges for groups of students correspond to regions on the map, which in turn are defined by skills which are typical of those regions, this gives the differences between these groups a substantive interpretation. This is illustrated in the map in Figure 4(b), which shows

the distributions of students in grades 3 and 5 in terms of their location on the map. For example, in the two-year span between grade 3 and grade 5, the average performance of students increases from just above Level 2, at which they had mastered such skills as "Uses simple sentences" and "Uses repetitive sentence structure", to the upper regions of Level 3, at which they were mastering such skills as "Controls simple sentence structure and attempts more complex structures".

Wright maps have also served as the central metaphor in a cut-point setting, scaling and linking system designed for the GSE (Wilson & Draney, 2000; Wilson & Draney, 2002). Within this system, they have been used both to understand the structure of the examinations, and to set the cut points for student achievement that define the different performance levels for each examination.

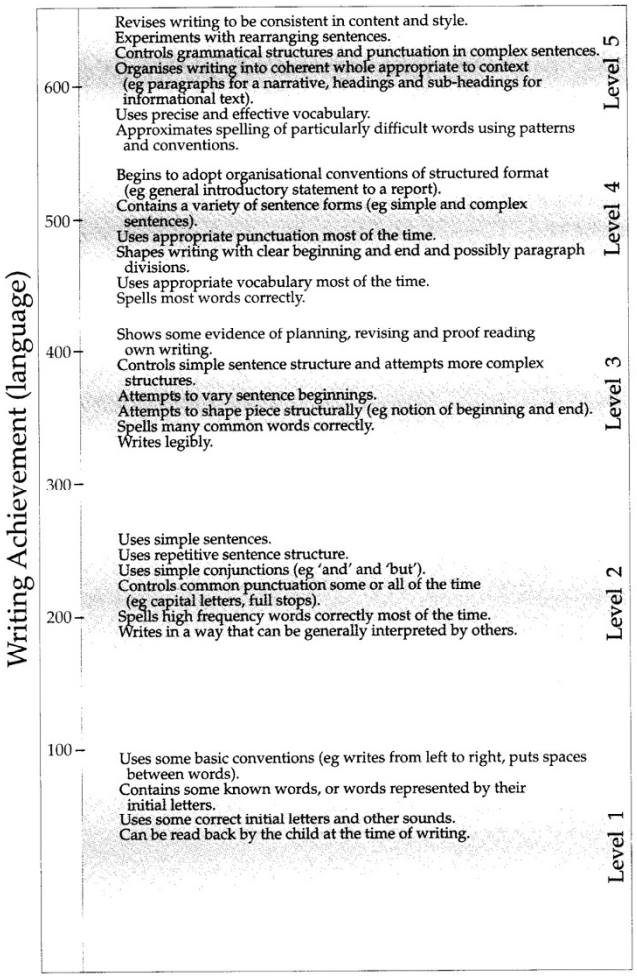


Figure 4(a). Map from the Australian National Literacy Survey.

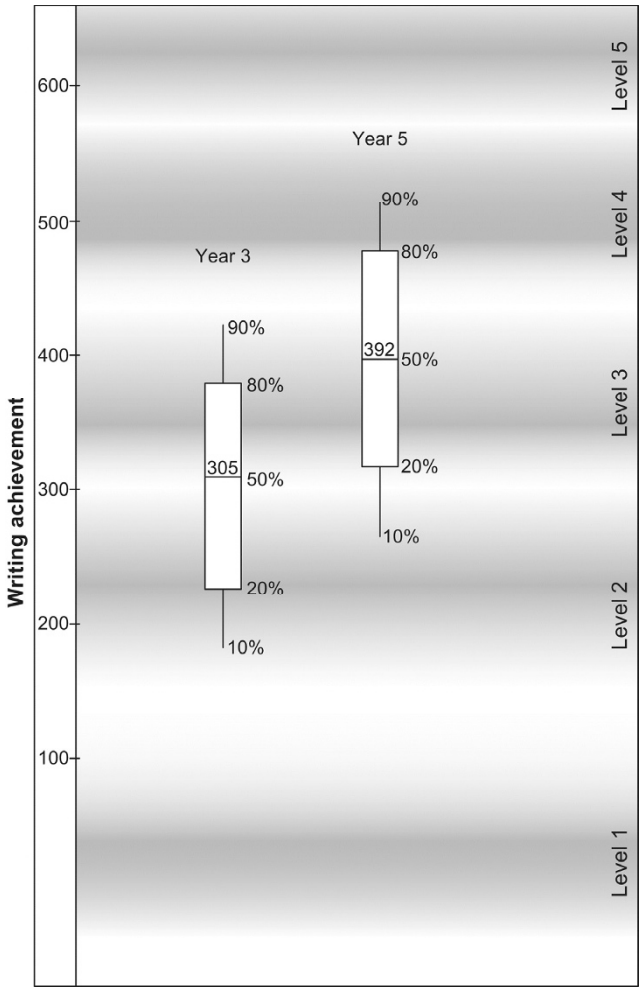


Figure 4(b). A second map from the Australian National Literacy Survey.

Figure 5 shows a map from the GSE in Economics. This map more closely resembles the traditional item and student map used in item response modeling.

Logit scale	Persons: Percentile	Persons: Histogram	Items: Fundamental Concepts	Items: Comparative Micro	Items: Comparative Macro	Items: Comparative Systems	Items: Comparative International	Items: Essay
2.20	99.9							
	99.9							5
	99.9	<						
	99.8	<						
1.80	99.4	<						
	99.2	<						
	99.1	X						
	98.5	X						
1.40	98.0	XX						
	97.2	XX						
	96.1	XXX						
	94.8	XXXX						
1.00	93.3	XXXX						4
	91.5	XXXXXX						
	88.9	XXXXXXX						
	86.1	XXXXXXXXXX		S	S			
.60	82.2	XXXXXXXXXX	K		A K			
	78.1	XXXXXXXXX						
	74.5	XXXXXXXXXXX		A	S S			3
	69.6	XXXXXXXXX	K	A	A			
.20	65.6	XXXXXXXXXXXX	A		A	A		
	60.2	XXXXXXXXXXXX	A	A S	A A		S	
	54.7	XXXXXXXXX			A K K	A S		
	51.2	XXXXXXXXXXX		A K S	K		S	
-.20	46.6	XXXXXXXXXXXXX		A				
	40.6	XXXXXXXXXXX	A	A				
	35.9	XXXXXXXXXXXXXX		A S			K S	
	29.3	XXXXXXXXXX	A K	A	A		A	
-.60	25.0	XXXXXXXXXXXXX	K				A	
	18.5	XXXXXXXXX						
	14.6	XXXXXXXX		A		A		
	11.0	XXXXXXXXX		A				
-1.00	7.2	XXXXX						2
	4.7	XXX	K					
	3.1	XXX						
	1.8	X					A	
-1.40	1.2	X						
	0.7	<						
	0.4	<						
	0.2	<						
-1.80	0.2	<						
	0.1	<		A				
	0.1	<						
	0.1	<						
-2.20	0.0							

S=Synthesis A=Application K= Knowledge

Figure 5. Map of Economics GSE by strand.

For the test represented in this map, there were 50 multiple-choice items and a single written response item scored on a scale of 1 to 5. The multiple-choice items represented five "strands" or important topic areas within the field of economics, represented by the columns under these headings: fundamental concepts;

microeconomics; macroeconomics; comparative systems and international economics. In addition, these items were designed to represent three different processes or areas of thought: knowledge, application and synthesis. Items representing these processes are labeled with a K, an A or an S, respectively.

From this representation, there are a number of things that can be learned about this examination. For example, it appears that the items on the comparative systems and international strands are, on average, somewhat easier than items on the other three strands. In addition, within each strand, the synthesis items tend to be on the hard side of the scale; no synthesis item falls below the horizontal line drawn at approximately -0.5 logits, although there are a number of knowledge and application items below this line. Item developers can examine item performance in this way in order to determine whether items representing the varying strands and processes are performing in accordance with their expectations.

Another important thing to note that is made clear by the Wright map is that proximal information about the upper end of the scale is provided primarily by the level of the written response item; the multiple-choice items cluster around the middle of the student distribution, and none appear to be above approximately 0.75 logits, or around the 86th percentile. This is especially important because the GSEs are honors examinations. The upper three performance levels are the ones for which students receive commendation; being in the lower three performance levels has few, if any, consequences. In general, only around 30% of students tend to fall into one of these top three performance levels; often, only 5% fall into the "high honors" level. Thus, this representation shows one reason why it is important to have written response items on examinations such as these – they provide us with information about parts of the scale that are not well measured by multiple-choice items.

Bringing the Building Blocks Together

The final ingredient in the BAS is the means by which the four building blocks discussed thus far are brought together into a coherent system – in this case, represented by assessment moderation and standard-setting. First, we will discuss assessment moderation. Moderation is the process by which instructors discuss student work and the scores they have given that work, making sure that equivalent scores are being awarded for similar work, and that scores are being interpreted in the same way by all instructors in the moderation group. Clearly, moderation is needed only for open-ended items. However, this process also helps to clarify the usefulness and validity of multiple-choice items by enforcing the meaning of scores through the sorts of Wright maps discussed above. In moderation sessions, instructors discuss the scoring, interpretation and use of student work that they all have read, and make decisions regarding standards of performance and methods for judging student work reliably in relation to these standards. Most importantly, moderation sessions provide an opportunity for instructors to discuss the implications of the assessment for their instruction, for example, by discussing ways to address common mistakes or difficult concepts in their subsequent instruction.

The assessment moderation process provides a vehicle for instructors to learn about how the assessment system works as a whole, to reflect upon their students' performance, to practice using the system and to learn to rely on a network of supportive colleagues who are also struggling with how to value and diagnose student work. The process is a continuous one, allowing instructors to learn, to apply, to reflect and to begin the cycle anew on a regular basis. It is hands-on, it models what the instructors are expected to do on their own, it is collegial and it must be sustained by the supply of suitable resources. As such, assessment moderation is a model of professional development in assessment that meets many of the standards proposed for professional development in curriculum and instruction (e.g., Frechtling, 1997; Ruskus & Luczak, 1995). It is costly in terms of the instructors' time, however, and special arrangements must be made to accommodate these costs, as with other forms of quality staff development. However, our experience suggests that, compared to the typical isolated workshops in assessment techniques or the provision of written materials alone, this professional development approach for instructors can help them to improve their students' performance (Wilson & Sloane, 2000).

Now we turn to the issue of cut-score setting, a function in large-scale assessment that in many ways parallels the role of assessment moderation in classroom assessment. For example, the most important use of maps in the GSE program has been for setting the cut scores between the six performance levels. The method we have developed, called "construct mapping" (Wilson & Draney, 2002), allows the standard-setting committee members to use the item response scale as a model of what a student at a given level knows and can do. The two types of items are scaled together in order to produce difficulty thresholds for the items and proficiency estimates for the students. This calibration is then used to create an item map combining the locations of the multiple-choice items and the score levels of all of the performance items. This map is represented in a piece of software (Wilson et al., 2010) that allows committee members to find out about the details of student performance at any given proficiency level, and to assist them in deciding where the cutoffs should be between performance levels.

An example showing a section of such a Wright item map is given in Figure 6, which uses the Written Composition GSE as the example. The column on the far left contains a numerical scale that allows the selection and examination of a given point on the map, and the selection of the eventual cut scores for the performance levels. This scale is a transformation of the original logit scale, designed to have a mean of 500, and to range from approximately 0 to 1000. The next two columns contain the location of the multiple-choice items (labeled by their order of appearance on the examination) and the probability that a student at the selected point would answer each item correctly (in this case, a student at 500 on the GSE scale – represented by the shaded band across the map). The next two sets of columns display the thresholds for the two written response items – for example, the threshold levels for scores of 2 and 3 on written response item 1 are represented by 1.2 and 1.3, respectively (although each item is scored on a scale of 1 to 5 on this particular examination, only the part of the scale on which a student would be

most likely to get a score of 2 or 3 on either item is shown) – and the probability that a student at 500 on the GSE scale would achieve a score at that particular level on each item. The software also displays, for a student at the selected point on the GSE scale, the expected total score on the multiple-choice section (the figure does not show this part of the display) and the expected score on each of the written response items. Later versions of this software allow the item content to be viewed, as well as, for open-ended items, scoring guides and exemplars as discussed above.

GSE scale	Multiple choice		WR 1		WR 2	
		P		P		P
620						
610						
600						
590						
580						
570	37	.30			2.3	.26
560	15	.34				
550						
540	28 39	.38				
530	27	.41				
520	19 38	.45				
510						
500	34 43 45 48	.50	1.3	.40		
490	17 18 20 40 50	.53				
480	4 31	.56				
470	11 32 33 44 47	.59				
460	5 9 12 46	.61				
450	3 6 7 10 16 29	.64				
440	36	.67				
430	8 14 22 23 26 35	.69				
420	13 24 25	.71				
410	41 42	.73				
400	1 21 30 49	.76				
390						
380					2.2	.56
370	2	.82				
360			1.2	.40		
350						

Figure 6. GSE cut-point setting map.

In order to set the cut points, the committee first acquaints itself with the test materials. The meaning of the various parts of the map is then explained, and the committee members, under the guidance of the trainers, carry out a series of exercises with the software, thus familiarizing themselves with the interpretation of different points on the scale. A detailed explanation of how this process was carried out in an English as a first foreign language testing situation in Germany is given by Draney, Kennedy, Moore and Morrell (2010).

The display of multiple-choice item locations in ascending order of difficulty, next to the written response thresholds, helps to characterize the scale in terms of what increasing proficiency "looks like" in the pool of test-takers. For example, if a

committee was considering 500 as a cut point between performance levels, they could note that at this point, items such as 34, 43, 45 and 48 are expected to be answered correctly about 50% of the time, a harder item like 37 is expected to be answered correctly about 30% of the time and easier items like 2 are expected to be answered correctly 80% of the time. The multiple-choice items near to any chosen point can be seen by the committee so that the members can relate these probabilities to their understanding of the content of the items. The committee could also note that a student at that point (i.e., 500) would be equally likely to score a 2 or a 3 on the first written response item (40% each) and more likely to score a 2 than a 3 on the second (56% vs. 26%). Examples of student work at these levels are also available to the committee for consideration of the interpretation of these scores. Committee members can examine the responses of selected examinees to both the multiple-choice and written response items, note their location on the map and judge the level.

The committee then, through a consensus-building process, sets up cut points on this map, using the item response calibrations to give interpretability in terms of predicted responses to both multiple-choice items and open-ended items. The location of students on the scaled variable are also available for interpretative purposes. This procedure allows criterion-referenced interpretations of cut scores as well as the traditional norm-referenced interpretations.

The use of the maps available from the item response modeling approach not only allows the committees to interpret cut-offs in a criterion-referenced way, it also allows the maintenance of similar standards from year to year through the linking of the item response scales.

DISCUSSION

A central tenet of the assessment discourse over recent years has been the WYTIWYG principle – "what you test is what you get". This principle has paved the way for nationwide assessment reforms at the state or district level. The assumption behind this principle is that assessment reforms will not only affect assessments per se, but that these effects will flow into the curriculum and instruction that students receive in their daily work in classrooms (see Black & Wilson, 2009 for a less rosy view.)

We have demonstrated a way in which large-scale assessments can be more carefully linked to what students are learning. The key here is the use of progress variables to provide a common conceptual framework across curricula. Variables developed and used in the ways in which we have described here can mediate between the level of detail that is present in the content of specific curricula and the necessarily less detailed content of standards documents. This idea of a "cross-walk between standards and assessments" has also been suggested by Eva Baker (Land, 1997, p. 6). These variables create a "conceptual basis" for relating a curriculum to standards documents, to other curricula, and to assessments that are not specifically related to that curriculum.

With the assessments (and instruction) in a specific curriculum structured by progress variables, the problem of item development is reduced – ideas and

contexts for assessment tasks may be adapted from other curricula that share progress variables. The cumulative nature of a curriculum is expressed through: (a) the increasing difficulty of assessments; and (b) the increasing sophistication needed to gain higher scores as expressed in the assessment scoring guides. The same structure clarifies to instructors the ultimate purpose of each instructional activity and each assessment, and also facilitates the diagnostic interpretation of student responses to the assessments.

The idea of a progress variable is not radically new – it has grown out of the traditional approach to test content – and most tests have a "blueprint" or plan that assigns items to particular categories, and hence justifies why certain items are present and others are not. The concept of a progress variable goes further by looking more deeply into why we use certain assessments when we do (i.e., by linking them to growth through the curriculum) and by calibrating the assessments with empirical information.

While the ideas inherent in the BAS are not unique, the combination of these particular ideas and techniques into a usable system does represent a new step in assessment development. The implications of this effort for large-scale tests, curricula and assessment reform on a broader level need to be explored and tested through other related efforts. We hope our efforts and experiences will encourage increased discussion and experimentation with the use of state of the art assessment procedures across a broad range of contexts, from classroom practice to large-scale assessments.

NOTE

[1] Named after Professor Benjamin Wright of the University of Chicago, who has worked so creatively to develop their usage.

REFERENCES

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning*. London: Open University Press.

Black, P., & Wilson, M. (2009). *Learning progressions to guide systems of formative and summative assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, April.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–63.

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping student understanding in chemistry: The perspectives of chemists. *Science Education, 93*(1), 56–85.

Draney, K., & Kennedy, C. (2010). The standard-setting criterion mapping method. In C. Harsch, H. A. Pant & O. Köller (Eds.), *Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany* (Vol. II) pp. 75–88. Münster: Waxmann.

Draney, K., Kennedy, C., Moore, S., & Morrell, L. (2010). Procedural standard-setting issues. In C. Harsch, H. A. Pant & O. Köller (Eds.), *Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany* (Vol. II) pp. 89–121. Münster: Waxman.

Frechtling, J. (1997). *Best practice in action: Final report of the multi-agency study of teacher enhancement programs.* Washington, DC: National Science Foundation.

Glaser, R. (1990). *Testing and assessment: O tempora! O mores!* Pittsburgh: LRDC, University of Pittsburgh.

Kennedy, C. A., Wilson, M., Draney, K., Tutunciyan, S., & Vorp, R. (2008). *ConstructMap Version 4* (computer program). University of, Berkeley, CA: BEAR Center.

Land, R. (1997). Moving up to complex assessment systems. *Evaluation Comment, 7*(1), 1–21.

Linn, R., & Baker, E. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education.* Chicago: University of Chicago Press.

Masters, G., & Forster, M. (1997). *Mapping literacy achievement: Results of the 1996 National School English Literacy Survey*. Hawthorn, Australia: ACER Press.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*(6), 675–698.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 37–76). Boston: Kluwer Academic Publishers.

Roberts, L., & Sipusic, M. (1999). *Moderation in all things: A class act* [Film]. (Available from the Berkeley Evaluation and Assessment Center, Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720-1670).

Ruskus, J., & Luczak, J. (1995). *Best practice in action: A descriptive analysis of exemplary teacher enhancement institutes in science and technology*. Washington, DC: National Science Foundation (Prepared under contract #SED 9255370).

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

Wilson, M., & Adams, R. J. (1996). Evaluating progress with alternative assessments: A model for Chapter 1. In M. B. Kane (Ed.), *Implementing performance assessment: Promises, problems and challenges.* pp. 39–60.. Hillsdale, NJ: Erlbaum.

Wilson, M., & Carstensen, C. (2007). Assessment to improve learning in mathematics: The BEAR Assessment System. In A. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 311–332). London: Cambridge University Press.

Wilson, M., & Draney, K. (2000). *Developmental assessment strategies in a statewide testing program: Scale interpretation, standard setting, and task-scoring for the Golden State Examinations.* Council of Chief State School Officers National Conference on Large Scale Assessment, Snowbird, UT, June.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12–14, 2000, pp 325–332). Tokyo: Springer-Verlag.

Wilson, M., & Scalise, K. (2006). Assessment to improve learning in higher education: The BEAR Assessment System. *Higher Education, 52*, 635–663.

Wilson, M., Scalise, K., Gochyyev, P., Lin, Y.-H., & Torres Irribarra, D. (2010). *Progress monitoring for real classroom contexts: The formative assessment delivery system.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO, May.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181–208.

Wolf, D., & Reardon, S. (1996). Access to excellence through new forms of student assessment. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities. Ninety-fifth yearbook of the National Society for the Study of Education* (Part I, pp. 52–83). Chicago: University of Chicago Press.

Mark Wilson & Karen Draney
Graduate School of Education,
University of California, Berkeley, USA

MICHAELA PFADENHAUER

COMPETENCE – MORE THAN JUST A BUZZWORD AND A PROVOCATIVE TERM?

Toward an Internal Perspective on Situated Problem-Solving Capacity

INTRODUCTION

Following the Progress in International Literacy Study (PIRLS), the Trends in International Mathematics and Science Study (TIMMS), and the Programme for International Student Assessment (PISA), moves are now afoot to extend the measurement of competencies to the higher education sector. This prompted me to take a critical look at "competence". Taking the current "competence boom" and the established concepts of competence as a starting point, the aim of this paper is to argue that competence – as an action-related category – must be conceptualized from the subjective perspective. Precisely because it is such a fragile thing from the individual's point of view, it is not unusual for actors to undertake self-appraisal in order to reassure themselves of their own competence.

THE "COMPETENCE" BOOM

"Competence" is ubiquitous nowadays. The concept enjoys such overwhelming popularity in a wide variety of contexts that hardly any sociological works on the subject in Germany fail to allude to its inflationary use (recent examples: Späte, 2011; Ott, 2010; Kurtz & Pfadenhauer, 2010). The emotional reactions that the term provokes in academic circles outside of the field of empirical educational research are a clear indication that the buzzword "competence" has become a term of provocation. The irritation it causes reflects the resentment felt toward developments in the education sector. In Germany, these developments are associated with keywords such as "G8",[1] "PISA", "the Bologna Process" and "outcome orientation".

In education policy in particular, this "competence-oriented shift" (Arnold, 1997) has been implemented so thoroughly that it is hard to imagine how people managed without the term in the past. Moreover, this "competence boom" has led to the massive displacement of hitherto established terms such as "qualifications", "learning goals" and "education" (in the sense of *Bildung*, i.e., self-formation). From a systems theory perspective, this development must be regarded as semantic displacement, which indicates a systemic change from an education system that emphasizes self-formation (*Bildung*) to one that stresses outcomes. Hence, the

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 81–90.

structural correlates of the transformed system are no longer the educational professions and humanities-oriented education science, but rather the organizations within the education system and empirical educational science whose representatives have joined forces in order to objectify competence:

> The organizations of the education system work on assessments of the competencies that individuals acquire during the periods they spend in the organizations. The empirical educational sciences develop scales that classify competencies and rank them at least ordinally and, ideally, also metrically (Hartig, 2007); [these scales] *measure* on the basis of populations the degree of competence that has actually been realized and that – depending on the interpretation – can be attributed to the individuals as a product of the educational work [of others], as a characteristic of the individuals themselves, or as a residual or confounding variable of [that part of] society that is beyond educational control (the milieu of origin) (Brosziewski, 2010, p. 131; our translation).

According to Richard Münch's institutional economic analysis, this transformation of the education system is due to the fact that two groups are working against each other: on one side, there is an increasingly powerful global elite made up of leading international scientists and economic operators; on the other side, there are increasingly disempowered regional authorities. As the agencies responsible for national educational institutions, the status of which was once unquestioned, these authorities are practically speechless in the face of the dominance of one global culture, the economistic guiding principles of which (for example, education as competence and human capital) are spreading throughout the education system. This confrontation has brought about a "hybrid educational system" (Münch, 2009, p. 31) that – at least in Germany – "is paralysed by massive contradictions" (ibid.). Münch points out that, analogous to the consequences that PISA brought about in schools, "hybrid modernization" with a growing pressure to perform is also to be expected in higher education institutes if elementary competencies are measured that are not, however, being imparted because professors continue to "[plague] students with expectations of academic excellence that, to a large extent, cannot be fulfilled, and thereby render academic studies a myth" (Münch, 2009, p. 52; our translation).

What is striking is that the social dimension – that which Odo Marquard (1981) calls "authority" and which is as much a characteristic of competence as the cognitive dimension (ability) and the non-cognitive dimension (willingness) – has been largely lost sight of in the competence debate.[2] In this paper, in addition to highlighting this social aspect, I argue that competence is by no means a stable – and thus relatively easily measurable – phenomenon, but rather a distinctly fragile thing that inevitably requires self-affirmation on the part of the bearer. Taking the established concepts of competence as my starting point, I identify the gap that exists from a phenomenological-action theoretical perspective and propose a definition that overcomes this shortfall. Against the background of the

accompanying shift in focus, I then explore the assessment of competence from the internal perspective.

THE CONFLICT OF COMPETENCE WITH REGARD TO "COMPETENCE"

The competence discourse is conducted mainly by two disciplines: pedagogy and psychology. In pedagogy, the concept gained relevance in the 1970s because it facilitated the circumvention of the long-standing dispute about the relationship between general and vocational education. The shift in the established semantics toward the concept of competence promised to overcome the narrow focus on the cognitive aspects of (vocational) education and an overly restrictive imparting of directly job-related skills in training and continued education, and to move in the direction of a more holistic form of competence development that takes into account an individual's whole personality. In the field of research into pedagogy and vocational education, a broad concept of competence prevails. Focusing on comprehensive ability and maturity, it comprises not only cognitive but also affective and motivational components (cf. Baethge et al., 2006; Fischer, 2010; Straka & Macke, 2010a).

In psychology (and in psychology-oriented educational science), on the other hand, one finds a narrower understanding of competence as "the ability (disposition) to master different demanding situations" (Jude & Klieme, 2008, p. 11; our translation). In contrast to the decontextualization which is symptomatic of intelligence testing, competencies are defined as "learnable, context-specific performance dispositions that relate functionally to situations and demands in certain domains" (Klieme & Hartig, 2007, p. 14; our translation). Hence, competence is related to concrete tasks, whereby the cognitive ability to master these tasks, which is acquired through the acquisition of specialized knowledge, is psychometrically modeled and measured.

In the field of vocational education research in particular, the problems ensuing from this reduction of the concept of competence to the specificities of the context, specialized knowledge and the cognitive dimension have been highlighted. These problems are also acknowledged by empirical education researchers. As "competence" in a broad sense resists measurement, the object of research is adapted to the logic of measurement – "operationalized", say the protagonists; "missed", say the critics.

A CONCEPT OF COMPETENCE BASED ON PHENOMENOLOGICAL ACTION THEORY

While highlighting the problem-solving aspect of competence, the following proposed definition does not confine competence to the cognitive dimension of being (mentally) capable of something because, at least as regards problem-solving action, understanding competence merely as the knowledge required to solve a particular problem is too restrictive. Ability based on actively acquired and socially imparted sedimented experiences must be seen in dialectic relationship to action

(cf. Fischer, 2010, S. 143). In other words, it is a question of "practical knowledge" (Knoblauch, 2010, S. 249), i.e., "know-how". As regards action, this ability must be accompanied by a willingness (for whatever reason) to master the problem at hand. Competent action also calls for motivation that stems from relevancies and interests. This motivation goes beyond the usual motivation to put a plan into action insofar as it requires an attitude toward penetrating a problem. This attitude is not "automatic", but must be assumed consciously.

However, the proposal advanced by vocational education researchers that competence should be regarded as "an entity [comprising] motive and the ability to act" (Straka & Macke, 2010b, p. 226; our translation) is also too restrictive. In the words of Christiane Hof (2002, p. 158; our translation), this concept also points to "the question of authority[3] and the assumption that competence manifests itself in the execution of a job in accordance with the expected standards." According to this view, competence should be perceived as an ascription in the sense that observable behavior is deemed to be in accordance with standardized expectations. Straka and Macke (2009, p. 16) also refer to the semantic content of competence as "socially ascribed authority" (which finds expression specifically in occupational profiles). However, in both cases, "authority" is formulated from the observer's perspective.

It was with reference to the law that Max Weber established the sociological connotations of competence in the sense of Zuständigkeit "as a basic category of rational authority" (Kurtz, 2010, p. 9). The legal distinction between formal and material competence indicates that authority can be formally ascribed (i.e., externally attributed) on the basis of a social position or an organizational function (competence by virtue of an office or position), or that it is the material outcome, as it were, of a subjective store of knowledge (competence by virtue of knowledge) (cf. Kühl, 2010 following Luhmann, 1964). In the latter sense at least, "authority" must be formulated from the subjective perspective.

According to Schulz-Schaeffer (2007, p. 14), the constitution of action through ascription can be viewed as a second way in which events are constituted as action. It can be seen as an "independent act of interpretation" which can "either supplement the constitution of action by the actor, compete with it, or be the only form of constitution of the event in question as action" (our translation). On the one hand, authority can be the result of external ascription ("being considered responsible for something"). On the other hand, however, it can also manifest itself as a subjective claim ("considering oneself to be responsible for something"). From an action theory point of view, this subjective claim develops in a complex manner and is related to the perception of a situation as one that concerns me. However, it does not concern me solely because of my individual motivational situation and my ability, but because of a prevailing interaction order. In both cases, the concept of competence connotes "responsibility", as defined by Alfred Schütz (1972, p. 256), or the established evolution and socialization theory considerations of Thomas Luckmann (2007). In the case of external ascription, I am responsible to someone, namely the person who made me responsible for something. In this case, my competence is a fragile thing, in the sense that it is not

I, but someone else, who decides whether or not I am competent. When competence is subjectively ascribed, I consider myself responsible for what I do or have done. In this case, competence is a fragile thing insofar as I need a frame of reference in order to decide whether or not I am competent, and this frame of reference must not itself be fragile.

The incorporation of authority into the definition of competence once again places greater emphasis on the social dimension of competence. The social aspect was already present in the linguistic concept of competence in the form of the normatively employed concept of acceptability. It is highlighted in all attempts to define communicative competence that emphasize the situational appropriateness of verbal and nonverbal utterances, whereby a real-time reference to performance, i.e., a reference to the situation and the prevailing interaction order, is implied.

A concept of competence that avoids addressing the action problem in a one-sided way includes three components: ability, willingness and authority. From this pespective, competent action is constituted through a capacity for iterative problem-solving that is characterized by "being able to", "wanting to", "being allowed to" and "being obliged to" do something, as perceived by the actor himself or herself.[4] Actors do not simply "have" this capacity habitually. Rather, they must bring it into the situation by applying an "action template" to an action goal. This capacity, which despite incorporation cannot simply be accessed like a construction kit, is the prerequisite for multifaceted and always domain-specific problem-solving action. It enables the actor to master problems in an intentional rather than a random way; in a systematic rather than "any old" way; and repeatedly rather than on a once-off basis. Moreover, this capacity is not visible from the outside. Indeed, even in the case of actors who are confident in their own competence, it manifests itself only when the action is being executed.

COMPETENCE FROM THE INTERNAL PERSPECTIVE

Composite terms such as social competence, media competence and information competence, to name but a few, indicate that competence is a multi-layered phenomenon. Efforts have been made to curb this inflation of competence types with the help of competence models. In his frequently cited classification, Heinrich Roth (1971) deconstructs the concept of competence into its experiential components: things, other people and the self. What is most striking about this classification is the fact that the aspect of language or speech, which is central to competence concepts that have recourse to Noam Chomsky, is not assigned particular importance. By contrast, Jürgen Habermas' (1984) distinction between cognitive, linguistic and interactive competence, which is based on the differentiation of the human environment into the regions of "external nature", "language" and "society", not only incorporates Piaget's developmental psychology-based concept of competence, but also Chomsky's concept of linguistic competence. Habermas expands Chomsky's concept from an action theory perspective and relates it to his understanding-oriented theory of communicative action. In Germany, the differentiation of action competence into subject-, method-,

social- and reflexive/personal/human competence (cf. Erpenbeck & Heyse, 1999) has risen to particular prominence. However, these stereotypical categorizations are not really reflected in the meaningful stratification of actors' experience, as I have demonstrated using organizational competence as an example (cf. Pfadenhauer, 2008b).

The latter study revealed that, from an internal perspective, a competent organizer is someone who divides organizational processes into various sub-projects, then breaks these sub-projects into "manageable" tasks, then divides these tasks into action steps which are as distinct as possible, then lays down the spatial and temporal order in which these steps are to be performed, and finally assigns the task of implementing the individual steps to the actor best suited to the task in question. A competent organizer of projects based on the social division of labor is someone who lays down in the most distinct and precise way possible what is to be done, by whom, when, where and in what way. A competent organizer is someone who, with these "rules of procedure", provides a binding, reliable basis for actions to be carried out by others that proves flexible even when unintended side effects occur. A competent organizer of this social division of labor is someone who is capable of ensuring that every individual involved in the realization of the project does what he or she is supposed to do and abides by the prescribed targets, forms of action and time limits. Finally, a socially competent organizer (i.e., an organizer who works on the basis of the given demands, or the demands that are accepted as given) is someone who also reflects on and evaluates the actions performed by the individual actors in terms of the adequacy of their contribution to the achievement of the target values.

Whereas, in the standard model, organizational competence is subsumed under "method competence" (cf. Schaeper & Briedis, 2004, p. 5), organizational competence encompasses all of the facets of action competence, which precludes the artificial division into categories. Competence – in this case, organizational competence – is linked to the inter-connected components of the process of organizing action, which entails providing the prerequisites for the actions of others, influencing their actions in a certain direction and evaluating these actions in terms of the target values (cf. Pfadenhauer, 2008a). In contrast to our everyday understanding of organizing, in which preparation and implementation activities are "mixed up", organizing is perceived scientifically as "higher order action" (Spann, 1969, p. 315); in other words, as meta-action that gives rise to other actions.

MEASURING COMPETENCE

Against this background, the definition of competence as the capacity to solve problems iteratively aims to take into serious consideration the action aspect, insofar as one must always clarify what characterizes the type of action to which the competence in question refers. This problem-solving capacity is classified as "situative" because competence refers in principle to a situation – a situation that is not simply "given", i.e., objective. In view of the fact that situations are generated

when certain parts of the *Lebenswelt* (lifeworld) acquire relevance, and by virtue of being situations, acquire distinct contours, Vonken (2005) defines competence in a fundamental way, namely as the capability to bring forth situations. According to Vonken, competence is "that which causes one to perceive and address – i.e., to generate – a situation" (Vonken, 2005, p. 186; our translation). As a rule, the usual understanding of competence as the ability to master situations neglects the fact that, to the actor, the situation appears to be both "given" and definable (cf. Hitzler, 1999). The actor usually experiences situations as a manifestation of social order structured by institutions, for example norms, or constructed in the course of interaction. In other words, situations are perceived as being characterized by a complex web of behavioral patterns, i.e., as being predefined with a claim to bindingness. In addition to these "given" conditions, the individual's subjective experiences and interests also enter into his or her definition of the situation, thereby giving the socially objectified definitions their specific importance for him or her as an actor.

Every situation in which action takes place has many aspects: (frequently, but not necessarily) other actors; "things" (in the broadest sense, i.e., techniques, language and knowledge); the self in his or her concrete mental (intentional, contra- or peri-intentional) orientation and physical condition (healthy/sick, sober/inebriated, etc.); surroundings (temperature, air, atmosphere, weather); spatiality (narrow/wide, good visibility/fog); sounds (noisy/quiet); smells, etc. In other words, every situation is equipped with correlates of sensory perceptions to which experience (perception and imagination) can – and to a certain extent must – be directed if the situation is to be mastered in accordance with one's own goals and responsibilities.

Moreover, as in the case of competent organizing (cf. once again Pfadenhauer, 2008b), every solution to a problem which is not exclusively cognitive is based on a conglomeration of elements of knowledge, techniques, strategies and reflections that can be broken down into various individual aspects. A considerable number of these elements are accessible to the conscious mind when one considers: (a) what one usually does (and has to do); and (b) how, using what practical knowledge, techniques (including physical techniques), social strategies, cognitive procedures, etc. does one manage to do what one does in a manner which is adequate in the situation at hand.

This does not dispute the fact that, during such an analysis of one's own problem-solving action, aspects which are relevant to the resolution of the problem, such as one's own impact, implicit knowledge, unintended side effects, etc., may be neglected. Although competence encompasses an individual's entire problem-solving ability, it goes without saying that someone who provides information about his or her competence may give information only about the components that he or she considers to be pragmatic or necessary for the situational mastering of problems. Occasions for such a disclosure of (personal) information are especially likely to arise when there are grounds for doubt – be it doubt on the part of others in the light of prior problem-solving actions, or self-doubt. One may doubt that one is actually able to master something that (for

whatever reason) one considers oneself to be "actually" prepared to do, "actually" capable of and, having generated the situation, for which one is "definitely" responsible.

CONCLUSION

Assessing one's own competence serves to reassure oneself. It allows one to objectify the "subjective and social capability for appropriate action" (Knoblauch, 2010, p. 248; our translation), which one does not simply have (like money in a bank account), but which (like stocks and shares) become manifest only on the point of transfer. The fact that one can access it and hold on to it for only as long as one uses it renders competence a fragile thing. Only the recollection of a prior problem-solving action, which is sedimented as experience, can provide evidence of an iterative problem-solving capacity. Hence, such considerations are not unusual, but are rather an everyday (albeit rarely explicit) process of assessing one's own competence.

NOTES

[1] "G8" refers to the change from the nine-year to the eight-year Gymnasium (secondary school leading to higher education entrance qualification).

[2] In addition, as will be shown below, in the rare cases in which this dimension is addressed, it is introduced as an observer category. In contrast, I argue that, like the other two dimensions, the social dimension must also be defined from the internal perspective.

[3] Henceforth, I use the word "authority" as hopefully the best translation of the German term "Zuständigkeit" (cf. Mulder, 2007).

[4] Due to its consistently internal perspective, it is only at first glance that this concept of competence fits in with the definition proposed by Straka and Macke (2009, p. 16; our translation), who argue that competence should be interpreted "as the product of an interaction between 'being allowed to act' (having been assigned competence) and 'being able and willing to act' (being able and willing to comply with the assigned competence)".

REFERENCES

Arnold, R. (1997). Von der Weiterbildung zur Kompetenzentwicklung. In Arbeitsgemeinschaft QUEM (Ed.), *Kompetenzentwicklung '97. Berufliche Weiterbildung in der Transformation – Fakten und Visionen* (pp. 253–299). Münster: Waxmann.

Baethge, M., Achtenhagen, F., Arends, L., Babic, E., Baethge-Kinsky, V., & Weber, S. (2006). *Berufsbildungs-PISA – Machbarkeitsstudie*. Stuttgart: Steiner.

Brosziewski, A. (2010). Von Bildung zu Kompetenz. Semantische Verschiebungen in den Selbstbeschreibungen des Erziehungssystems. In T. Kurtz & M. Pfadenhauer (Eds.), *Soziologie der Kompetenz* (pp. 119–134). Wiesbaden: VS Verlag für Sozialwissenschaften.

Erpenbeck, J., & Heyse, V. (1999). *Die Kompetenzbiographie. Strategien der Kompetenzentwicklung durch selbstorganisiertes Lernen und multimediale Kommunikation*. Münster: Waxmann.

Fischer, M. (2010). Kompetenzmodellierung und Kompetenzmessung in der beruflichen Bildung – Probleme und Perspektiven. In M. Becker, M. Fischer & G. Spöttl (Eds.), *Von der Arbeitsanalyse zur Diagnose beruflicher Kompetenzen* (pp. 141–158). Frankfurt amMain.: Peter Lang.

Habermas, J. (1984). Notizen zur Entwicklung der Interaktionskompetenz. In J. Habermas (Ed.), *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns* (pp. 187–225). Frankfurt a.M.: Suhrkamp.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 83–99). Weinheim: Beltz.

Hitzler, R. (1999). Konsequenzen der Situationsdefinition. Auf dem Weg zu einer selbstreflexiven Wissenssoziologie. In R. Hitzler, J. Reichertz & N. Schröer (Eds.), *Hermeneutische Wissenssoziologie*. Konstanz: UVK.

Hof, C. (2002). (Wie) lassen sich soziale Kompetenzen bewerten? In U. Clement & R. Arnold (Eds.), *Kompetenzentwicklung in der beruflichen Bildung* (pp. 289–308). Opladen: Leske + Budrich.

Jude, N., & Klieme, E. (2008). Einleitung. In N. Jude, J. Hartig & E. Klieme (Eds.), *Kompetenzerfassung in pädagogischen Handlungsfeldern. Bildungsforschung Band 26* (pp. 11–15). Bonn: BMBF.

Klieme, E., & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Eds.), *Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaften. Sonderheft 8* (pp. 11–31). Wiesbaden: VS Verlag für Sozialwissenschaften.

Knoblauch, H. (2010). Von der Kompetenz zur Performanz. Wissenssoziologische Aspekte von Kompetenz. In T. Kurtz & M. Pfadenhauer (Eds.), *Soziologie der Kompetenz* (pp. 237–255). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kühl, S. (2010). Ächtung des Selbstlobs und Probleme der Kompetenzdarstellung. In T. Kurtz & M. Pfadenhauer (Eds.), *Soziologie der Kompetenz* (pp. 275–291). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kurtz, T. (2010). Der Kompetenzbegriff in der Soziologie. In T. Kurtz & M. Pfadenhauer (Eds.), *Soziologie der Kompetenz* (pp. 7–25). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kurtz, T., & Pfadenhauer, M.. *Soziologie der Kompetenz*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Luckmann, T. (2007). Zur Entwicklung und geschichtlichen Konstruktion persönlicher Identität. In T. Luckmann (Ed.), *Lebenswelt, Identität und Gesellschaft* (pp. 231–253). Konstanz: UVK.

Luhmann, N. (1964). *Funktionen und Folgen formaler Organisation*. Berlin: Duncker & Humblot.

Marquard, O. (1981). Inkompetenzkompensationskompetenz. In O. Marquard (Ed.), *Abschied vom Prinzipiellen* (pp. 23–38). Stuttgart: Reclam.

Münch, R. (2009). *Globale Eliten, lokale Autoritäten. Bildung und Wissenschaft unter dem Regime von PISA, McKinsey & Co*. Frankfurt am Main: Suhrkamp.

Mulder, M. (2007). Competence – the essence and the use of the concept in ICVT. *European Journal of Vocational Training, 40*(1), 5–22.

Ott, M. (2010). *Aktivierung von (In-)Kompetenz. Praktiken im Profiling – eine machtanalytische Ethnographie.* Konstanz: UVK.

Pfadenhauer, M. (2008a). *Organisieren. Eine Fallstudie zum Erhandeln von Events*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Pfadenhauer, M. (2008b). Doing phenomenology: Aufgrund welcher Merkmale bezeichnen wir ein Handeln als "kompetentes Organisieren"? In J. Raab, M. Pfadenhauer, P. Stegmaier, J. Dreher & B. Schnettler (Eds.), *Phänomenologie und Soziologie. Positionen, Problemfelder, Analysen* (pp. 339–348). Wiesbaden: VS Verlag für Sozialwissenschaften.

Pfadenhauer, M. (2010). Kompetenz als Qualität sozialen Handelns. In T. Kurtz & M. Pfadenhauer (Eds.), *Soziologie der Kompetenz* (pp. 149–172). Wiesbaden: VS Verlag für Sozialwissenschaften.

Roth, H. (1971). *Pädagogische Anthropologie. Volume 2: Entwicklung und Erziehung*. Hannover: Schroedel.

Schaeper, H., & Briedis, K. (2004). *Kompetenzen von Hochschulabsolventinnen und Hochschulabsolventen, berufliche Anforderungen und Folgerungen für die Hochschulreform*. Retrieved from http://www.forschung.bmbf.de/pub/his_projektbericht_08_04.pdf

Schütz, A. (1972). Einige Äquivokationen im Begriff der Verantwortlichkeit. In A. Schütz (Ed.), *Gesammelte Aufsätze* (Vol. 2, pp. 256–258). Den Haag: Nijhoff.

Schulz-Schaeffer, I. (2007). *Zugeschriebene Handlungen. Ein Beitrag zur Theorie sozialen Handelns*. Weilerswist: Velbrück.

Späte, K. (2011). *Kompetenzorientiert Soziologie lehren.* Leverkusen: Barbara Budrich.

Spann, O. (1969). Gesamtausgabe. Band 4: Allgemeine Gesellschaftslehre [zuerst 1914]. In W. Heinrich, H. Riehl, R. Spann & F. A. Westphalen (Eds.), *Gesamtausgabe* (pp. 503–523). Graz: Akademische Druck- u. Verl.-Anst.

Straka, G. A., & Macke, G. (2009). Berufliche Kompetenz: Handeln können, wollen und dürfen. Zur Klärung eines diffusen Begriffs. *Zeitschrift in Wissenschaft und Praxis BWP, 3*, 14–17.

Straka, G. A., & Macke, G. (2010a). Kompetenz – nur eine „kontextspezifische kognitive Leistungsposition“? *Zeitschritt für Berufs- und Wirtschaftspädagogik, 106*(3), 444–451.

Straka, G. A., & Macke, G. (2010b). Sind das „Dogma vollständige Handlung“ und der „Pleonasmus Handlungskompetenz“ Sackgassen der bundesdeutschen Berufsbildungsforschung? Ein kritisch-polemischer Essay. In M. Becker, M. Fischer & G. Spöttl (Eds.), *Von der Arbeitsanalyse zur Diagnose beruflicher Kompetenzen* (pp. 215–229). Frankfurt am Main: Peter Lang.

Vonken, M. (2005). *Handlung und Kompetenz. Theoretische Perspektiven für die Erwachsenen- und Berufspädagogik.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Michaela Pfadenhauer
Institute of Sociology
Karlsruhe Institute of Technology (KIT), Germany

PART 2

INSTRUMENTS AND STUDIES

SIGRID BLÖMEKE

THE CHALLENGES OF MEASUREMENT IN HIGHER EDUCATION

IEA's Teacher Education and Development Study in Mathematics (TEDS-M)

In 2008, a comparative study was carried out that focused on the outcomes of teacher education with standardized testing. Representative samples of primary and secondary mathematics teachers in their final year of teacher training from 17 countries in Africa, the Americas, Asia and Europe were examined, as well as representative samples of teacher educators and training institutions. The "Teacher Education and Development Study: Learning to Teach Mathematics" (TEDS-M) was carried out under the supervision of the International Association for the Evaluation of Educational Achievement (IEA).[1] TEDS-M looked at how teachers of mathematics were trained and what kinds of competences they had acquired at the end of their training. More than 24,000 prospective teachers were assessed on their mathematics content knowledge (MCK), mathematics pedagogical content knowledge (MPCK) and general pedagogical knowledge (GPK). In addition, they were surveyed regarding their beliefs about mathematics as well as about the teaching and learning of mathematics.

TEDS-M was a challenging enterprise. The variability of higher education institutions, programs and processes led to heated discussions about the comparability of the results (see section 3.1 for more details). It also led to a new approach to the definition of teacher competences. The traditional IEA approach to developing assessments as used in the "Third International Mathematics and Science Survey (TIMSS)" based on students' opportunities to learn (OTL) did not seem feasible (see section 1.1 for more details). The limitations of paper-and-pencil approaches and multiple-choice items increased the challenge of providing accurate indicators of the teachers' competences (see section 1.2).

The sampling process was also difficult, because several levels of aggregation had to be taken into account as well as the problem of unstable group membership (see section 2.1 for details). In TEDS-M, adults were surveyed who were only loosely associated with a classroom structure. As a result, countries struggled with their response rates (see section 2.1). Finally, the scaling was less straightforward than in other studies because it was necessary to deal with broad constructs and inherent multidimensionality (see section 2.2 for more details).

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 93–112.

This paper describes the theoretical framework of TEDS-M and the study design in order to illustrate how the study dealt with these challenges. Then, the core results are documented in order to illustrate the potential of such studies in higher education. Finally, conclusions are drawn with regard to further research on the outcomes of higher education.

THEORETICAL FRAMEWORK

Teachers' Professional Competence

TEDS-M was conceptualized closely to the notion of professional competence in general as it was defined by Weinert (2001), and specifically with regard to teaching, as outlined by Bromme (1997). Competence in this tradition means having the knowledge and skills to successfully solve core job-related problems. Weinert (2001) divides competence into the cognitive ability to solve such problems and the motivational, volitional and social willingness to successfully and responsibly apply these solutions in various situations. In TEDS-M, cognitive ability was categorized as teachers' professional knowledge. Motivational, volitional and social willingness were categorized as teachers' professional beliefs (see Figure 1).

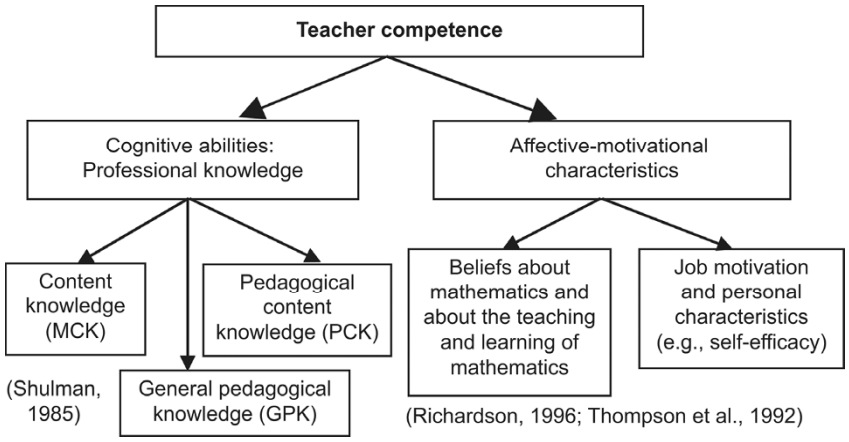


Figure 1. Theoretical framework of teacher competence in TEDS-M.

As frequently discussed in the literature, professional knowledge can be divided into three facets: content knowledge; pedagogical content knowledge; and generic pedagogical knowledge (Shulman, 1985; Blömeke, 2002). In the context of TEDS-M, content knowledge is the knowledge of mathematics. Pedagogical content knowledge refers to knowledge about teaching and learning mathematics. Pedagogical knowledge, finally, is the knowledge typically acquired in a teacher training program that is not subject-matter related (Blömeke & Paine, 2008).

An important implication of this definition of teachers' professional competence is its situated nature and applicability (Blumer, 1969). Teachers' knowledge is therefore only fully evaluated if it is applied to different classroom situations. The problems and situations in the TEDS-M test were set by constitutive features of the profession. In order to determine which features are, in fact, constitutive, TEDS-M referred to existing standards in various national teacher training programs (KMK, 2004; NCTM, 2000). Table 1 exemplifies the problems which mathematics teachers are expected to solve with their pedagogical content knowledge in the TEDS-M test, based on these standards.

Table 1. Core situations which mathematics teachers are expected to manage (Tatto, Schwille, Senk, Ingvarson, Peck, & Rowley, 2008)

Mathematical curricular knowledge	Establishing appropriate learning goals Knowing about different assessment formats Selecting possible pathways and seeing connections within the curriculum Identifying the key ideas in learning programs Knowledge of the mathematics curriculum
Knowledge of planning for mathematics teaching and learning [pre-active]	Planning or selecting appropriate activities Choosing assessment formats Predicting typical student responses, including misconceptions Planning appropriate methods for representing mathematical ideas Linking didactical methods and instructional designs Identifying different approaches for solving mathematical problems Planning mathematics lessons
Enacting mathematics for teaching and learning [interactive]	Analyzing or evaluating students' mathematical solutions or arguments Analyzing the content of students' questions Diagnosing typical student responses, including misconceptions Explaining or representing mathematical concepts or procedures Generating fruitful questions Responding to unexpected mathematical issues Providing appropriate feedback

The understanding of "competence" used in TEDS-M includes the instructional, professional and personal beliefs held by future teachers. Beliefs were defined by Richardson (1996, 103) as "psychologically held understandings, premises, or propositions about the world that are felt to be true". Teachers' beliefs are crucial to their perception of classroom situations and to their decisions on how to act (Leder, Pekhonen, & Törner, 2002; Leinhardt & Greeno, 1986). If beliefs are operationalized specifically to both the content being taught and the challenges a specific classroom situation presents, empirical evidence exists for a link between teacher beliefs and student achievement (Bromme, 1994).

Beliefs have a vital function with regard to orientation as well as action (Grigutsch, Raatz, & Törner, 1998). Therefore, they connect knowledge and action. In this sense, they are also an indicator of the type of instruction that teachers will use in their future teaching (Brown & Rose, 1995). Several types of teacher beliefs are distinguished in TEDS-M (Calderhead, 1996), which are mainly epistemological beliefs about the nature of mathematics and beliefs about the teaching and learning of mathematics (Thompson, 1992).

Predictors

Teacher competence is the core criterion for effective teacher training in TEDS-M. In order to explain which factors may influence the development of teacher competence, potentially influential factors were divided into three categories: the individual characteristics of future teachers; the institutional characteristics of teacher training; and the systemic characteristics of the country in question (see Figure 2).

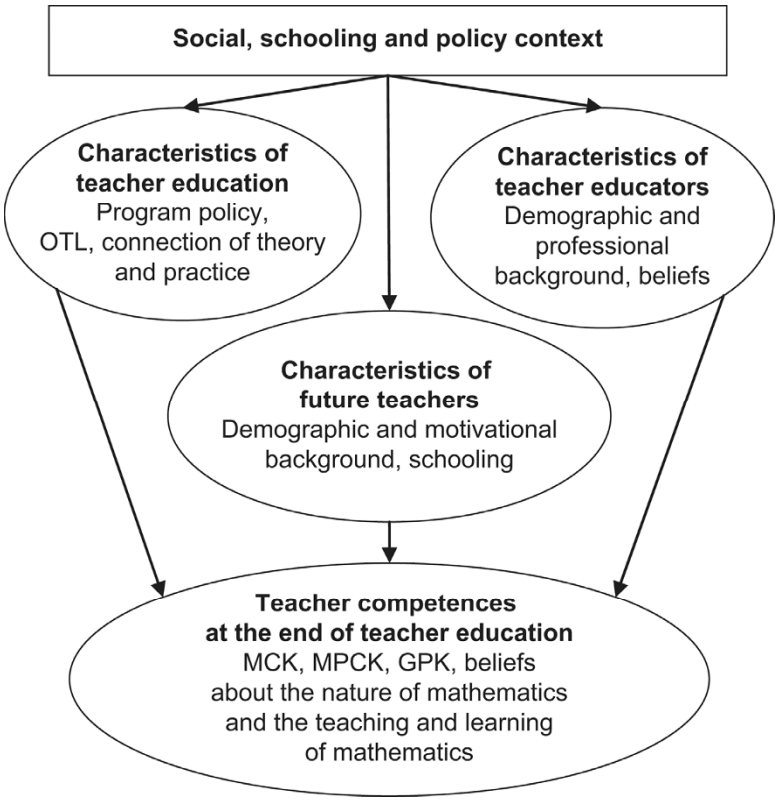


Figure 2. Theoretical framework of features influencing the acquisition of competence (see also Tatto et al., 2008).

It is known from instructional research that school achievement is influenced by the students' prior knowledge and their motivation (Helmke, 2004). Such *individual* prerequisites and characteristics may also play a role in teacher training. The present state of research regarding *institutional* characteristics suffers from the fact that only crude data exist about opportunities to learn mathematics, mathematics pedagogy and general pedagogy. Therefore, unsurprisingly, findings about the effects of institutional features in teacher training on outcomes are inconsistent (Blömeke, 2004; Cochran-Smith & Zeichner, 2005; Wilson, Floden, & Ferrini-Mundy, 2002).

One can imagine a wide array of institutional features which may influence the acquisition of competence: selectivity; program content; teaching methods; the characteristics of teacher educators; the accountability of institutions and educators; and the location and organization of teacher training. TEDS-M additionally distinguishes between the intended and the implemented features of teacher training and adopts a multifaceted approach by analyzing a representative sample of course offerings as well as asking future teachers about their actual opportunities to learn in order to overcome the methodological problems of prior studies.

STUDY DESIGN

Sampling

The target population of TEDS-M consisted of students in their final year of teacher training who were on track to receive a license to teach mathematics either in primary or secondary schools (Tatto et al., 2008). A teacher training program was included if it prepared primary teachers for one of grades 1 through 4 as the common denominator of level 1 education in the "International Standard Classification of Education" (primary or basic education, cycle 1; UNESCO, 1997) or if it prepared lower secondary teachers for grade 8, as the common denominator of level 2 education (lower secondary or basic education, cycle 2).

The TEDS-M International Study Center at Michigan State University established sampling guidelines and procedures for the National Research Centers (NRCs) of the participating countries. The NRCs provided the sampling frames, while the selection of sample elements, weighting, the determination of the participation rate and adjudication were conducted by the IEA Data Processing and Research Center (DPC). The NRCs were responsible for data collection and the DPC was responsible for assembling and processing the international dataset released to the NRCs in December 2009 (for more details, see Tatto et al., 2008).

In a two-stage process, random samples were drawn from the target population in each participating country. The samples were stratified according to important teacher training features like "route" (consecutive vs. concurrent programs), "type" of program (grade span the license includes, e.g., grades 5 through 9 vs. 5 through 12), "focus" of opportunities to learn (with or without extensive mathematics) and "region" (e.g., federal state) in order to reflect accurately the distribution of future primary and lower secondary teachers' characteristics at the end of training. The goal

for each country's sample size was a simple random sample of 400 future primary and 400 future lower secondary teachers in their final year of teacher training on the national level. However, in most countries, the sample size was lower, so that a census was taken because the target population was relatively small.

In 2008, approximately 15,000 future primary and 9,000 future lower secondary teachers from 17 countries (see Table 2) were tested on their knowledge of mathematics and mathematics pedagogy by a standardized paper-and-pencil assessment. All countries had to meet the IEA's quality requirements, as set out in TIMSS or PIRLS. This included controlling the translation processes, monitoring test situations and meeting the required participation rates. The TEDS-M quality standards require minimum participation rates for all target populations of the survey. The aim of these standards is to ensure that bias resulting from non-response is kept within acceptable limits.

The data collection procedures and response rates were evaluated by the IEA Sampling Team. The evaluation resulted in adjudication comments and recommendations with regard to how the data were to be reported (see Dumais & Meinck, in press a): "reporting without any annotation" was used if all participation rate requirements were met, the exclusion rate was below 5% and full coverage of the target population was observed. Besides other countries, the German results were reported in this way.

"Annotation because of low participation rates" was used if the participation rate did not meet the required minimum. This applied to four countries on the primary level (Chile, Norway, Poland and the USA) and five countries on the lower secondary level (Chile, Georgia, Norway, Poland and the USA). Annotation was also advised if reduced coverage of the target population was observed (this applied to Poland, Switzerland and the USA), the sample composition did not fully meet the TEDS-M definition of the target population (this applied to Norway) or a substantial proportion of missing values was observed (this applied to the USA). "Unacceptable" was the verdict if the combined participation rate dropped below 30%. Canada had to be excluded from the study for this reason.

Table 2. Countries participating in TEDS-M

Botswana	Chile	Germany	Georgia
Canada	Malaysia	Norway	Oman
Philippines	Poland	Russia	Switzerland
Singapore	Spain	Taiwan	Thailand
U.S.A.			

Instruments

TEDS-M sought to measure future teachers' MCK and MPCK at the end of teacher training. For this purpose, a 60-minute paper-and-pencil assessment was completed

during a standardized and monitored test session. The items were designed to depict classroom performance as accurately as possible.

In order to capture the desired breadth and depth of teacher knowledge, a matrix design was applied. Five primary and three lower secondary test booklets were developed with rotating blocks of items (using a balanced incomplete block design). Scaled scores were calculated using item response theory. The achievement scores were transformed onto a scale with an international mean of 500 and a standard deviation of 100 test points. The 76 items of the MCK test covered numbers (e.g., whole numbers, fractions and decimals), algebra (e.g., equations/formulae and functions) and geometry (e.g., geometric shapes, location and movement/rotation) with approximately equal weight and (to a lesser extent) data. In addition, three cognitive dimensions were covered: knowing (e.g., recalling and computing); applying (e.g., representing and implementing) and reasoning (e.g., analyzing and justifying). A third heuristic emphasized different levels of expected difficulty (novice, intermediate and expert). A sample item is given in Figure 3 (the full set of released items is available from tedsm@msu.edu).

The 32 items of the MPCK test covered two subdimensions:

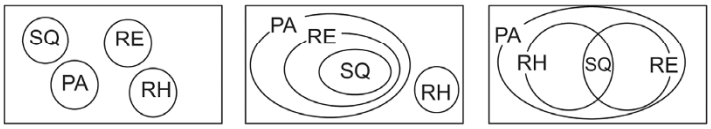
- Knowledge of curricula and planning, which is necessary before a teacher enters the classroom (e.g., establishing appropriate learning goals, knowing about different assessment formats or linking didactic methods);
- Instructional designs (identifying different approaches to solving mathematical problems) as well as interactive knowledge about how to enact mathematics for teaching and learning (e.g., diagnosing typical student responses, including misconceptions, explaining or representing mathematical concepts or procedures, providing appropriate feedback).

In line with the MCK test, three levels of expected difficulty and four content areas were distinguished. An example is given in Figure 4.

The item development was based mainly on the MT21 study (Schmidt, Blömeke, & Tatto, 2011), as well as on the two Michigan studies entitled "Knowing mathematics for teaching algebra" (KAT; Ferrini-Mundy, Floden, McCrory, Burrill, & Sandow, 2005) and "Learning mathematics for teaching" (LMT; Hill, Loewenberg Ball, & Schilling, 2008). Three item formats were used: multiple choice, complex multiple choice and open constructed response.

Three countries added a national option in which they measured general pedagocial knowledge (GPK) with a 30-minute paper-and-pencile assessment. The countries were Germany, Taiwan and the US. See Figure 6 for an example.

Three students have drawn the following Venn diagrams showing the relationships between four quadrilaterals: rectangles (RE), parallelograms (PA), rhombuses (RH) and squares (SQ).



[Tian] [Rini] [Mia]

Which student's diagram is correct? *Check <u>one</u> box.*

A. [Tian] o_1

B. [Rini] o_2

C. [Mia] o_3

Figure 3. Sample item from the TEDS-M primary test of MCK.

Prove the following statement:
If the graphs of linear functions

$f(x) = ax + b$

and

$g(x) = cx + d$

intersect at a point P on the x-axis, the graph of their sum function $(f + g)(x)$ must also go through P.

Figure 4. Sample item from the TEDS-M lower secondary test of MCK

When teaching children about length measurement for the first time, Mrs. [Ho] prefers to begin by having the children measure the width of their book using paper clips, and then again using pencils.

Give **TWO** reasons she could have for preferring to do this rather than simply teaching the children how to use a ruler?

Figure 5. Sample item from the TEDS-M primary test of MPCK.

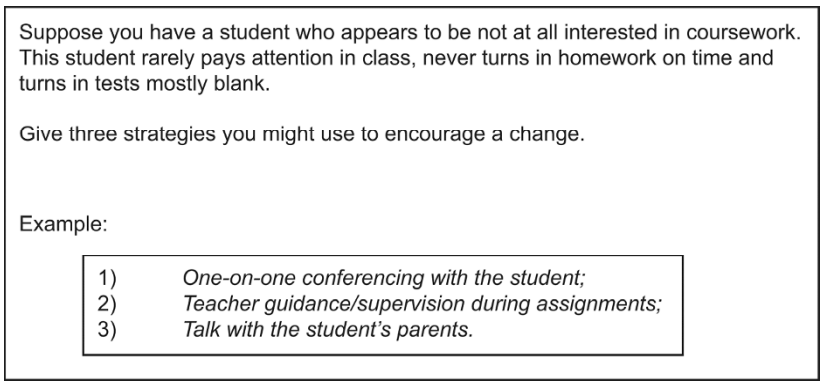


Figure 6. Sample item from the TEDS-M test of GPK with sample answer.

Data Analysis

Due to the complex sampling design, standard errors were estimated using balanced repeated replication (BRR) (Dumais & Meinck, in press b). Weights were determined by Statistics Canada according to the sampling design and adjusted for non-participation and non-response. Parameter estimations were determined using the International Database Analyzer provided by the IEA. Therefore, the results give a sound picture of the professional competences of future mathematics teachers who in 2008 were in their final year of teacher training.

Regarding the report on teacher training outcomes in TEDS-M, it is necessary to distinguish between an evaluation of the national teacher training system and an evaluation of specific teacher training programs within countries. Both approaches have their benefits and their limitations. Due to the traditional policy orientation of the large-scale assessments of the IEA and the OECD, studies such as TIMSS, PISA or PIRLS focus on the national level. This is a valuable approach, because it stresses the overall educational effectiveness of a nation, regardless of the structure of its education system. Therefore, TEDS-M followed this approach on the one hand. From this perspective, with regard to international competitiveness, it is important to consider what a nation accomplishes as a whole.

However, it is obvious that additional information is to be gained by looking into program types. Only then is it possible to learn about pathways to success without confounding variables like cultural or societal features or the economic status of a country. Therefore, TEDS-M followed an alternative approach on the other hand by reporting outcomes on the level of program types within countries. However, this approach must be used carefully: the relatively small sample sizes in the case of teachers (compared to students) become even smaller when types of programs are examined, the precision of estimates is probably lower because the sampling target was mainly on the national level, and the sets of countries with similar programs are fairly small, while their comparability is limited.

RESULTS

Structure of Primary Teacher Education

In many TEDS-M countries, primary school consists of grades 1 through 6. Germany is an exception, as primary school in most federal states has only four grades. In principle, teacher training can be organized in a concurrent or consecutive way. Germany is also an exception in this respect as its teacher training system combines important features of both approaches ("hybrid system").

Features of Incoming Primary Students

As can be seen across all TEDS-M countries, a typical primary teacher at the end of teacher training is around 24 years old and female. Her parents have a degree of level 3 or 4 on the International Standard Classification of Education (ISCED), and there are between 25 and 100 books in her parents' home, as well as a computer. The average teacher's prior knowledge from schooling is extensive, namely 12 years of mathematics classes and "good" or even "very good" grades compared to her age cohort. The language of teacher training fits with the language spoken in her home. Intrinsic motives, like a desire to educate young people, and intellectual motives dominated her decision to become a teacher much more than extrinsic motives like status rewards.

Unsurprisingly, huge variation exists between countries with regard to these characteristics of primary teachers. It seems that teachers from consecutive programs are older on average than those from concurrent programs. Teachers in the Philippines and Georgia are, on average, only 21 years old at graduation, whereas teachers in Germany are nearly 27 years old. The higher age in Germany is an accumulated consequence of many different societal, schooling and teacher training aspects. Male graduates do not represent the majority of graduating primary teachers in any of the TEDS-M countries. However, there seems to be a tendency for their proportion to increase if the program requires more academic mathematics classes or if they will receive a teaching license to teach higher grades.

In many TEDS-M countries, the educational background of the teachers' mothers and fathers is roughly equal. However, this does not apply to all countries. In Germany, Switzerland and Spain, mothers have, on average, lower degrees than fathers, while in Russia, Poland and Georgia, mothers have higher degrees than fathers. These differences are probably related to the role of women in the respective societies.

The cultural capital of the future primary teachers was measured using several indicators, which pointed in approximately the same direction: in Germany and Norway, the teachers' cultural capital is especially high. The cultural capital of teachers in Georgia and Russia is also strikingly high given their rank on the United Nations' Human Development Index. We hypothesize that this result

reflects high educational aspirations in these societies. In general, it is noteworthy that the teachers' cultural capital is much higher in most countries than the capital of their students (see, for example, the teacher survey in TIMSS). This result points to a possible selection effect.

With regard to the language spoken at home compared to the official language of teacher training (i.e., the test language of TEDS-M), there is a distinct difference between two groups of countries. In one group, which consists of Botswana, Malaysia and the Philippines, future teachers were tested in English as the language of instruction, although this was the language spoken at home only for a small minority. We found that a substantial proportion of teachers speak a different language at home compared to at teacher training in Singapore, Thailand and Taiwan, too. In contrast, there are many countries in which almost every teacher speaks the official language of testing at home, although there is sometimes substantial language diversity in these countries as well (e.g., in Germany and the USA).

Primary teachers in Germany had received lower grades in school than future teachers in other TEDS-M countries – however, this result is probably an artifact. Germany has a highly stratified school system with many options. Therefore, the reference group was more selective than in most other countries.

With regard to motivation, we can once again notice differences between countries. As regards the full set of motives presented in TEDS-M, future teachers in the U.S., Switzerland, Norway, Germany, Spain and Chile have particularly strong educational motives in relation to intellectual or extrinsic motives. We assume that this is due to the long-standing tradition of child-orientated pedagogy in these countries (see, for example, the reception of Ellen Key's famous book *The century of the child* in 1909). In contrast, future teachers in Asian and Eastern European countries specifically stress the intellectual challenge of teaching. We assume that this is due to the high value assigned to mathematics in these countries (and, in the case of some Asian countries, their Confucian heritage and the overall value of teachers). It seems that striving for a license for teaching in higher grades supports more intellectual motives.

The future teachers were asked about the extent to which they felt limited by financial or familial constraints during their studies. The result is once again striking, because there is another split between countries: on the one side, we have countries in which future teachers stress family obligations more than financial worries. This applies to all Asian countries in TEDS-M, as well as Botswana and Chile. On the other side, we have the Western countries and Poland, where financial limitations dominate family issues. It is probably not far-fetched to relate this result to cultural differences, as reflected in Hofstede's (1983, 1993) continuum of collectivism and individualism.

Opportunities to Learn in Primary Teacher Education

The extent of opportunities to study mathematics during primary teacher training varies a great deal between the TEDS-M countries. In Thailand, where specialists

are trained for this particular subject, primary teachers cover most topics. Germany is one of the countries in which the number of opportunities to learn mathematics is significantly below the international average. This result is mainly a function of the fact that mathematics is strongly neglected in one out of four study programs (primary and lower secondary teachers who do not specialize in mathematics, a program that is offered in half of the 16 federal states). Graduates from the other three types of program – primary and lower secondary teachers who specialize in mathematics, pure primary teachers who specialize in mathematics but also pure primary teachers who do not specialize in mathematics (offered in the other half of the 16 federal states) – covered significantly more mathematical topics during their training.

It is possible to identify an international profile of OTL in mathematics which applies to Germany as well: number is a dominant field of study in primary teacher training, followed by data and, within certain limits, geometry. Calculus is of very little importance in most countries. Another commonality across countries is the relatively high amount of OTL taken in general pedagogy, with regard to theoretical as well as practical topics. There seems to be a normative consensus that GPK is a vital part of teacher knowledge. Less agreement exists with regard to MPCK, and especially its theoretical aspect. Germany is one of the countries with the lowest extent of OTL in this field.

Teacher Educators

On average, more than half of the educators in primary teacher training are female. The proportion of teacher educators with a degree at ISCED level 6 (at least a PhD) varies a great deal between the TEDS-M countries: from 0% in Botswana to 82% in Georgia.

Outcomes of Primary Teacher Education

With regard to MCK, future primary teachers from Taiwan achieved the most favorable results of all of the TEDS-M countries (see Table 3). The difference to the international mean of 500 test points was fairly large – more than one standard deviation, which is a highly relevant difference. The achievement of primary teachers from the U.S. was slightly above the international mean and roughly on the same level as the achievement of teachers in Germany and Norway. Their difference to the international mean was significant but of low practical relevance. These groups of teachers also reached significantly lower performance levels than Swiss and Thai teachers. If we take into account the Human Development Index used by the U.N. in order to indicate the social, economic and educational developmental state of a country, the high performance of teachers from Russia and Thailand is especially striking.

With regard to MPCK, the achievement of future primary teachers from the U.S. was roughly on the same level as the achievement of teachers in Norway, which was significantly above the international mean. In this case, the

difference to the international mean was of practical relevance. Teachers from two other countries outperformed the U.S. The difference of the U.S. to the achievement of teachers from Singapore and Taiwan was, however, still highly relevant.

Interesting differences exist with regard to achievement in MCK and MPCK, which require more research. Whereas Singapore is behind Taiwan in the case of MCK, these countries are on the same level in the case of MPCK. With regard to MPCK, Norway and the U.S. are only half of a standard deviation behind the two East Asian countries, whereas this difference reaches one standard deviation with regard to MCK. Malaysia's score for MPCK is around the international mean, while its score is below the mean for MCK. Russia, Thailand and Germany perform significantly lower in MPCK than in MCK. These differences are worth examining in detail. They may point to specific strengths and weaknesses.

Table 3. Knowledge of future primary teachers by country

MCK			*MPCK*		
Country	*Mean*	*SE*	*Country*	*Mean*	SE
Taiwan	623	4.2	Singapore	593	3.4
Singapore	590	3.1	Taiwan	592	2.3
Switzerland*	543	1.9	Norway[1 n]	545	2.4
Russia	535	9.9	U.S.A.[*** 1 3]	544	2.5
Thailand	528	2.3	Switzerland*	537	1.6
Norway[1 n]	519	2.6	Russia	512	8.1
U.S.A.[*** 1 3]	518	4.1	Thailand	506	2.3
Germany	510	2.7	Malaysia	503	3.1
International	**500**	**1.2**	Germany	502	4.0
Poland[** 1]	490	2.2	**International**	**500**	**1.3**
Malaysia	488	1.8	Spain	492	2.2
Spain	481	2.6	Poland[** 1]	478	1.8
Botswana	441	5.9	Philippines	457	9.7
Philippines	440	7.7	Botswana	448	8.8
Chile[1]	413	2.1	Chile[1]	425	3.7
Georgia	345	3.9	Georgia	345	4.9

[1] Combined participation rate <75%
[3] High proportion of missing values
[*] Colleges of education in German-speaking regions
[**] Institutions with concurrent programs
[***] Public universities
[n] The results for Norway are reported by combining the two datasets available in order to present an accurate mean for this country.

With regard to the achievement of primary teachers from different program types, MPCK is taken as an example in this summary (for results concerning MCK, see Blömeke, Kaiser, & Lehmann, 2010a). Unsurprisingly, primary teachers trained as mathematics specialists show the best performance. None of the mean MPCK

scores for this program type are significantly below the international mean of 500 test points. The results of teachers from other programs are more striking. In Taiwan, Singapore and Norway, future teachers from non-specialist programs achieved high scores in MPCK. At the same time, there are huge differences within countries. Poland and Germany are examples of this phenomenon. In these two countries, it is possible to teach mathematics in primary schools with a license from either a general (without an emphasis on mathematics) or a specialist teacher training program (with an emphasis on mathematics). The average MPCK achievement of these programs differs by approximately one standard deviation.

A new field of research is the assessment of teachers' GPK. TEDS-M is the first comparative study to address this concept. Germany, Taiwan and the U.S. assessed knowledge about lesson planning, classroom management and motivation, dealing with heterogeneity and student assessment, with each dimension subdivided into three cognitive tasks (recalling, understanding and creating). The main result on the primary level indicates that German teachers significantly outperform U.S. teachers. The overall difference is approximately one standard deviation as is the difference on each sub-dimension. All differences are therefore highly relevant. Within Germany, graduates from pure primary programs perform significantly better than students from joint primary and lower secondary programs.

Finally, beliefs were captured as teacher-education outcomes in TEDS-M. There is extensive variation between and within countries – however, it is possible to identify profiles which seem to be influenced by cultural features, specifically according to Hofstede's (1983, 1993) continuum of individualism and collectivism. In countries with an individualistic orientation like Germany, future teachers specifically stress the dynamic aspects of mathematics with regard to the static aspects and constructivist principles of teaching and learning in relation to transmission-oriented principles. In contrast, in countries with a collectivistic orientation, the support of static and transmissive aspects is relatively high compared to the support of dynamic and constructivist aspects. Countries which seem to develop from collectivism to individualism according to Hofstede's index are positioned in the middle of the TEDS-M scale as well. If a TEDS-M country deviates from Hofstede's index (e.g., Poland), the specific mathematics tradition in that country may be an explanation. Within Germany, the profile of beliefs varies according to teacher training program type. The more mathematics education a future teacher has received, the more he or she supports dynamic and constructivist beliefs.

Structure of Lower Secondary Teacher Education

Lower secondary school consists in many TEDS-M countries of grades 7 through 9. In principle, teacher training can be organized in a concurrent or consecutive manner. Germany is an exception, as its teacher training system combines important features of both approaches ("hybrid system").

Features of Incoming Students

In a comparison of all TEDS-M countries, mathematics teachers for lower secondary schools are very similar to primary teachers and correspondingly there is a similar degree of variation between countries. One difference applies to Germany, where future lower secondary teachers are nearly 30 years old when they finish teacher training. Another difference applies to the university entrance characteristics of German lower secondary teachers. In many respects, it is necessary to distinguish between types of program. Future teachers at the "Gymnasium" (prepared for teaching grades 5 through 13) – who are trained in longer programs than other lower-secondary teachers (prepared for teaching up to grade 10) and employed as senior civil servants, in contrast to other lower-secondary teachers who are employed as junior civil servants – had, on average, better grade point averages in their high school exit exam "Abitur" and were more likely to have taken advanced mathematics in high school.

Opportunities to Learn in Lower Secondary Teacher Education

The extent of opportunities to learn mathematics, mathematics pedagogy and general pedagogy varies a great deal between the TEDS-M countries, but it is possible to identify an international profile of OTL with regard to the relationship between mathematics and the two pedagogical dimensions. At the end of teacher training, lower secondary mathematics teachers in Poland, Russia, Georgia, Taiwan and Oman, as well as (within limitations) those in Germany and Thailand had specifically extensive OTL in mathematics compared to OTL in mathematics pedagogy and general pedagogy. In contrast, lower-secondary teacher training in Norway, the U.S., Chile and Botswana is more strongly dominated by pedagogical topics. One could say that in the first set of countries, content is emphasized, while in the second set of countries, the teaching of the content is stressed.

With regard to the subfields of mathematics, it is interesting to examine the relationships between OTL in numerics, calculus, data and geometry during teacher training. The largest variation exists regarding calculus. Whereas in Botswana, Singapore, Georgia, Malaysia, Oman and Taiwan, calculus is dominant in relation to the other three fields (which probably implies an orientation of teacher training toward upper secondary school grades), the extent of OTL in calculus is specifically low in Norway, Switzerland, the U.S. and Chile, which probably indicates an orientation toward the lower secondary school grades.

Overall, future teachers who are also going to be licensed to teach mathematics on the upper secondary level had significantly more OTL than future mathematics teachers on the lower secondary level only. Besides Norway and Chile, where lower secondary teachers are trained as generalists, a particularly low amount of OTL was reported in Germany and Singapore, where lower secondary teachers are trained in two subjects.

Outcomes of Lower Secondary Teacher Education

With regard to MCK, future lower secondary teachers from Taiwan, Russia, Singapore, Poland and Switzerland significantly outperformed teachers from the other countries. If we take into account the Human Development Index used by the U.N., the performance of lower secondary mathematics teachers from Russia and Poland is remarkable. With regard to MPCK, the achievement of Taiwanese and Russian teachers is outstanding. The achievement of teachers from Singapore, Switzerland and Russia is also well above the international mean.

Tables 4 & 5. Knowledge of future secondary mathematics teachers by country

MCK of future lower secondary mathematics teachers	
Country	**Mean (S.E.)**
Taiwan	667 (3.9)
Russia	594 (12.8)
Singapore	570 (2.8)
Poland** 1	540 (3.1)
Switzerland*	531 (3.7)
Germany	**519 (3.6)**
U.S.A.*** 1 3	505 (9.7)
International	500 (1.5)
Malaysia	493 (2.4)
Thailand	479 (1.6)
Oman	472 (2.4)
Norway2 n	444 (2.3)
Philippines	442 (4.6)
Botswana	441 (5.3)
Georgia1	424 (8.9)
Chile1	354 (2.5)

MPCK of future lower secondary mathematics teachers	
Country	**Mean (S.E.)**
Taiwan	649 (5.2)
Russia	566 (10.1)
Singapore	553 (4.7)
Switzerland*	549 (5.9)
Germany	**540 (5.1)**
Poland** 1	524 (4.2)
U.S.A.*** 1 3	502 (8.7)
International	500 (1.6)
Thailand	476 (2.5)
Oman	474 (3.8)
Malaysia	472 (3.3)
Norway2 n	463 (3.4)
Philippines	450 (4.7)
Georgia1	443 (9.6)
Botswana	425 (8.2)
Chile1	394 (3.8)

* Colleges of education in German-speaking regions
** Institutions with concurrent programs
*** Public universities
1 Combined participation rate < 75%
2 Combined participation rate < 60%
3 High proportion of missing values
n The results for Norway are reported by combining the two datasets available in order to present an accurate mean for this country.

On this level, a great deal can be learned by distinguishing between program types in addition to an evaluation of national teacher training systems. However, the same caution has to be exercised as on the primary level due to the relatively

small sample sizes, lower precision of estimates, smaller sets of countries and consequently limited comparability. Once again, pedagogical content knowledge is taken as an example (for MCK, see Blömeke, Kaiser, & Lehmann, 2010b).

The TEDS-M data do not necessarily support the hypothesis that teachers from consecutive programs perform better than teachers from concurrent programs. In contrast, the TEDS-M data support the hypothesis that more mathematics leads to better results. For example, German lower secondary teachers who will also teach mathematics in upper secondary schools have outstanding MPCK, which is, on average, on the same level as the MPCK of Russian teachers and significantly higher than the MPCK of teachers from Singapore with a license to teach in lower and upper secondary grades. German mathematics teachers with a license up to grade 10 perform less well. Their average achievement in MPCK is only slightly higher than the international mean for comparable programs.

With regard to GPK, the data revealed that future lower secondary school teachers from Germany and Taiwan significantly outperformed their counterparts from the U.S. The achievement of U.S. future teachers was more than one and a half standard deviations lower than the achievement of German or Taiwanese future teachers. This is a difference which is of high practical relevance. There was no statistically significant difference between teacher achievement in Germany and Taiwan.

Table 6. GPK of future lower secondary teachers

Country	*M*	*SE*	*SD*
Germany	576	4.9	85
Taiwan	572	3.2	52
International	500	2.2	100
U.S.A.	440	3.0	66

CONCLUSIONS

If we are to summarize the main lessons that we have learned from TEDS-M, methodological and substantive aspects have to be mentioned. TEDS-M showed that studies in the field of teacher education are challenging and difficult. Several levels of aggregation must be considered, and each one has its own benefits and limits. From a substantive point of view, we learned that achievement in different domains of teacher knowledge (MCK, MPCK and GPCK) can differ a great deal, as can the achievement of teachers from different programs within a country. Here we can learn the most for policy efforts within countries to improve the effectiveness of a teacher training system. Overall, teacher competence does not seem to be a main function of socio-demographic features, the beliefs of incoming students or the length, structure or content of teacher training programs alone, but a complex amalgam of these influences.

Intertwining of Societal, Educational and Institutional Aspects

Like everyone else, researchers are embedded in their own culture, and so they often overlook matters of culture. This is particularly the case for teacher training, given the unique way in which it incorporates or touches upon many different levels of education and stands at the intersection of education and other social, economic and political forces. This embedded character of the system of teacher training in any one country makes looking beyond that country's experience mandatory in order to recognize the assumptions which drive it, which are all too often taken for granted. The investigation of another teacher training system in a foreign country, for example, and the discovery that it is possible to organize the training differently sheds new light on the domestic system. The recognition of this cultural boundedness of teacher training is an argument for approaching a comparative study in ways that maximize opportunities for cross-cultural communication and the direct examination of concepts (LeTendre, 1999).

As such, language problems become important and are far more demanding to resolve than a "simple" translation of instruments or responses (National Research Council, 2003). Of course, at one level, this is a common, familiar and well-studied aspect of cross-cultural studies, for which there are now widely-used conventions of translation, back translation and so on (Hambleton, 2002). In TEDS-M, the back-translation approach was used: a first translation was done within the countries participating, and then the instrument items were translated back into English by the IEA. Differences between this translation and the initial version were discussed and resolved by agreement between experts.

In teacher training, we would argue that there are more language-related challenges that require attention. The resultant language problem is not only one of costs (which we do not intend to minimize). Instead, it is more fundamentally a problem of cultural boundaries. Many terms from native languages cannot be translated because adequate English terms are missing and vice versa. In the field of education, this problem arises frequently. It is even difficult to name the process by which future teachers learn their profession: is it teacher education, is it teacher training or is it perhaps teacher preparation? These questions relate to deeper and often tacit assumptions about schooling, teaching and learning to teach. As these terms connect to broadly shared cultural beliefs, the uniqueness of their meaning often is not explicit and can easily escape scrutiny unless outsiders to the cultural community stumble over them and begin to enquire about them. Behind the apparently simple choice of whether to refer to the practice as teacher education, teacher training, teacher preparation, or something else, lie other aspects of history, policy, social values and cultural norms.

NOTE

[1] TEDS-M was funded by the IEA, the National Science Foundation (REC 0514431) and the participating countries. In Germany, the German Research Foundation funded TEDS-M (DFG, BL 548/3-1). The instruments are copyrichted by the International Study Center at Michigan State

University, USA (ISC). The views expressed in this paper are those of the author and do not necessarily reflect the views of the IEA, the ISC, the participating countries or the funding agencies.

REFERENCES

Blömeke, S. (2002). *Universität und Lehrerausbildung* [University and teacher education]. Bad Heilbrunn/Obb.: Klinkhardt.

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit der Lehrerbildung [Empirical evidence for the effectiveness of teacher education]. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Eds.), *Handbuch Lehrerbildung* [Handbook Teacher Education] (pp. 59–91). Bad Heilbrunn/Braunschweig: Klinkhardt/ Westermann.

Blömeke, S., & Paine, L. (2008). Getting the fish out of the water: Considering benefits and problems of doing research on teacher education at an international level. *Teaching and Teacher Education, 24*(4), 2027–2037.

Blömeke, S., Kaiser, G., & Lehmann, R.. (2010a). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* [Cross-national comparison of the professional competency of and learning opportunities for future primary school teachers]. Münster: Waxmann.

Blömeke, S., Kaiser, G., & Lehmann, R.. (2010b). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* [Cross-national comparison of the professional competency of and learning opportunities for future secondary school teachers of mathematics]. Münster: Waxmann.

Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Berkeley: University of California Press.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers [Competencies, objectives and classroom performance of teachers]. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie: Psychologie des Unterrichts und der Schule. Bd. 3* [Encyclopedia psychology: Psychology of Teaching, Learning and Schools] (pp. 177–212). Göttingen: Hogrefe.

Bromme, R. (1994). Beyond subject matter: A psychological topology of teachers' professional knowledge. In R. Biehler, R. W. Scholz, R. Straesser & B. Winkelmann (Eds.), *Mathematics didactics as a scientific discipline: The state of the art* (pp. 73–88). Dordrecht: Kluwer.

Brown, D. F., & Rose, T. J. (1995). Self-reported classroom impact of teachers' theories about learning and obstacles to implementation. *Action in Teacher Education, 17*(1), 20–29.

Calderhead, J. (1996). Teachers: Beliefs and knowledge. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 709–725). New York: Macmillan.

Cochran-Smith, M., & Zeichner, K. M.. (2005). *Studying teacher education. The report of the AERA Panel on Research and Teacher Education.* Mahwah, NJ: Lawrence Erlbaum.

Dumais, J., & Meinck, S. (in press a). Sampling design. In M. T. Tatto (Ed.), *Teacher Education Study in Mathematics (TEDS-M) technical report*. East Lansing, MI: Michigan State University.

Dumais, J., & Meinck, S. (in press b). Estimation weights, participation rates, and sampling error (draft). In M. T. Tatto (Ed.), *Teacher Education Study in Mathematics (TEDS-M) technical report*. East Lansing, MI: Michigan State University.

Ferrini-Mundy, J., Floden, R., McCrory, R., Burrill, G., & Sandow, D. (2005). *A conceptual framework for knowledge for teaching school algebra.* East Lansing, MI: Authors.

Grigutsch, S., Raatz, U., & Törner, G. (1998). Einstellungen gegenüber Mathematik bei Mathematiklehrern [Epistemological beliefs of mathematics teachers about mathematics]. *Journal für Mathematik-Didaktik, 19*, 3–45.

Hambleton, R. K. (2002). Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58–79). Washington: National Academy Press.

Helmke, A. (2004). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern* [Quality of instruction: Measurement, evaluation, improvement] (3rd ed.). Seelze: Kallmeyersche Verlagsbuchhandlung.

Hill, H. C., Loewenberg Ball, D., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualising and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

Hofstede, G. (1983). Culture's consequences: International differences in work-related values. *Administrative Science Quarterly, 28*(4), 625–629.
Hofstede, G. (1993). Cultures and organizations: Software of the mind. *Administrative Science Quarterly, 38*(1), 132–134.
Key, E. (1909). *The century of the child.* New York: Putnam.
KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland). (2004). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss (Jahrgangsstufe 10)* [Standards for mathematics in the middle school exit exam (grade 10)]. München: Wolters Kluwer.
Leder, C., Pehkonen, E., & Törner, G.. (2002). *Beliefs: A hidden variable in mathematics education?* Dordrecht: Kluwer Academic Publishers.
Leinhardt, G., & Greeno, G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology, 78*(2), 75–95.
LeTendre, G. K. (1999). The problem of Japan: Qualitative studies and international educational comparisons. *Educational Researcher, 28*(2), 38–45.
National Research Council. (2003). Understanding others, educating ourselves: Getting more from international comparative studies in education. Washington, DC: National Academies Press.
NCTM. (2000). *Principles and standards for school mathematics*. Reston: NCTM.
NOKUT (Nasjonalt Organ for Kvalitet i Utdanningen). (2006). *Evaluering av Allmennlærerutdanningen i Norge 2006. Hovedrapport* [Evaluation of general teacher education in Norway: Main report]. Retrieved from http://www.nokut.no/Documents/ NOKUT/ Artikkelbibliotek/Norsk_utdanning/SK/alueva/ALUEVA_Hovedrapport.pdf
Richardson, V. (1996). The role of attitudes and beliefs in learning to teach. In J. Sikula, T. Buttery & E. Guyton (Eds.), *Handbook of research on teacher education* (2nd ed., pp. 102–119). New York: Macmillan.
Schmidt, W. H., Blömeke, S., & Tatto, M. T. (2011). *Teacher education matters. A study of the mathematics teacher preparation from six countries.* New York: Teacher College Press.
Shulman, L. (1985). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 3–36). New York: Macmillan.
Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics: Conceptual framework.* East Lansing, MI: College of Education, Michigan State University.
Thompson, A. G. (1992). Teachers' beliefs and conceptions: A synthesis of research. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 127–146). New York: Macmillan.
UNESCO. (1997). *International standard classification of education – ISCED*. Paris: UNESCO.
Weinert, F. E. (2001). Concepts of competence: A conceptual clarification. In D. S. Rychen & L. H. Salgnik (Eds.), *Defining and selecting key competencies* (pp. 45–66). Göttingen: Hogrefe.
Wilson, S., Floden, R., & Ferrini-Mundy, J. (2002). Teacher preparation research: An insider's view from the outside. *Journal of Teacher Education, 53*(3), 190–204.

Sigrid Blömeke
Institute of Education,
Humboldt-University of Berlin, Germany

KARINE TREMBLAY

OECD ASSESSMENT OF HIGHER EDUCATION LEARNING OUTCOMES (AHELO):

Rationale, Challenges and Initial Insights from the Feasibility Study

INTRODUCTION

This paper summarizes a presentation of the OECD AHELO initiative which was delivered at the Conference on Modeling and Measurement of Competencies in Higher Education organized in Berlin (Germany), February 24–25, 2011.

RATIONALE OF AHELO INITIATIVE

Higher Education in Global Mutation

Higher education today is a critical factor in innovation and human capital development and plays a central role in the success and sustainability of the knowledge economy. Countries and individuals around the world recognize the strategic value of investing in higher education, and as a result higher education has been fast expanding globally for the last few decades. Nowadays, it is estimated that some 135 million students are enrolled worldwide in more than 17 000 higher education institutions. Ten years into the third millennium, however, higher education systems worldwide are faced with profound mutations, as illustrated by the following trends.

First and foremost, higher education has over the past four decades moved from elite access to mass participation in all OECD countries – and increasingly so in non-OECD countries as well. This trend is exhibited by changes in attainment rates of different age cohorts within countries, whereby younger age groups are more likely to hold a post-secondary degree than older generations[1] (OECD, 2010a). With a number of emerging economies now catching up with their OECD peers in terms of educational attainment, this trend is likely to continue in the years to come, and analysts at a recent UNESCO conference estimated that an additional 125 million students would have to be accommodated by 2025 worldwide (UNESCO, 2011). To put the figures in perspective, this would be the equivalent of building about four new universities every week!

Despite this massive increase in higher education participation and completion, post-secondary qualifications still have a very high economic value for degree

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 113–126.

holders, whether in terms of labor market opportunities and employability vis à vis secondary graduates, or net income over one's lifetime[2] (OECD, 2010a). The reason for this has to do with changing skill demands in the move to knowledge economies. Indeed, manual and routine cognitive skills are on the decline in the twenty-first century whereas there has been an increasing and accelerating demand for non-routine analytic and interactive skills since the 1990s (Levy & Murnane, 2004).

Although higher education credentials retain high economic value on the labor market, there are persistent concerns related to the quality and relevance of higher education. Indeed, the context for contemporary higher education is characterized by greater heterogeneity of students than in the past in terms of abilities and expectations, the multiplication of new institutions of higher education to accommodate the massification of participation – including some rogue providers – and significant non-completion in most OECD countries. In response, quality assurance systems have been adopted and developed in most OECD countries over the past two decades (OECD, 2008a). Yet despite significant progress in this area, quality assurance systems have not eradicated failure and inefficiencies in the learning process. Nowadays, an average three out of ten students entering a post-secondary program in the OECD will drop out without obtaining at least a first tertiary degree[3] (OECD, 2010a). The economic and social costs of dropout and underachievement are substantial, given the high costs per student in higher education, the social outcomes of education and the contribution of higher education to social mobility and equity.[4] Available evidence on this front suggests that disadvantaged students are not only less likely to access higher education, they are also more likely to drop out, thereby reinforcing inequality (OECD, 2008a).

Meanwhile, higher education systems are increasingly globalized and internationalized. This is most visible in the considerable growth in international student numbers over the past three decades – that is, students enrolled outside their country of origin[5] (OECD, 2010a). A recent review of internationalization policies in OECD countries revealed that other forms of internationalization such as the mobility of programs and institutions and the internationalization of curricula and teaching in English were also on the rise (OECD, 2008a).

There is also evidence of an increasing internationalization of the labor markets for the highly-skilled. On the one hand, there is growing international organization of high-skilled professions. On the other hand, migration policies across the OECD are increasingly skewed towards high-skilled immigrants and it is estimated that an average 21% of international students stay on to work in their country of study, thereby contributing to high-skilled intakes of immigrants (OECD, 2010b).

THE FEASIBILITY STUDY CONCEPT

The above trends all point to the need for greater attention to quality, transparency and the international recognition of higher education credentials. In this context, policy-makers as well as the public devote much attention to the outcomes of higher education, but face a considerable information gap in their efforts to improve the quality of teaching and enhance the learning outcomes of students.

There is no reliable information which enables comparative judgments to be made about the capabilities of students in different countries and different institutions, or about the quality of teaching. Proxies of higher education quality do exist, but none are perfect. The reputations of institutions are highly subjective. Likewise, international rankings are biased towards available data on inputs and research excellence and fail to provide an accurate picture of teaching at various institutions. Other proxies such as satisfaction rates or labor market outcomes are also sensitive to cultural norms or conjuncture and local economic conditions.

Developing measures that give due weight to teaching practices and learning outcomes has thus become essential, but the uncertainties and doubts of some actors remain as to whether it is scientifically and operationally feasible to measure learning outcomes across higher education institutions of very different types, and in countries with different cultures and languages.

In response, the Organisation for Economic Co-operation and Development (OECD) launched in 2008 a feasibility study to explore the viability of developing an international Assessment of Higher Education Learning Outcomes (AHELO) that would measure learning outcomes in ways that are valid across cultures and languages – but also taking into account the diversity of institutional settings and missions.

The underlying motivation of AHELO is that information on learning outcomes could contribute to higher education institutions' knowledge of their teaching performance, and thereby provide a tool for development and improvement. Consequently, the AHELO emphasizes improving teaching and learning and providing higher education leaders with tools to empower them and foster positive change.

OVERVIEW OF THE FEASIBILITY STUDY

Purpose

The goal of the AHELO feasibility study is to assess whether it is possible to measure at the international level what undergraduate degree students know and can do upon graduation. The implication of this feasibility study methodology is a research approach to the work.

In exploring the feasibility of an international assessment of learning outcomes in higher education, the OECD recognizes that quality is a multidimensional concept which cannot be aggregated to a single measure which could then be used for a unidimensional ranking of institutions. On the contrary, AHELO aims to develop a range of different performance metrics ranging from summative performance indicators to measures that capture the learning gain or "value-added" at an institution. The context dimension is also prominent in the AHELO approach and the feasibility study will identify context factors that are important in contextualizing performance and putting results into perspective. Efforts will also focus on the identification of context variables which are associated with higher learning outcomes with a view to shedding light on effective teaching and learning strategies and approaches.

The feasibility study aims to test both the science of the assessment and the practicality of implementation. To do so, the work will unfold in two phases:

- The first phase from August 2010 to June 2011 has consisted in providing an initial proof of concept. In this phase, the goal was to develop provisional assessment frameworks and testing instruments suitable for an international context for each of the three strands of work, generic skills, economics and engineering, and to validate those tools through small-scale testing (cognitive labs and think-aloud interviews) in participating countries in order to get a sense of cross-linguistic and cross-cultural validity. The focus has been on the feasibility of devising assessment frameworks and instruments that have sufficient face validity in various national, linguistic, cultural and institutional contexts.
- In a second phase from March 2011 to December 2012, the goal is to evaluate the scientific and practical feasibility of an AHELO by focusing on the practical aspects of assessing students' learning outcomes. During this phase, the implementation of assessment instruments and contextual surveys in small groups of diverse higher education institutions will explore the best ways to implicate, involve and motivate leaders, faculty and students, encouraging them to take part in the testing, but will also look at the relationships between context and learning outcomes, and the factors leading to enhanced outcomes. This second phase will address issues of practical feasibility and assess data reliability.
- Should these two phases demonstrate the feasibility of assessing student learning outcomes in countries with different cultures and languages, a further phase will consist in developing a value-added measurement strand to explore methodologies and approaches to capture value-added or the contribution of higher education institutions to students' outcomes, irrespective of students' incoming abilities.

With the completion of the feasibility study, the information collected on student performance and the analysis of the results will help to assess from both scientific and practical standpoints whether a fully-fledged AHELO study could feasibly be taken forward. The outcomes of the AHELO feasibility study will guide the decision to be made by the OECD member countries on whether to launch a fully-fledged study in the longer term.

AHELO'S FOUR STRANDS OF WORK

Given the two key aims of the AHELO feasibility study, the focus of AHELO is on providing proof of concept by exploring different approaches, methodologies and instruments that might eventually be envisaged as parts of a fully-fledged assessment. Consequently, the AHELO feasibility study has been designed to cover four different strands of work, as illustrated in Figure 1.

Figure 1. AHELO's four strands of work.

In each strand, the approach chosen is not to develop comprehensive assessment instruments, but rather to explore how best to assess student performance in higher education institutions around the world.

– A cross-discipline strand: the generic skills strand

  The generic skills strand is an essential component of the feasibility study. Indeed, transversal higher-order competencies such as critical thinking, analytic reasoning, problem-solving or the generation of knowledge and the interaction between substantive and methodological expertise are widely viewed as critical for the success of individuals in the information age. It is therefore important for an AHELO to measure those generic skills, and not only cognitive knowledge, which are necessary for success in both academic and business contexts.

  For the purpose of the feasibility study, the decision has been made to rely on an existing instrument – the US-based Collegiate Learning Assessment – and to gauge whether it can be applied in other national and cultural contexts with adequate translation and adaptations. The CLA measures focus on skill sets that students will need as they graduate and enter the workforce, namely critical thinking, analytical reasoning, problem-solving and written communication. These skills are intertwined. Thus the CLA requires students to use these skills together to respond to tasks. CLA performance tasks use open-ended prompts that require constructed responses. Each task has an accompanying library of information which students are instructed to use in preparing their answers. These tasks often require students to marshal evidence from diverse quantitative and qualitative sources such as letters, memos, summaries of research reports, maps, diagrams, tables, etc., and to exercise judgment on their reliability or relevance (e.g. scientific evidence vs. rumor, misinterpreted data, etc.).

– Two discipline strands: Economics and Engineering

  Despite the fact that generic competencies underlie most facets of undergraduate education, institutions and learners invest most of their effort in discipline-specific knowledge and skills. There would thus be strong limitations in adopting an approach entirely restricted to generic competencies insofar as it would not assess the kind of subject-matter competencies that most higher education departments or faculties would consider their primary work.

  For the purposes of the feasibility study, AHELO focuses on assessing learning outcomes in two contrasted disciplines representing both scientific and social sciences domains. Economics and engineering have been chosen because these are common disciplines among higher education institutions in

OECD countries, they are relatively different in terms of substance and context, they are less likely to be influenced by unique cultural features, and they reflect the dynamics of disciplinary change. The economics and engineering assessments will thus help to gauge the viability of measuring discipline-specific skills and proving the AHELO concept, with a view to expanding the number of disciplines covered over time should AHELO evolve towards a fully-fledged AHELO main study in the future.

The focus of the disciplinary assessments is to measure competencies that are not only fundamental but also "above content," i.e. competencies indicating students' capacity to extrapolate from what they have learned and apply their knowledge in novel contexts unfamiliar to them. In this regard, the AHELO approach follows the PISA dynamic model of lifelong learning in which new knowledge and skills necessary for successful adaptation to a changing world are continuously acquired throughout life. AHELO focuses on aspects that higher education students will need in the future and seeks to assess what they can do with what they have learned. With this approach, the development of the assessment instruments is not constrained by the common denominator of program curricula which are very diverse in higher education. This is an important prerequisite if an AHELO is to be relevant to a range of different types of institutions. Instead, the disciplinary assessments examine students' ability to reflect, and to apply their knowledge and experience to novel and real-world tasks and challenges.

The development of frameworks and instruments is done by *Education Testing Service* (ETS) in the case of Economics and the *Australian Council for Education Research* (ACER) in cooperation with Japanese and European research groups in the case of Engineering.[6] The instruments mix multiple choice and open-ended questions in proportions of 30 open-ended and 60 multiple choice questions in the case of Economics and 60 open-ended and 30 multiple choice questions in the case of Engineering. As a result, the discipline strands will also help in assessing the optimal mix of question item types and the balance that should be given to open-ended and multiple choice question items in an eventual AHELO main study.

– A research-based strand: the value-added measurement strand

Should the assessment instruments demonstrate the feasibility of assessing student learning outcomes across different countries, languages, cultures and types of institutions, a third phase of the feasibility study might be launched with a value-added measurement strand to explore methodologies and approaches to capture value-added or the contribution of higher education institutions to students' outcomes, irrespective of students' incoming abilities.

Measuring value-added in higher education, i.e. the learning gain that takes place during the higher education experience, imposes layers of complexity that, though theoretically well understood, are still challenging in the context of large-scale assessments. Given the complexity of measuring marginal gain, the feasibility study would first scrutinize possible methods

for capturing marginal learning outcomes that can be attributed to attendance at different higher education institutions, both from a conceptual/theoretical perspective and in terms of psychometric approaches. It will build upon similar work carried out at school level by the OECD (OECD, 2008b) and review options for value-added measurement in higher education. Researchers will be invited to study potential data sources, methodologies and psychometric evidence on the basis of datasets existing at the national level, with a view to providing guidance towards the development of a value-added measurement approach for a fully-fledged AHELO main study.

Figure 2 below summarizes the substantive focus and main features of the feasibility study's various strands of work.

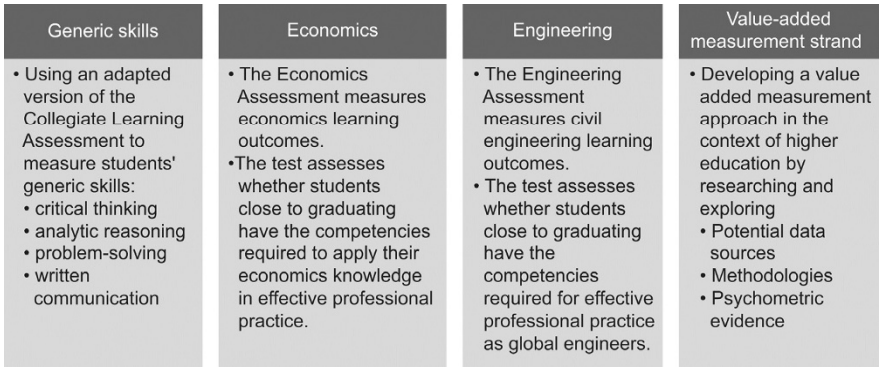


Figure 2. Substantive focus of the four AHELO strands of work.

Adding a Contextual Dimension

Although the main focus of the AHELO feasibility study is to gauge the feasibility of assessing learning outcomes, it is also necessary to assess the feasibility of gathering contextual variables that will be needed to interpret performance measures and help institutions understand the performance of their students and improve their teaching accordingly. The goal of contextual data is a dual one. A primary objective of contextual data is to insure fair comparisons between peer institutions, i.e. compare like with like. A secondary but equally important objective is to enable better understanding of the black box of teaching and learning in higher education with a view to understanding what works, and for whom and in which contexts.

To this end, the contextual variables will allow for disaggregation of assessment results by different kinds of institutional/program characteristics and student populations, and they will provide information to help construct appropriate comparisons across institutions. Further, the contextual data collected through

student, faculty and institution instruments will be used to rehearse some psychometric analyses to identify relevant contextual variables for longer-term development and demonstrate the analytical potential of AHELO for institutional improvement.

This aspect of the feasibility study is led by the *Center for Higher Education Policy Studies* (CHEPS) at the University of Twente in the Netherlands and the *Center for Postsecondary Research* (CPR) at Indiana University in the United States. It also requires assurance that the contextual surveys developed are internationally valid and reflect the cultural context of the countries in which the AHELO feasibility study is implemented. The contextual information will be collected from existing documentation at the country level and through three surveys: a student survey, a faculty survey and an institution survey – each of them taking about 10 minutes to complete.

Remarks on Feasibility Study Data Collection

The data collection period for the AHELO feasibility study will take place in the first half of 2012. It will focus on students who are nearing completion of their first three to four-year degree (Bachelor type of program).

Although students will be tested individually, the unit of analysis and focus of reporting is the higher education institution in the AHELO approach. Differently from the OECD Programme for International Student Assessment (PISA), which focuses on 15-year-old students, AHELO is not expected to develop comparative data at national or sub-national level. The focus is really on institutions and programs to assist them in identifying their strengths, weaknesses and areas for institutional policy intervention.

For a feasibility study, it is important to involve a range of diverse countries and institutions in the work, as a way to prove the AHELO concept. Therefore, the AHELO feasibility study has been designed to insure the participation for each strand of work of a minimum of four or five countries with cultures and languages as diverse as possible. Similarly, each participating country has been instructed to select a range of higher education institutions that should reflect the diversity of higher education at national level. The required numbers of participants are large enough to assess the measurement properties of the various instruments, and small enough to keep the process manageable and avoid validity gains being sacrificed to efficiency gains.

In this respect, AHELO has been quite successful in attracting participating countries from all continents and from a variety of language families as the distribution of countries across strands illustrates:

- **Generic skills:** Colombia, Egypt, Finland, Korea, Kuwait, Mexico, Norway, the Slovak Republic and the US (Connecticut, Missouri, Pennsylvania).
- **Economics:** Belgium (Fl.), Egypt, Italy, Mexico, the Netherlands, the Russian Federation and the Slovak Republic.

– **Engineering:** Australia, Canada (Ontario), Colombia, Egypt, Japan, Mexico, the Slovak Republic and Sweden.

CHALLENGES

The key challenges of the feasibility study can be framed in terms of a set of questions to be examined and addressed.

With respect to scientific feasibility, a first initiative is to assess whether it has been possible to develop assessment instruments to capture learning outcomes that are perceived as valid in diverse national and institutional settings. In a second stage, the feasibility study will also check whether the test items perform as expected and meet predefined psychometric standards of validity and reliability. The latter is tied to the issue of scoring and whether it is possible to score higher-order types of items consistently across countries and institutions. Given the ultimate goal of AHELO to provide a tool for development and improvement, the feasibility study will also assess whether it has been possible to capture information on teaching and learning contexts that contributes to explaining differences in student performance.

As regards practical feasibility, a first question relates to whether participating countries and institutions have managed to put in place effective strategies to secure institutional and student participation in the feasibility study assessments. A corollary is whether it has been possible to motivate students to take the tests seriously and give of their best. Since AHELO ultimately aims to support institutions in their improvement efforts, it would also be important to know whether the implementation of the feasibility study has brought valuable information, insights and benefits to participating institutions, including in demonstrating the value of such assessments to improve teaching and in building support for an AHELO.

INITIAL INSIGHTS

Overall Progress

As of August 1, 2011, the assessment frameworks and instruments for the three strands of work have been validated through small-scale qualitative testing and are currently being reviewed by expert groups and participating countries. The contextual dimension framework and the three survey instruments have also been developed and are under review by participating countries.

The field implementation of the assessment instruments and contextual surveys is scheduled to start at the beginning of 2012. The field testing of those instruments and surveys will involve about one to two hundred students per institution, and about ten higher education institutions per country. By the end of 2012, the information collected on student performance and the analysis of the results will help to assess from both scientific and practical standpoints whether a fully-fledged AHELO study could be taken forward.

Although results from field testing are yet to come, some intermediate findings and insights can be reported now.

Generic Skills Strand

This part of the study is well on its way to proving that an international tool to measure generic skills can indeed be developed. Two performance tasks, chosen by participating countries, and their supporting materials have been translated and culturally adapted in the different languages of participating countries. The tasks have also been put to the test by students from countries as different as Finland, Korea, Kuwait, Mexico and Norway, who have taken the CLA test in their own language and provided qualitative feedback on validity.

With respect to instrument development, the international cooperation on the cultural adaptation and translation of the CLA performance tasks has already provided valuable lessons on instrument development and translation/adaptation processes. Indeed, the initial adaptation of the performance tasks by participating countries was minimal (names, city/government structures, date format), but although the translation process went smoothly, it brought some adaptation issues to light.

The conduct of cognitive labs in participating countries subsequently indicated that the performance tasks functioned as anticipated and can be considered valid, with slight modifications to help with understanding. It was noted however that the performance task concept was unfamiliar in some countries, and should AHELO develop into a main study, it would probably be necessary to provide students with an example prior to test administration, including an exemplary answer, to get a sense of the expected response format. Finally, one of the unexpected findings from the cognitive labs was that the perceived confidence and trust in the sources of information provided to students as accompanying test materials seemed to differ according to different national contexts. In particular, the reliability of information from government sources was interpreted very differently depending on the prevalence of corruption across countries.

These initial insights have thus brought to light some cultural issues that would have to be addressed in an AHELO main study, but no major hurdle. All in all, they suggest that measuring generic skills across languages, cultures and types of institutions seems feasible.

Disciplinary Strands

Likewise, the development of the economics and engineering instruments and their testing with focus groups of students should provide further insights into the feasibility of developing internationally valid assessment instruments in the disciplines, as well as some initial data on the reliability of the measures. As with the generic skills strand, the focus will primarily be on whether an instrument can be developed that has face validity across different cultural and linguistic contexts.

A prerequisite for this is to reach international agreement on expected learning outcomes in the two disciplines considered in the feasibility study, to provide proof of the concept that it is possible to develop domain assessment frameworks in the disciplines within the context of great curriculum diversity in higher education programs.

Early progress on this front has been made with the Tuning approach, which has been successfully applied in Europe in many disciplinary fields and is now being piloted in other parts of the world. In this respect, it is noteworthy that the outcomes of the Tuning-AHELO project have already demonstrated that reaching agreements on expected learning outcomes can be achieved across diverse national and cultural settings, and in contrasting disciplines (Tuning Association, 2009a, 2009b).

On this basis, economics and engineering assessment frameworks have been developed by international groups of experts, with a view to demonstrating that agreements on domain definition can be reached in two disciplinary fields as distinct as economics and engineering. The frameworks are based on the Tuning-AHELO documents and in the case of Economics on the United Kingdom's "QAA subject Benchmark Statement for Economics 2007." These frameworks have provided the basis for instrument development and their subsequent translation and adaptation.

There was a lot of uncertainty about the feasibility of getting academics from different countries to agree on which learning outcomes to measure in the disciplines, and to agree on an assessment instrument. These doubts were part of the rationale for including an economics strand in the feasibility study, to gauge whether agreement was possible in a social science field. One of the remarkable findings of the feasibility study to date is that it has in fact been easier than anyone thought to get economics experts to agree on what an AHELO should cover and measure. The reason for this is that AHELO goes above content knowledge and focuses rather on the "language" of economics.

The qualitative validation of both instruments has been conducted with focus groups involving small numbers of students at a range of institutions within each country. Both students and faculty were invited to provide feedback on the instruments. Initial feedback from the focus groups suggests that the authentic scenario tasks that have been developed stimulate students' interest in the tasks and engage them.

Although these initial insights only address some of the questions raised by the feasibility study, they provide encouraging signals about the viability of an AHELO which the second phase of the work will now explore in greater detail.

NEXT STEPS AND LONGER-TERM POTENTIAL OF AHELO DATA

The outcomes of the AHELO feasibility study in late 2012 will inform the decision by OECD countries on whether to proceed to the launch of a fully-fledged AHELO main study. Should a decision be made to launch an AHELO main study upon completion of this feasibility study, AHELO would offer great potential for the higher education sector in the longer term.

First and foremost, the availability of information on students' actual learning outcomes would provide institutions with an objective diagnosis and a benchmarking tool to gauge their teaching and learning performance in absolute terms and relative to peer institutions. Such diagnosis and benchmarking are critical for identifying areas of institutional strength, and also those areas where improvements could be made. Objective evidence and diagnosis are the first step towards a quality improvement plan at the institutional level.

In addition, the rich set of data that an AHELO would generate would permit comprehensive analyses of the relationships between outcomes and a range of context variables with a view to identifying some of the features associated with higher outcomes. In this respect, AHELO has the potential to move the education research frontier further as regards effective teaching and learning approaches in higher education. In the longer term, the outcomes of this research will equip institutions with evidence-based recommendations and tools to improve their teaching and the learning outcomes of their students.

With this knowledge base at hand, AHELO could contribute to addressing some key challenges of higher education:

– In terms of equity, to the extent that AHELO could help identify effective teaching and learning strategies for specific groups of students – including those from disadvantaged backgrounds – thereby promoting success and completion for all students.
– In terms of responsiveness, to the extent that the development of information on actual learning outcomes could initiate a debate with stakeholders on whether higher education develops adequate skills and competencies vis à vis societal demands.
– In terms of effectiveness, to the extent that objective information on learning outcomes and institutional areas of strength could help prospective students in choosing an institution which is appropriate for them.
– In terms of impact, to the extent that objective comparative information on learning outcomes would facilitate credit transfers and degree recognition and reduce the risk inherent in international student mobility, thereby fostering mobility of students and highly skilled workers.

There are many more outcomes for the higher education sector in the short term, however.

First and foremost, the launch of the AHELO feasibility study has raised awareness of and advocated the use of quality data in higher education and by implication contributes to shifting emphasis from the research performance of institutions towards placing greater weight on their teaching mission. A longer-term outcome in this respect will be to spur reflection – once learning outcomes are defined and measured – on their relevance to the needs of the workforce.

For participating institutions, a short-term outcome will be to obtain objective information and benchmarks on their teaching and learning practices and on their outcomes. In the longer term, the conduct of the feasibility study will allow them to build capacity by assessing learning outcomes and using quality data to improve

student performance. They will also see the benefits of international exchange and discussions in this area.

IN CONCLUSION…

Despite the fact that higher education has been expanding fast and is key to success in the knowledge economy, there is still a significant gap of information about its quality. There are no tools available to compare the quality of teaching and learning in higher education institutions on an international scale. The few studies that do exist are nationally focused, and international university rankings are based on reputation and research performance, and do not reflect the quality of teaching and learning, nor the diversity of institutions' missions and contexts.

In this respect, the AHELO initiative is a unique and innovative attempt to fill this gap and to develop criteria that would make it possible to evaluate the quality and relevance of what students learn in institutions around the world. For frontline higher education practitioners – from academics to institutional leaders – AHELO would provide valuable information on effective teaching strategies to enhance learning outcomes. Students, governments and employers would also stand to benefit since AHELO would shed light on whether the considerable resources invested in higher education are being used effectively, and on the capacities of graduates to enter and succeed in the labor market.

This paper has shared some emerging insights of the work which are promising and encouraging, but the final conclusions from the feasibility study will not be available until late 2012. It is only on this basis that OECD member countries will decide whether to launch a fully-fledged study in the longer term.

Information about AHELO and the unfolding of the feasibility study can be found on the AHELO website at www.oecd.org/edu/ahelo.

NOTES

1 Education at a Glance, indicator A1.
2 Education at a Glance, indicators A6 and A7.
3 Education at a Glance, indicator A4.
4 Education at a Glance, indicators B1 and A9.
5 Education at a Glance, indicator C2.
6 The National Institute for Education Research (NIER) in Japan and, in the case of engineering, the Florence School of Engineering on behalf of the EUropean and Global ENgineering Education (EUGENE) network.

REFERENCES

Collegiate Learning Assessment (United States). Retrieved from www.cae.org

Levy F., &. Murnane, R. (2004). *The new division of labor: How computers are changing the way we work*. Princeton, NJ: Princeton University Press and Russell Sage Foundation.

Organisation for Economic Co-operation and Development (OECD). (2008a). *Tertiary education for the knowledge society*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2008b). *Measuring improvements in learning outcomes: best practices to assess the value-added of schools*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2010a). *Education at a glance*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2010b). *International migration outlook*, Paris: OECD.

Tuning Association. (2009a). *A Tuning-AHELO conceptual framework of expected/desired learning outcomes in the Science of Economics, OECD*. Retrieved from www.oecd-ilibrary.org/education/tuning-ahelo-conceptual-framework-of-expected-and-desired-learning-outcomes-in-economics_5kghtchwb3nn-en

Tuning Association. (2009b). *A Tuning-AHELO conceptual framework of expected/desired learning outcomes in Engineering, OECD*. Retrieved from www.oecd-ilibrary.org/education/a-tuning-ahelo-conceptual-framework-of-expected-desired-learning-outcomes-in-engineering_5kghtchn8mbn-en.

UNESCO. (2011). *UNESCO global forum on rankings and accountability in higher education: Uses and misuses*, 16–17 May 2011, UNESCO, Paris.

Karine Tremblay
Directorate for Education
Organisation for Economic Cooperation and Development (OECD), France

ROGER BENJAMIN

THE PRINCIPLES AND LOGIC OF COMPETENCY TESTING IN HIGHER EDUCATION

INTRODUCTION

The aim of this chapter is to argue that more attention should be devoted to student learning assessment in higher education, using multiple types of instrument, some of which link directly to teaching and learning in the future. Examples from the case of the U.S. will be used throughout, as this case is illustrative of the underlying trends faced by all countries, to one degree or another. The essay moves through these steps in order to justify the following assertions. A number of trends suggest that the next few decades will bring substantial restructuring in the higher education sector. This restructuring will require much more evidence-based decision-making because the stakes will be high. In turn, this places the focus on the quality of student learning, which is a critical outcome of higher education institutions. Second, the challenges ahead are sufficiently serious that widespread debate will occur about how to resolve them. Third, the work required to generate appropriate responses in order to assess student learning will be discussed, including the central role of faculty. Fourth, the case for performance assessment, now being widely explored in the U.S., will be presented, followed by a short set of recommendations.

THE RATIONALE FOR EVIDENCE-BASED DECISION-MAKING

A combination of factors has created an unprecedented crisis in undergraduate education in the U.S.[1] Access deficits caused by 47 million high school dropouts (equaling one-sixth of the U.S. population), a college readiness gap signified by 40% of new students who cannot read, write or perform math at college level, and only 57% of students graduating within six years present enormous challenges to higher education in the U.S.[2]

The 47 million high school dropouts alone constitute a massive deadweight on the economy. These citizens have been denied the tools they need in order to be productive; they are largely employed in minimum wage service jobs or are not even in the labor force. Their economic and social prospects are bleak, and as they represent such a large percentage of society, America's prospects for both economic growth and reducing inequality are also becoming increasingly problematic.

Rising costs, now combined with declining revenues in higher education, make it much more difficult to reverse this situation. Instead, they exacerbate what has already become an overall market failure, an example of what political economists such as recent Nobel laureate Elinor Ostrom term a "common pool problem" (CPP).

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 127–136.

CPPs arise whenever there is confusion or conflict over a public good, e.g., regarding who pays what proportion of the costs and who gets what proportion of the benefits, or where one person's use affects another's ability to use the assets, or when groups, public or private, fail to provide the resources, over-consume them and/or fail to replenish them. American higher education is a public good. As with other public goods, an attempt has been made to shift a great deal of the responsibility for funding public higher education (which is attended by the majority of students) from the state governments to the students through higher tuition fees. However, while tuition has increased significantly, imposing higher financial burdens on students and their families, annual funding increases have not kept pace with cost increases in higher education, which have been, on average, 1% or more higher than the Cost of Living Index (CPI) (see Griswold, 2006). The net effect has been accelerating cuts to academic programs, student aid and infrastructure for colleges. Under these conditions, little headway has been made in providing training opportunities for the large high school dropout group which keeps increasing. When these kinds of conditions apply to a public good like higher education, it warrants the CCP label. When the CPP becomes acute, as in today's higher education sector, either bold action must be taken to solve it or the CPP will become a permanent crisis (see Grant Hardin's tragedy of the commons, 1968).

Efforts made thus far to deal with the CPP have been defeated because the system of incentives that guide the behavior of college administrators, faculty, staff and other supporting stakeholders is not organized to consider it a problem or, in any case, is not focused on solving it. The incentive system of research universities, and the model for faculty and administrators in the rest of the postsecondary education sector, privileges research and scholarship, not teaching and learning. High school dropouts are not seen as coming under the scope of the postsecondary education sector's mission.

All major institutions in society are highly resistant to change from within; the postsecondary education sector is no exception. The imperative for redesign to deal with the CPP, if it arises, is most likely to stem initially from external sources. The elements that make the imperative possible are plain to see. The CPP, framed by a combustible mixture of rising costs and declining revenues, is now viewed through the emerging consensus that human capital (the knowledge, experience and education levels of a nation's citizens) is clearly the principal resource in the U.S. This consensus should eventually lead to widespread agreement among public and private leaders that as the principal duty of the state is to guarantee the security of its citizens, this now means preserving and enhancing the quality of its human capital.

In short, there will soon be a significant national debate in the U.S. about the undergraduate CCP that will cut across political party lines. We can see this presaged in debates about K-12 education. If human capital is the principal national resource, education should be recognized as the key to success in all other policy areas, such as health, economics, the environment, energy, agriculture and national security. This means that the quality of education should be the central priority of the national government. The human capital argument will create the basis for new and higher limits for the role of education, because leaders will

connect the dots and come to understand the critical importance of dealing with the common problem facing K-16 education, which dwarfs all other issues that America is currently facing.

THE CHALLENGES AHEAD

Will the combination of rising costs and declining resources, coupled with the growing perception of the centrality of human capital, provide strong enough external forces to compel the resolution of the CPP? This combination is strong enough to provoke a national debate about the need to tackle the CPP, but this does not mean that the CPP will be resolved.

Unless a way is found to create an effective institutional redesign strategy that faculty and administrators in postsecondary education will buy into, the next decade will be a period of turmoil, with continuing cost and resource problems accompanied by growing quality and access deficit issues. For example, because 80% of the potential growth in access to postsecondary education will come from the Hispanic population, many of whom are high school dropouts, improvements to access, retention and graduation rates will be problematic without new approaches to the problem.

The CPP, combined with the effects of the disruptive force of Internet-based education solutions, has created the prospect of substantial restructuring and redesigning of the postsecondary education sector over the coming decade. Examples include the following:

- Mission differentiation in order to address the need to sharpen the focus of colleges instead of pursuing multiple missions simultaneously;
- Identification of the gaps in the quality of student learning between African-American and Hispanic students on the one hand, and non-Hispanic Whites and Asian-Americans on the other hand, with analysis-based recommendations about what to change in order to reduce the gaps;
- Description of the extent and nature of student learning deficits at the national, state and institutional level, as defined by the Carnegie classification. Development and implementation of recommendations to improve student learning;
- Identification and implementation of advances in pedagogy that will improve student learning outcomes;
- Assessment of the benefits and costs of advances in educational technology with regard to student learning growth;
- Description and analysis of the impact of resources on student learning outcomes.

The benchmarking of student learning outcomes, which is only now becoming widespread, is a necessary but not sufficient prerequisite for these and all other prospective attempts to restructure and redesign postsecondary education, in an effort to respond to the CPP. This is because one needs a metric against which to

evaluate the benefits over costs against the dependent variables that one is attempting to solve or improve. The outcomes of student learning in undergraduate education are appropriate candidates for this role.[3]

Regardless of what actual scenario plays out over the difficult decades while lie ahead for postsecondary education, empirical evidence, including that which is based on the assessment of student learning, will play a much greater role. Without evidence based on credible research, little progress will be made in dealing with the CPP, because it will not be possible to generate accurate descriptions and analyses upon which to base recommendations and solutions.

THE RESPONSE NEEDED TO ASSESS STUDENT LEARNING OUTCOMES

Due to the size and importance of the issue, we need all hands on deck. Advocates of portfolios are creating important best practice models for faculty in the classroom to emulate. There are many other efforts to provide ways for individual faculty members to directly assess the results of their teaching and learning. We need also to recognize the contributions of cognitive scientists, who now have much to teach us about how the brain learns (Miller, 2003). Education technologists will be needed in order to scale up the ideas which are created (see, for example, the Open Education Resource (OER) movement). Measurement scientists are also needed because they insist upon measurement instruments with demonstrable validity, meaning they measure what the instrument is intended to measure, are given to students under the same conditions and are based on reliable scoring rubrics. Due to the high stakes involved, direct measures of student learning that meet the highest standards of reliability and validity will be required in order to provide the systematic evidence needed to make the many decisions that will affect resource distribution and the improvement of the quality of student learning over the coming decades.

THE CENTRAL ROLE OF THE FACULTY

The discussion about assessment and accountability tends to focus on policy issues or the reliability and validity of assessment instruments. These are, of course, important issues. However, a discussion of the relevance of the assessment instrument to teaching and learning is either completely absent or approached as an afterthought. The threshold question is the instrument's relevance to the faculty in the classroom. In addition, the relevance of the assessment instrument to the faculty in the classroom should take precedence over its technical dimensions and larger policy debates over whether or how assessment or accountability should occur. In other words, the assessment instrument must be known to be reliable and valid, but this should only be a necessary, and not sufficient, condition for its adoption by a college or university. Do the faculty find the assessment instrument useful? That should be the most important question.

The faculty should be the focus of assessment, because individual instructors are at the center of matters relating to teaching, learning and the curriculum. The

implication of this point is that faculty buy-in is critical to the future of assessment and accountability in the academy. Until it is clear that testing organizations have developed assessment instruments that are accepted by the faculty as valuable aids to their instruction, it is unlikely that we will move forward in the policy debates on assessment and accountability in higher education. Thus, our focus should be on encouraging the faculty to use assessment instruments that are in line with their teaching and learning goals.

If the faculty buy in to using assessment instruments as central tools to monitor and improve teaching and learning, this will increase the probability of positive developments on other fronts, such as accountability and the use of assessment-based evidence for internal governance and diagnostic purposes, because it will be possible to base these other activities on assessments which faculty members perceive to be authentic.[4] However, this process must begin with the faculty recognizing the inherent value of assessment to their own work as teachers. This will occur only if the assessment tools themselves are proven to be effective for the cycle of teaching, learning and assessing for continuous improvement. Of course, additional significant changes are needed in order to make this equation work. The faculty must have incentives to encourage them to focus on student learning rather than research alone, and students need support and encouragement to learn.

THE RATIONALE FOR PERFORMANCE ASSESSMENT[5]

The current assessment regime, dominated by multiple-choice tests, is no longer sufficient in the knowledge economy. For a century, multiple-choice tests have been the principal assessment method in education. This probably made sense in the industrial era of development, as this method mirrors the focus on the mastery of content demonstrated by students' ability to recall facts. Today, we live in an economy dominated by information and services rather than physical goods. In the knowledge economy, it is more important to be able to access, structure and use information than merely recall facts. This places a premium on the ability of students to reason, assess the relevance of information and make arguments; in short, to think critically. This effort to focus on critical thinking skills is being implemented in classrooms across the country, in which faculty are arming their students to navigate a constantly changing world defined by an ever-increasing volume of information. The manner in which we assess students must reflect these interests. Multiple-choice tests may present examples of correlations and causation and ask students to identify whether each is correctly or incorrectly applied. However, responding passively to such choices is very different from asking students in performance assessments to actively critique a case study that presents an argument about data in which the concepts of correlation and causation are misused. It is also important to underline the requirement in the knowledge economy for citizens to actively shape the information at their disposal, rather than simply to respond passively to choices put before them.

Assessment must therefore catch up with an emerging reform agenda in higher education, resulting from our new understanding of student learning. At the most

basic level, this involves understanding that the meaning of knowledge itself is undergoing a significant shift. New theories from the field of cognitive science stress the importance of improving students' ability to structure learning experiences that help them to use what they have learned in new settings. Simon argues that the meaning of "knowing" has changed from being able to recall information to being able to find and use it (Simon, 1996, p. 43). Under these conditions, the proponents of the new learning theories argue that active learning is critical, because students must learn to recognize when they understand a subject and when they need more information (Pellegrino, Cudowsky, & Glaser, 2001). The implications for higher education are profound. If we consider the assumptions which structured higher education in the industrial era, the lecture format was the norm, with students seen as passive receptacles receiving the content provided by lecturers. The role of higher education was to transmit knowledge. Faculty and administrators were comfortable with these assumptions, because even though it was understood that knowledge was progressing in multiple fields, most shared the view that there was a stable, enduring stock of knowledge that graduating seniors should know. Under these circumstances, content was emphasized and multiple-choice tests were the preferred assessment tool. However, although I am arguing for a greater emphasis on performance assessment in the 21st century knowledge economy, this does not mean that multiple-choice tests are inferior. Good performance assessments, just like good multiple-choice tests, must be constructed on clear definitions of what students should know and be able to perform. They then must measure that domain. There are no silver bullets in assessment instruments, just as there are no universal solutions to the improvement of student learning in higher education. Now that the Internet and computer-assisted scoring have made it feasible to scale up performance assessment, it is time to explore the practical possibilities of this testing paradigm more fully.

Over the past two decades, it has become clear that a new vision of undergraduate education is developing in response to the changing definition of knowledge. It is comprised of three parts:

- A shift from the lecture format to a student-centered approach that emphasizes analytical writing. Faculty are much more interested in active student participation in the learning process, and students appear to be equally interested in doing so. Although evidence is still in the formative stage, it appears that colleges that emphasize analytical writing produce students who do well in assessments that benchmark higher-order skills;
- There has been a change in emphasis from the pre-existing focus in curricula and texts on content to case- and problem-based materials that ask students to apply what they know to new situations. This is reflected in curriculum reform and is also resulting in textbook publishers producing solely content-filled volumes. The graduate business school emphasis on the case approach to learning may be an early example of this strategy;

– There has also been a change in assessment from multiple-choice and short answer formats to open-ended essays that are better aligned with the first two parts of the reform.

Performance assessments (constructed responses that require students to demonstrate their ability to perform tasks) appear to be better aligned with the focus of this education reform movement. Performance assessments are congruent with recent theories of learning and knowledge that focus on applying what one knows to new situations, typically including the ability to think critically, solve problems and write effectively. The Internet and computer-assisted scoring have enabled performance assessments to be administered, scored, analyzed and reported to students and their instructors in increasingly cost-effective and accurate ways. Faculty do not like multiple-choice tests, but perceive performance assessments as being authentic. This means that performance assessment, uniquely, can be used in both the standardized and formative assessment space. This is encouraging, because the only way to create a sustainable assessment system in postsecondary education is to create a more systematic, continuous system of teaching and learning improvement based on assessment instruments that all parties (the faculty, administrators, governmental authorities) can use in order to achieve their objectives.[6]

CONCLUSION

In today's knowledge economy, it is more important to be able to access, structure and use information than merely to accumulate facts. Performance assessments are appropriate for benchmarking and stimulating the development of the necessary critical thinking skills. As there are no "correct" answers to performance tasks, they are worth teaching to. Moreover, in the case of the Collegiate Learning Assessment (CLA, a performance assessment used in the U.S., 2010), because it is not focused on discipline- and content-based knowledge, its use does not narrow the curriculum.

Recommendation One: Consider adopting performance assessments for higher education, because they can be used in both formative and standardized assessment applications. They are unique in this sense. This does not mean that performance assessment of critical thinking skills should be the sole approach to assessment in higher education. Performance assessment should be combined with multiple other forms of assessment.

Recommendation Two: Technology is a key enabler for possible advances in assessment. Performance assessment has been around for a long time, but the Internet unlocked its large-scale potential because complex tasks can be placed on an Internet website and administered, scored, analyzed and reported back to the student and their college with fewer errors and in a cost-effective way. The advent of computer-assisted scoring has created the opportunity for on-demand testing in the classroom, because the rapid turnaround of test results is now possible (see the description of computer-assisted scoring in Appendix 1). In addition, consider the

potential use of sharing best practices for teaching, learning and assessing on the new OER Internet platforms now being developed.[7]

Recommendation Three: Consider collaboration between cognitive scientists, educational technologists, formative assessment experts and measurement scientists in order to create more unified approaches to teaching, learning and assessment.[8] A much greater effort than is currently being made anywhere in higher education will be required to improve teaching and learning, which is the fundamental goal of assessment.

APPENDIX[9]

Computer-assisted Scoring

Computer-assisted scoring employs computer models to score open-ended assessment responses. These models are created from the scores assigned by trained and calibrated graders. The computer uses these grades to operationally infer the rubric and scoring scale. The computer-assisted scoring process does not deal with the content of a complex performance assessment; instead, it is dependent on the scoring of human experts. Several hundred expert-scored student responses are used to train the computer-assisted scoring engine. The computer-assisted scoring engine "learns" the features and characteristics of the scoring rubric and each score from the expert-scored responses, which it uses to evaluate student responses. The engine relies on the collective wisdom of the expert scorers, reflected in the scores they assigned to a representative set of actual student responses. Much like the training of human scorers, the engine "learns" how to score student responses through repeated exposure to expert-scored examples.

Once approximately 500 student responses have been double scored by experts and the quality of the task has been verified, the results of the experts' scoring are used to generate the computer-scoring model. The computer-assisted scoring engine is presented with the complete text of approximately 500 student responses, along with the experts' scores. The engine examines the content and structure of each response and associates the information with the score assigned in order to create a model of what each point "looks like."

In sum, the computer-assisted scoring approach has been shown to be as accurate as expert scorers, and in some cases, more accurate than expert scorers. The use of computer-assisted scoring allows testing organizations to offer accurate, fast and cost-effective value-added assessment services to institutions of higher education.[10]

NOTES

[1] This chapter is based on examples from the American postsecondary education sector, set in the American social, economic and political context. However, most political leaders now understand the vital need to ensure and enhance the skills of their workforce and the human capital of their citizens. Many countries are also facing major CPPs as a function of immigration, population

growth, a lack of resources or the absence of postsecondary education institutions of a sufficient size and quality to improve the human capital in their country. See Benjamin (2012) for a more extensive version of the points made in this chapter.

[2] Moreover, a recent Social Science Research Council study found that student learning in colleges was not as high as previously thought (Arum & Rotska, 2011).

[3] For research, the other principal public good produced by colleges and universities, there are a number of empirical-based metrics that permit serious examination of the factors that produce stronger research programs.

[4] For example, faculty may be able to reclaim governance over the undergraduate curriculum (Benjamin, 2007).

[5] Performance assessments are designed to evaluate the ability of students to apply what they know to new situations. For an example of a performance assessment, see http://starttest.com/7.0.0.1/programs/clacross/Practice%20Test%20Page.htm, which displays a collegiate learning performance task used in many colleges in the U.S.

[6] The performance assessment paradigm has recently been embraced by the U.S. Department of Education and the Gates Foundation. Due to a number of grants, performance assessments are being developed for college readiness tests (see website of the SMARTER Balanced Assessment Consortium http://www.k12.wa.us/smarter/). One performance assessment, the CLA, with which the author is associated, is used extensively by U.S. higher education institutions. For a description of the assessment instrument see the paper of R. Shavelson in this volume (Part 1). See also *The Architecture of the CLA* (Benjamin et al., 2009).

[7] See the description of the OER on the Hewlett Foundation website (http://www.hewlett.org/programs/education-program/open-educational-resources). The purpose of the OER is to place intellectual property such as textbooks, curriculum materials and disclosed assessment instruments on the Internet as open resources which are available to use for free by faculty, teachers or any interested parties.

[8] A best practice example of such an approach that combines cognitive science, subject matter specialists, assessment specialists and education technologists is the Open Learning Initiative at Carnegie Mellon University (see http://oli.web.cmu.edu/openlearning/).

[9] This description is an excerpt from an essay on computer-assisted scoring by S. Elliot (2011).

[10] I would like to thank the conference organizers Sigrid Blömeke and Olga Zlatkin-Troitschanskaia for organizing such a stimulating conference. I would also like to thank the two referees for their helpful comments. Finally, I would also like to thank Christiane Kuhn for her gracious hospitality before, during and after the Berlin conference.

REFERENCES

Arum, R., & Roska, J. (2011). *Academy adrift: Limited learning on our campuses*. Chicago: University of Chicago Press.

Benjamin, R. W. (2007). Recreating the faculty role in governance in research universities. In J. Burke (Ed.), *Fixing the fragmented university* (pp. 70–98). Bolton: Anker Publishing.

Benjamin, R. W. (2012). The new limits of education policy: Avoiding a tragedy of the commons. London: Edward Elgar.

Benjamin, R., Chun, M., Kugelmass, H., Nemeth, A., & Shavelson, R. (2009) The Architecture of the CLA. New York: CAE. Retrieved from collegiatelearningassessment.org/files.

Carnegie Mellon University. *Open Learning Initiative*. Retrieved from http://oli.web.cmu.edu/openlearning/.

Council for Aid to Education. (2010). *CLA performance task*. Retrieved from http://starttest.com/7.0.0.1/programs/clacross/Practice%20Test%20Page.htm.

Elliot, S. (2011). *Computer-assisted scoring of performance tasks for the CLA and CWRA*. Retrieved from http://www.collegiatelearningassessment.org/files/ComputerAssistedScoringofCLA.pdf

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243–1248.

Griswold, J. (2006). *Higher Education Price Index (HEPI) report.* Retrieved from http://www.commonfund.org (Common Fund Institute).

Miller, G. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences, 7*(3), 141–144.

Pellegrino, J., Chudowsky, N., & Glaser, R.. (2001). *Knowing what students know: The science and design of educational assessment*. Washington DC: The National Academy Press.

Simon, H. (1996). *The sciences of the artificial*. Boston: MIT Press.

The William and Flora Hewlett Foundation. *Open educational resources*. Program of Grant Making. Retrieved from http://www.hewlett.org/programs/education-program/open-educational-resources.

Roger Benjamin
Council for Aid to Education (CAE), USA

RAFAEL VIDAL URIBE

MEASUREMENT OF LEARNING OUTCOMES IN HIGHER EDUCATION

The Case of Ceneval in Mexico

INTRODUCTION

In mid-2010, the National Center for the Evaluation of Higher Education in Mexico (Ceneval) received an invitation from the German Federal Ministry of Education and Research (through the Johannes Gutenberg University at Mainz and Humboldt University) to deliver a keynote lecture about the operation of the Mexican Higher Education Exit Assessments Tests (EGEL) at a conference entitled "Modeling and Measurement of Competencies in Higher Education" that was held in Berlin in early 2011. The following lines present the highlights of that presentation.

CENEVAL

Ceneval is a non-profit institution which was founded in 1994 by the National Association of Universities and Higher Education Institutions (ANUIES). Ceneval's mission is the design and administration of tests and assessments for academic purposes. Ceneval's tests target mainly students from high schools and universities. Since its inception, Ceneval has delivered a number of assessment reports, results and benchmarks to academic institutions, education officials and students. All reports produced by Ceneval within this project focus on the improvement of education. Ceneval does not perform institutional or program evaluations, nor peer-review evaluations, but serves hundreds of public and private academic institutions and several governmental agencies and administers its instruments all over the country, and in the U.S.A. and Ecuador.

Ceneval operates more than 240 different instruments, which are continuously maintained by more than 80 expert groups. Thousands of teachers and experts from all over the country are in charge of writing and reviewing the items. More than 3 million people take Ceneval tests every year. Around 1,800 testing events relating to Ceneval tests take place every year, and around 40,000 people take online tests using proprietary online software. More than 500 full-time staff work at Ceneval.

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 137–146.

Ceneval's instruments can be grouped into four broad categories:

- Admissions or entrance tests (EXANI);
- EGEL;
- Generic knowledge and skills instruments (EXDIAL);
- Instruments developed for other agencies (on demand).

In the following, we shall focus on categories 2 and 3.

HIGHER EDUCATION EXIT ASSESSMENTS (EGEL)

EGEL are tests for the evaluation of higher education (HE) learning outcomes. In the Spanish language, these instruments are called *Examen General Para el Egreso de la Licenciatura*, hence the acronym "EGEL".

EGEL tests are standardized instruments which target students who are about to finish their higher education (a bachelor's degree in North America or a Licenciatura in Mexico). These higher education tests have been developed for 33 different subjects or university programs. Table 1 shows the 33 EGEL subjects.

Table 1. Complete list of EGEL tests in operation (as of April 2011)

Life Sciences and Behavioral sciences (11)		*Social Sciences and Humanities (10)*		*Engineering and Technology (12)*	
1	Agricultural Sciences	12	Accounting Education	22	Chemical Engineering
2	Biology	13	Business Administration	23	Civil Engineering
3	Chemistry	14	Communication	24	Computer Engineering
4	Clinical Chemistry	15	Economics	25	Computer Science
5	Dentistry	16	International Business Administration	26	Electrical Engineering
6	Medicine	17	Law	27	Electronic Engineering
7	Nursing	18	Marketing	28	Industrial Engineering
8	Nutrition	19	Pedagogy/ Education	29	Information Systems
9	Pharmacy	20	Social Work	30	Mechanical Engineering
10	Psychology	21	Tourism	31	Mechanical Electrical Engineering
11	Veterinary Medicine			32	Mechatronics Engineering
				33	Software Engineering

As previously mentioned, there are only 33 EGEL subjects, and although higher education institutions (HEIs) use to have many more different programs (some have up to 50 or 60 bachelor's degree programs or Licenciaturas), with EGEL tests a HEI could cover (assess) a very large part of the graduation pool. This is because EGEL tests target the higher education programs with the highest enrolment rate, and also because there are programs that, under different denominations, actually teach almost the same domains. For instance, in the case of the Business Administration EGEL Test (EGEL-ADMON), we found in Mexico up to 94 different higher education programs with different names that are closely related in purpose and subject; therefore, students from those 94 programs could all be evaluated with the EGEL-ADMON Test.

The 33 EGEL tests cover up to 75% of the country's graduation pool for one year. Up to 104,806 students from 490 HEIs in Mexico took EGEL tests during 2010. The EGEL Tests Program has been in operation since the creation of Ceneval in 1994. Over the past 17 years (1994–2010), 739,733 higher education students have taken EGEL tests.

The Nature of EGEL

The purpose of EGEL tests is to identify whether or not graduating students have the minimum knowledge, skills and competencies to go into professional practice. An EGEL test is a test of minimums (minimum knowledge required). It is the "floor"; therefore, the test is not an in-depth evaluation of each individual student, but instead provides a rough idea of what a student knows and is able to do in order to start his or her professional life.

EGEL tests are made to assess only the basics, and not everything which is taught in schools. EGEL tests offer students and HEIs an indicator of students' knowledge and skills.

EGEL tests are not a requirement for students, unless the HEI decides to make it so. However, on many occasions, EGEL tests have been administered to everyone who is graduating from a particular program. When every student in the graduation pool takes an EGEL test, the HEI gains a good indication of the performance of the program in question within the national context. As EGEL tests are not mandatory according to the law, HEIs are free to ask for the administration of these tests to their students. In general, HEIs pay for the administration of EGEL tests.

In a Nutshell, What is The Purpose of EGEL Tests?

- EGEL tests allow students to know whether or not they have reached a certain national standard set by a group of experts. If they achieve a satisfactory or an outstanding result, they also receive a Ceneval Diploma that can be used for job-seeking purposes.
- EGEL tests help principals, deans, program administrators and advisors to benchmark their higher education programs.

– For assessment and accreditation agencies, EGEL tests provide ways to improve the delivery of their recommendations and accreditations.

EGEL Characteristics and Development Processes

All EGEL tests are multiple-choice, nation-wide, domain-specific and not mandatory. As diagnostic tests, EGEL tests are all criterion-referenced assessments (Shrock & Coscarelli, 2007); this means that a student's performance is judged or measured against a standard or criterion. Moreover, this also means that the overall results are expressed not on a numerical scale, but on a categorical scale (outstanding, satisfactory or not yet satisfactory). The reports will be discussed in greater depth later in this article.

All items on the tests are oriented to practical situations. Items ask students to pronounce over situations that a young or fresh professional is likely to encounter.

In guiding the process of defining the construct domain, producing the test specification, building the item bank, creating and administering the actual tests, scoring the results and delivering reports, Ceneval follows the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Each EGEL test is always supervised by a technical expert group (TEG) which is made up of university professors, industry experts and professionals from disciplines which are associated with the test subject. TEG members must reach an agreement on the minimum level that students must attain in order to enter the job market; consequently, nothing below that minimum should be acceptable. The TEG must endorse all of the activities and decisions that affect the test, and also in order to permanently enhance the test validity and reliability TEG must review statistical reports and evidence collected after every test administration.

Like many well-known tests (ACT, 2007; General Education Development, 2002, and so on), each EGEL test has a Technical Manual; within it, there is a complete description of the purpose of the test, its objectives, target population, the conditions required in order to take the test, the psychometric techniques needed to verify the quality of items and testing forms, algorithms used to obtain test scores, and any information needed to interpret the individual and institutional reports.

Trained specialists write EGEL items. Norms and specifications for writing the items are set by Ceneval and by every EGEL TEG (Osterlind, 1998; Downing, 2006; Haladyna, Downing, & Rodriguez, 2002). All items are developed in a proprietary bank platform (*e*-BRAE).

The qualitative review process is time-consuming; however, no EGEL items are ready for an operative test until they have been field-tested with representative student samples. Each item must meet the minimum quality criteria in terms of the difficulty and discrimination indices of classical test theory. In addition, tests with more than 300 examinees are analyzed using the two-parameter item response theory model (Crocker & Algina, 1986; DeVellis, 2010; Embreston & Reise, 2000; Raykov & Marcoulides, 2010).

As soon as an EGEL test produces data, Cronbach's alpha is calculated as a reliability measure (Nunnally & Bernstein, 1994). In 2010, 16 out of 33 tests reached alpha coefficients above 0.90, and the rest were above 0.80, indicating that the EGEL subjects have good levels of internal consistency reliability.

EGEL Administration and Scoring

Nowadays, almost all EGEL tests are administrated using paper and pencil forms, with only a few delivered online; however, all tests will be available in full online in 2012.

Ceneval organizes four national testing events every year in more than 60 sites all over the country. The printed test forms are administrated in two four-hour sessions in the same day (morning and evening). EGEL tests are usually between 200 and 250 items in length.

Proctors trained by Ceneval administer the tests. In order to control variables that could affect the test results, instructions for proctoring the tests are standardized across the administration events (McCallin, 2006).

As previously mentioned, all EGEL tests are criterion-referenced, meaning that the students' attainment is compared against a criterion which has been previously defined by the TEG. EGEL tests may have three, four or five subscales, and it is the decision of the TEG which combination of results from the subscales produces the overall result. For EGEL tests, there are only three possible overall results: *Outstandin*g, *Satisfactory* and *Not yet satisfactory*. Student and institutional reports also show the students' performance on each test subscale. Test subscale scores are expressed in categorical and numerical terms. The numerical scale is called the Ceneval Index and has limits of 700 and 1,300 (600 points). Table 2 shows the equivalence of the Ceneval Index to performance levels for the subscale reports.

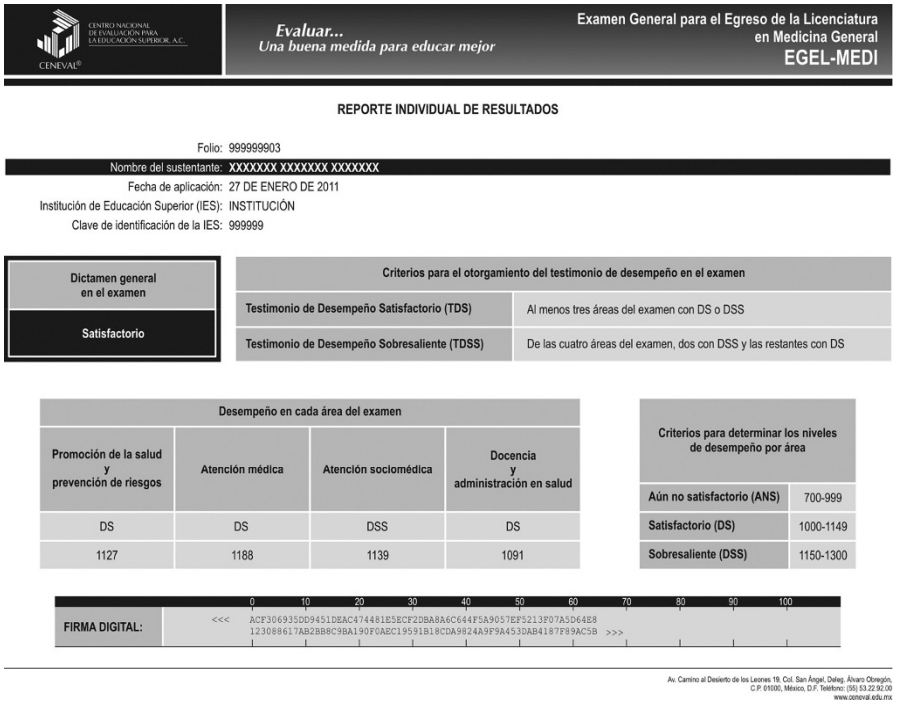
Table 2. Equivalence of Ceneval Index to performance levels

Ceneval Index range	*Performance level*
700–999	Not yet satisfactory
1,000–1,149	Satisfactory
1,150–1,300	Outstanding

EGEL Reports

After each test administration event, Ceneval prepares two types of report: individual (student) and institutional reports.

Individual report. Individual reports show the student's overall results and his or her results for each subscale of the test. The reverse of the individual report shows a description of the student's performance levels on each subscale. Figure 1 shows an individual report for the Medicine EGEL Test (front side). In addition, all students that achieve a satisfactory or outstanding overall result receive a Ceneval Diploma called a *Testimonio*.


CENEVAL®

Evaluar...
Una buena medida para educar mejor

Examen General para el Egreso de la Licenciatura en Medicina General
EGEL-MEDI

REPORTE INDIVIDUAL DE RESULTADOS

Folio: 999999903
Nombre del sustentante: XXXXXXX XXXXXXX XXXXXXX
Fecha de aplicación: 27 DE ENERO DE 2011
Institución de Educación Superior (IES): INSTITUCIÓN
Clave de identificación de la IES: 999999

Dictamen general en el examen
Satisfactorio

Criterios para el otorgamiento del testimonio de desempeño en el examen	
Testimonio de Desempeño Satisfactorio (TDS)	Al menos tres áreas del examen con DS o DSS
Testimonio de Desempeño Sobresaliente (TDSS)	De las cuatro áreas del examen, dos con DSS y las restantes con DS

Desempeño en cada área del examen			
Promoción de la salud y prevención de riesgos	Atención médica	Atención sociomédica	Docencia y administración en salud
DS	DS	DSS	DS
1127	1188	1139	1091

Criterios para determinar los niveles de desempeño por área	
Aún no satisfactorio (ANS)	700-999
Satisfactorio (DS)	1000-1149
Sobresaliente (DSS)	1150-1300

0 10 20 30 40 50 60 70 80 90 100

FIRMA DIGITAL: <<< ACF306935DD9451DEAC474481E5ECF2DBA8A6C644F5A9057EF5213F07A5D64E8
123088617AB2BB8C9BA190F0AEC19591B18CDA9824A9F9A453DAB4187F89AC5B >>>

Av. Camino al Desierto de los Leones 19, Col. San Ángel, Deleg. Álvaro Obregón, C.P. 01000, México, D.F. Teléfono: (55) 53.22.92.00 www.ceneval.edu.mx

Figure 1. EGEL test individual report.

Institutional report. Institutional reports are very helpful tools for HEIs. Ceneval provides HEIs with reports of the performance of their students in the EGEL tests. Of course, information about the results of an individual student is sent only to her or his HEI. Every institutional report is a performance résumé for all of the students of one HEI. Institutional reports are sent to the program head or academic dean a few days after a testing event. No-one else receives the institutional report, as they are confidential.

Comparative institutional report. In addition, once a year, a complete comparative institutional report (CIR) with the aggregate results of all of the students who have taken EGEL tests during the past calendar year, is sent to every HEI involved in EGEL. This CIR shows the results of EGEL tests for all students and HEIs which have participated in the EGEL tests during that calendar year (more than 104,000 students and more than 490 HEIs went into the aggregated CIR for 2010); however, the names of all of the institutions are coded. The idea behind these CIRs is to provide a rough idea of the performance of a program in a specific HEI in relation to similar programs in other institutions.

Figure 2 shows an excerpt from a CIR (for medicine only).

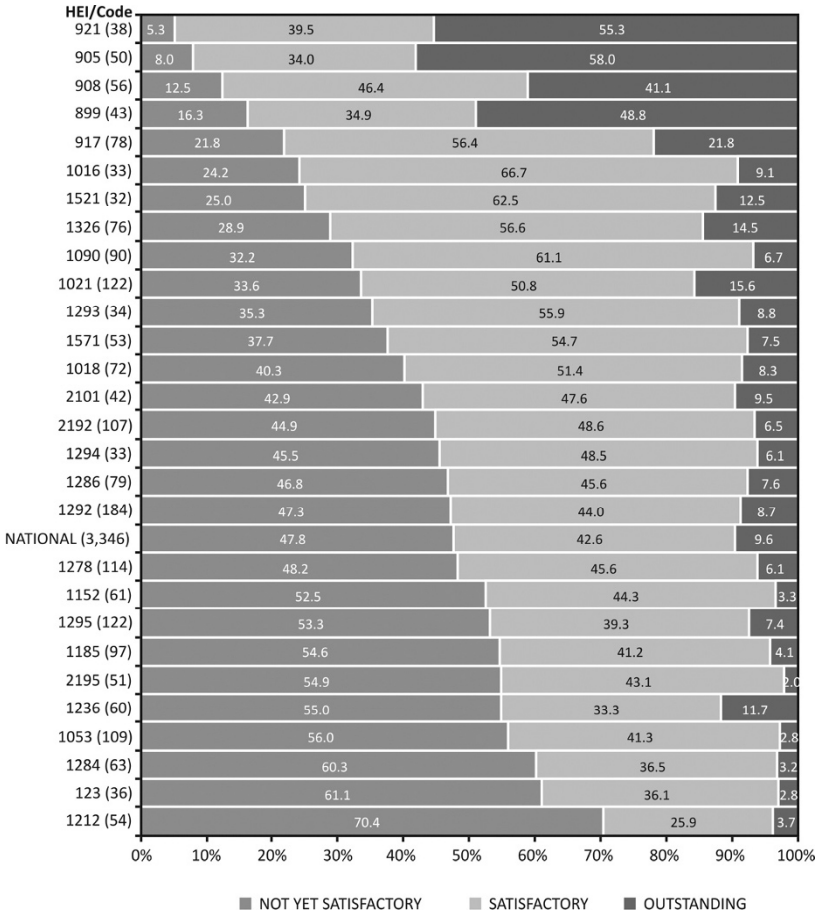


Figure 2. EGEL CIR.

GENERIC KNOWLEDGE AND SKILLS INSTRUMENTS

Ceneval also provides HEIs with a set of four generic knowledge and skills instruments; these testing instruments do not target students from a specific higher education program, but the students of several or all higher education programs. These tests are not necessarily taken at the end of the undergraduate program. These generic instruments are as follows:

Basic Science for Engineering Programs Test (EXIL-CBI)

This test was developed for students who are finishing the basic science subjects in the engineering faculties (it is usually taken at the end of the second academic year). Table 3 shows the content of the EXIL-CBI Test.

Table 3. Scales and subscales of the EXIL-CBI Test

Content	
1. Mathematics	1. Algebra 2. Calculus 3. Differential equations 4. Probability and statistics
2. Physics	1. Mechanics 2. Thermodynamics 3. Electromagnetism
3. General chemistry	1. Pure substances and mixtures 2. Chemical reactions

Statistics Test (EXTRA-ES)

The EXTRA-ES Test was developed in order to reach all students with statistical topics in their higher education programs. As not all students taking statistical courses take statistics at the same level, the EXTRA-ES test is organized with a common core recommended for all statistics students and three optional modules which are included at the discretion of the program director or student advisor. Table 4 shows the content of the EXTRA-ES Test.

Table 4. Scales and subscales of the EXTRA-ES Test

Common core
Fundamentals of statistical thinking and data structure generation
Description, organization and interpretation of data
Notions of inference

Optional modules
Inferential methods and modeling
Sampling
Experimental statistics

Communication and Critical Thinking Test (ECCyPEC)

Communication and critical thinking skills are necessary for every student finishing an undergraduate program. Ceneval has developed a test that aims to measure these skills in a broad and general way. This instrument targets any undergraduate student towards the end of his or her years in higher education. Table 5 shows the content (scales and subscales) of the ECCyPEC.

Table 5. Scales and subscales of the ECCyPEC Test

Content
Reading comprehension
Knowledge of written expression
Critical thinking

Written Expression Test (EEE-II)

The EEE-II is an essay test that aims to measure actual writing skills. It is not an objective test. Higher education students are asked to write an essay of two to three pages on a dilemmatic topic. Each EEE-II essay is scored by at least two expert Spanish teachers with the help of a rubric. Table 6 shows the content of the *EEE-II Test.*

Table 6. Content of the EEE-II Test

Content
Conventions of language
Syntactic knowledge
Lexical variety
Thematic progression of the text
Global consistency
Planned speech
Information sources
Creativity

For more detailed information (in Spanish) about EGEL and the generic skills instruments from Ceneval, please go to http://www.ceneval.mx/.

REFERENCES

ACT (2007). *The ACT technical manual*. Retrieved from http://www.act.org/aap/pdf/ ACT_Technical_ Manual.pdf.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth.

DeVellis, R. F. (2010). *Scale development: Theory and applications* (3rd ed.). Applied social research methods series. Thousand Oaks, California: SAGE Publications, Inc.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Embreston, S., & Reise, S. P. (2000). *Item response theory for psychologists.* Multivariate application series. Mahwah, New Jersey: Lawrence Erlbaum Associates.

General Educational Development. (2002). *Technical Manual: 2002 series GED tests*. Washington, DC: American Council on Education. Retrieved from http://www.acenet.edu/Content/ NavigationMenu/ged/pubs/TechnicalManual_2002SeriesGEDTests.pdf.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 625–652). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Evaluation in education and human services. Boston, MA: Kluwer Academic Publishers.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York: Taylor and Francis Group, LLC.

Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-referenced test development: Technical and legal guidelines for corporate training and certification*. (3rd ed.). San Francisco, CA: John Wiley and Sons.

Rafael Vidal Uribe
National Center for the Evaluation of Higher Education (Ceneval), Mexico

HILDEGARD SCHAEPER

THE GERMAN NATIONAL EDUCATIONAL PANEL STUDY (NEPS)

Assessing Competencies Over the Life Course and in Higher Education

INTRODUCTION

The German National Educational Panel Study (NEPS) is an exceptional and unique research endeavor which aims to gain new insights into the acquisition of competencies across the entire life span, to describe crucial educational transitions, to study educational careers, to identify the determinants of competence development and educational decisions and to analyze the impact of education and competencies on the life course. This article gives a brief overview of the conception and structure of the NEPS. It then describes in more detail the general approach to modeling and measuring competencies used by the NEPS, as well as the way of addressing the issue of subject-specific competencies in higher education.

RESEARCH QUESTIONS, RESEARCH DESIGN AND ORGANIZATION OF THE NEPS

Overview

The NEPS, funded by the Federal Ministry of Education and Research, is an instrument for studying education over the life course and addresses a wide range of questions, including:

- How do competencies develop over the life course?
- What are the central factors in the process of competence acquisition and educational decision-making? What role do educational institutions, non-formal and informal learning environments play? How important are social characteristics, the cultural context and economic living conditions?
- What does the relationship between competencies and educational credentials look like? To what extent do certificates (grades) reflect levels of competence?
- Which competencies are decisive for educational and labor market success? To what extent do labor market outcomes depend on acquired competencies, credentials, social origins, social and cultural capital and personality traits?
- What are returns to education and competencies in terms of income, occupational career, subjective well-being, social, political and cultural participation and health?

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 147–158.

– Which factors favor participation in continuing education in later life? What conditions are unfavorable for lifelong learning?

The preeminent theoretical orientation of the NEPS is the life course perspective, meaning that the study aims to investigate the process of education, learning and competence development over the entire life span. In order to provide relevant data quickly, the NEPS uses a specific methodological approach (see Figure 1). We decided not to observe a single birth cohort over several decades, but to start with different cohorts at the same time and to follow them over a longer period of time. The cohorts are either age-based (newborns, adults) or defined by a specific point in their educational career (e.g., entry into higher education). Each of these cohorts focuses on one or two stages of education. In accordance with the structure of the German education system, the NEPS distinguishes between eight educational stages, e.g., stage 1 "From birth to early child care"; stage 5 "From upper secondary school to higher education, vocational training and the labor market"; and stage 7 "From higher education to the labor market" (see Figure 2).

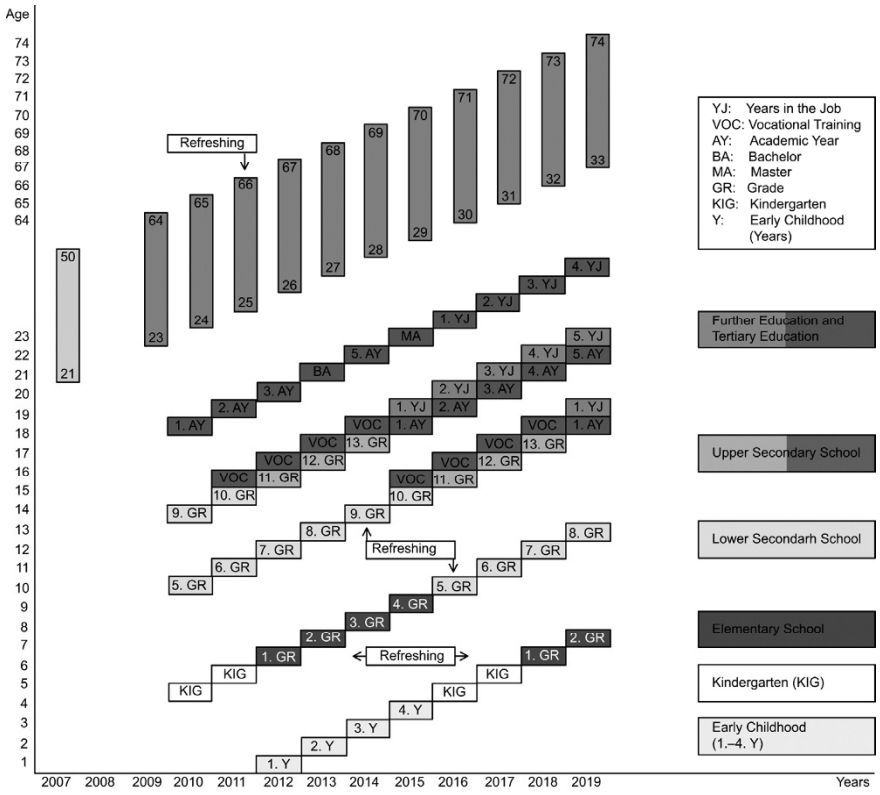


Figure 1. The multi-cohort sequence design of the NEPS.

As the life course perspective is central to the NEPS, a coherent conceptual framework is necessary that covers the entire life course and integrates the study of different stages of education and cohorts. This integration is ensured by a theoretical orientation towards five major dimensions, which are called "pillars" (see Figure 2): Pillar 1 is concerned with competencies and has the task of modeling competence development over the life span and constructing corresponding tests (for details, see Weinert et al., 2011). Pillar 2 focuses on the conceptualization of different learning environments and the operationalization of central contextual characteristics that are expected to have an impact on competence acquisition and educational decisions (for details, see Bäumer et al., 2011). Pillar 3 addresses social and gender inequalities and the question of how inequalities are reproduced and transformed by educational decisions (for details, see Stocké et al., 2011). As the individual migration history and an individuals' ethnic or cultural origin have an effect on competence development and educational decisions that goes beyond the mechanism of social inequality, pillar 4 addresses the acquisition of education across the life course of migrants and their descendants (for details, see Kristen et al., 2011). The central issue of pillar 5 is returns to education, both in monetary and non-monetary terms (e.g., income, risk of unemployment, subjective well-being, social, cultural and political participation, health; for details, see Gross et al., 2011).

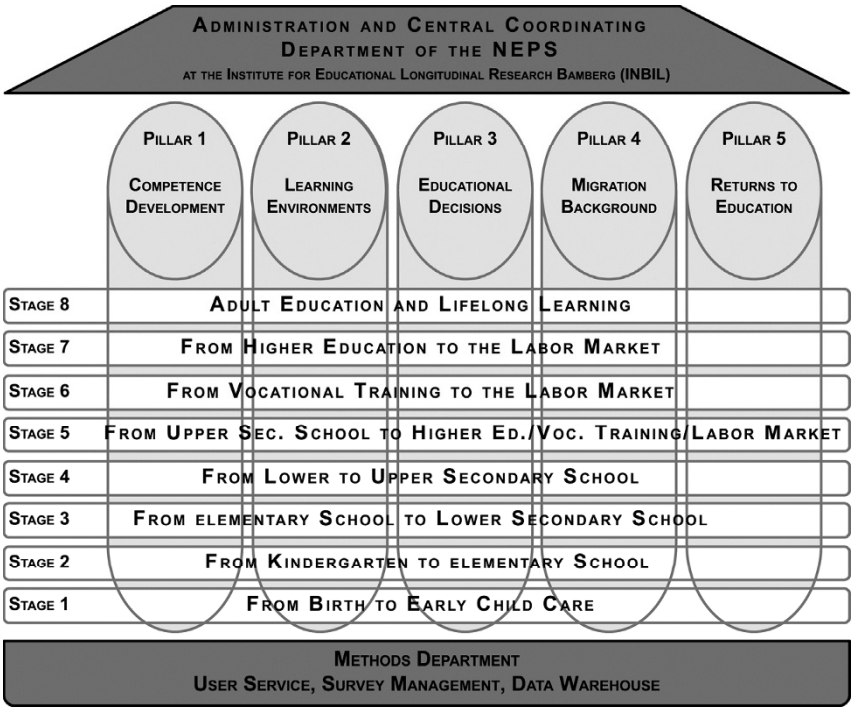


Figure 2. The five basic pillars and eight stages of the NEPS.

It goes without saying that such an ambitious and interdisciplinary project as the NEPS which approaches the research questions from different theoretical and disciplinary angles, cannot be carried out by a small research group. Therefore, experts and expert groups from all over Germany collaborated in order to form an effective network of excellence. This consortium links researchers from various disciplines (e.g., psychology, educational science, sociology, economics, demography, migration studies, statistics and survey methods) and major research institutions and is headed by the principal investigator Hans-Peter Blossfeld. The central coordination and administration facility of the NEPS is located at the Institute for Longitudinal Educational Research at the University of Bamberg (INBIL). The HIS-Institute for Research on Higher Education, based in Hanover, is responsible for stage 7: "From higher education to the labor market".

More detailed information on the objectives, design and structure of the NEPS is given by Blossfeld, von Maurice, & Schneider (2011).

The Sub-study: "From Higher Education to the Labor Market"

Within the conceptual framework of the NEPS, the longitudinal sub-study entitled "From higher education to the labor market" (stage 7) follows a cohort of approximately 16,500 randomly selected new entrants to higher education through their student days and beyond (for further information on this sub-study, see Aschinger et al., 2011). Of course, the key research areas for the higher education stage center on the overall questions of the NEPS, but focus to an extent on specific aspects. With regard to educational decisions, for example, stage 7 of the NEPS pays special attention to dropping out, entering a master's program, starting a dissertation and entering employment. As regards competencies, we will not only examine the domains that are assessed in all stages of education covered by the NEPS and, therefore, constitute a common core of competence assessment (see below), but we will also collect data on stage-specific competencies in particular fields of study ("subject-specific competencies of higher education students/graduates").

The higher education stage of the NEPS also pays special attention to particular groups of students who have previously been neglected in higher education research or are of special interest to education policy. For example, teacher training students are oversampled in order to provide detailed large-scale data on what is considered to be a key profession for the quality of school education. In addition, we tried to include the entire population of first-year students without a school-leaving certificate qualifying them for higher education (so-called "nontraditional students"; Schuetze & Wolter, 2003).

Data is being collected using several modes, e.g., self-administered questionnaires, computer-assisted telephone interviewing, online surveys and group tests in classroom settings. As regards the frequency and timing of the panel waves, up to three (but usually two) short surveys or tests will take place every year. Data collection started in autumn 2010.

MODELING AND MEASURING COMPETENCIES WITHIN THE NEPS

The General Approach

According to a well-established definition which often forms the basis of competence research in Germany, competencies are "context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains" (Koeppen et al., 2008, p. 62). In other conceptualizations however, the term "competence" is not restricted to the cognitive dimension, but also includes motivation, volition, affection and attitudes. Weinert (2001, p. 2433), for example, refers to competencies as "combinations of those cognitive, motivational, moral, and social skills available to (or potentially learnable by) a person [...] that underlie the successful mastery through appropriate understanding and actions of a range of demands, tasks, problems, and goals".

The NEPS takes a broad view on competencies, but has decided to distinguish systematically between cognitive and non-cognitive components and to assess them separately (for detailed information on the selection, rationale and conceptualization of competencies within the NEPS, see Weinert et al., 2011). This decision was made for several reasons: From an analytical point of view, the advantage of modeling and assessing different competence dimensions separately lies in the possibility of analyzing the interplay and relationships between them. From the perspective of a longitudinal reconstruction of competence development, the limitations and challenges of modeling and measuring competencies in a coherent way across the entire life span have to be taken into account. While the internal dynamics of development and change of some competencies can and will be reconstructed over the life course, the longitudinal reconstruction of other competencies is difficult and does not lie at the heart of the NEPS.

The definition of competencies as context-bound and domain- as well as demand-specific implies that competencies are the result of learning processes and that they can be acquired. Therefore, they must be distinguished from generalized, context-free cognitive dispositions (such as intelligence) which are learnable only to a limited extent (cf. Koeppen et al., 2008). Nonetheless, the NEPS includes these domain-general abilities in order to analyze their impact on the acquisition of domain-specific competencies (cf. Weinert et al., 2011).

In sum, the NEPS addresses the following competence domains and generalized abilities (for details, see Weinert et al., 2011):

A) Domain-general cognitive abilities (e.g., "fluid intelligence", "cognitive mechanics") which are assessed using perceptual speed and figural reasoning as relatively culture-fair and language-free indicators;

B) Domain-specific cognitive competencies, e.g., mathematical literacy, which are subject-bound during school age and become basic, cross-curricular competencies in later life. Three of these competencies, i.e., German language competencies, mathematical literacy and scientific literacy, will be assessed consistently and coherently across all stages of education throughout the life course. In addition, indicators of foreign language competencies will be measured;

C) Meta-competencies and social competencies: While the competence areas mentioned above focus on cognitive, educationally relevant competencies in a narrow sense, the third category refers to metacognitive and non-cognitive competencies. Featuring metacognition, self-regulation, ICT literacy and social competencies, this category includes those competencies that are often labeled "key competencies" (cf. Rychen, 2003).
D) Stage-specific (curriculum- or job-related) attainments, skills and outcome measures.

While all of the competencies or abilities in areas A to C are addressed in every educational stage and cohort of the NEPS – either directly using tests or indirectly through the collection of self-report data – the fourth competence area (stage-specific (curriculum- or job-related) attainments, skills and outcome measures) will be included only for selected stages of education. In the higher education stage of the NEPS, for example, we will carry out a test of subject-specific competencies in selected subject areas.

In addition to these competencies and general abilities, the NEPS collects data on stable personality dimensions (e.g., the Big Five, self-esteem) and on motivation (e.g., achievement motivation, personal goals, general interest orientation and topic-related interests; a complete overview is given by Wohlkinger et al., 2011).

Subject-specific Competencies in Higher Education

For the higher education stage, the NEPS uses two approaches to measure subject-specific competencies, that is, competencies that refer to a particular field of study. On the one hand, we are using self-report instruments that are applicable to the entire sample of higher education students; these self-report measures will be collected several times (for details, see Aschinger et al., 2011). On the other hand, we are also employing tests of subject-specific competencies in selected fields of study.

Several reasons led to the decision to gather self-report data: First, whereas tests of the subject-specific competencies of higher education students hardly exist in Germany and are yet to be developed, self-assessment questionnaires are more common. Second, self-report instruments are relatively economical in terms of administration, time and money, and they can be administered to a large sample at low cost. Third, although self-assessments are criticized for being unreliable and invalid, several studies have found a systematic correlation between self-rated competencies and alternative measures of the same construct (for an overview and references, see Braun et al., 2008). Fourth, when self-report data on competencies and data from achievement tests are collected simultaneously, it is possible to test the validity of the self-assessment instrument.

As regards subject-specific competence tests, the NEPS will start with business administration and teacher education and will include additional subject areas in future cohorts of the panel. The reason for selecting teacher education lies in the fact that the quality of schools, the competencies of teachers and teacher education continue to be central issues in educational research and education policy. The

choice of business administration for measuring disciplinary competencies is justified by the quantitative importance of this field of study. In addition, curricula and intended learning outcomes are relatively comparable across higher education institutions and even across countries.

The tests for both subject areas will be administered at the end of the study program. Therefore, they will not allow us to analyze the dynamics of competence development, but they will measure learning outcomes after the students have passed the degree course. While we will use or perhaps adapt existing tests for the subject-specific competencies of future teachers, the test for business administration students is yet to be developed.

The competence model we will use as a basis for test selection and construction is informed by well-known and elaborate conceptualizations that were primarily developed in research on the professional competencies of teachers. For example, the COACTIV study[1] (cf. Baumert & Kunter, 2006; Kunter et al., 2007), the PaLea study[2] (cf. Bauer et al., 2010) and the TEDS-M study[3] (cf. Blömeke, Kaiser, & Lehmann, 2010a, 2010b; Schmidt, Blömeke, & Tatto, 2011) proposed a competence model that is hierarchically structured and adopts a broad, multidimensional concept of teachers' professional competencies, including both cognitive and noncognitive dimensions.

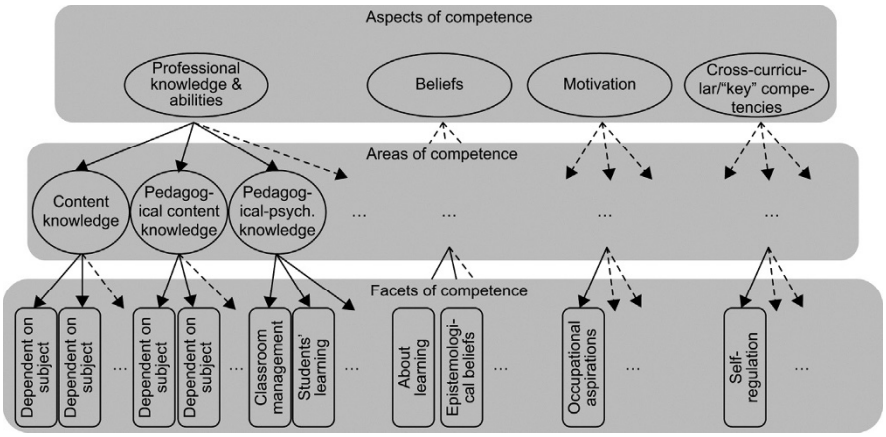


Figure 3. Professional competence of teachers (adapted from Bauer et al., 2010; Baumert & Kunter, 2006).

On the highest level of abstraction, non-cognitive prerequisites of professional competence – such as value commitments and beliefs, motivation and cross-curricular/"key" competencies – are distinguished from cognitive subject-specific competencies (see Figure 3). As research in the field of cognitive psychology has shown that experts and novices differ first and foremost in their declarative, procedural and strategic knowledge and that generic abilities play a less important role (Weinert, 1998), knowledge is central to professional competence and should

form the focus of competence assessment. Research on teachers' competencies differentiates between several areas of professional knowledge, for example – and most importantly – content knowledge, pedagogical content knowledge and general pedagogical-psychological knowledge (see the second level of Figure 3). Both content knowledge and pedagogical content knowledge refer to specific subjects taught in school, either in terms of subject matter knowledge, i.e., the teacher's understanding of the structures of the domain, or knowledge on how best to present the subject to students (for details on the concepts of content knowledge and pedagogical content knowledge, see, for example, Krauss et al., 2008). Their conceptualization and assessment are, therefore, domain-specific. As the NEPS includes (future) teachers of all subjects and because, at present, the final size of the net sample is unknown, it is uncertain whether we will be able to test content knowledge and pedagogical content knowledge. Nevertheless, it will be possible to address selected facets of general pedagogical knowledge, for example, teachers' knowledge of classroom management and student learning in general.

It will also be possible to collect data on facets of value commitment and beliefs (e.g., epistemological beliefs, subjective theories on learning and instruction), motivation (e.g., occupational aspirations), "key" competencies and personality (e.g., self-regulation, self-efficacy, social competencies). Some of these facets belong to the "core" survey program of the NEPS and will definitely be addressed (e.g., occupational aspirations, interests, social competencies, the "Big Five" personality traits, general self-concept). Whether or not we will be able to include additional aspects that are particularly relevant to the professional competence of teachers is not yet known.

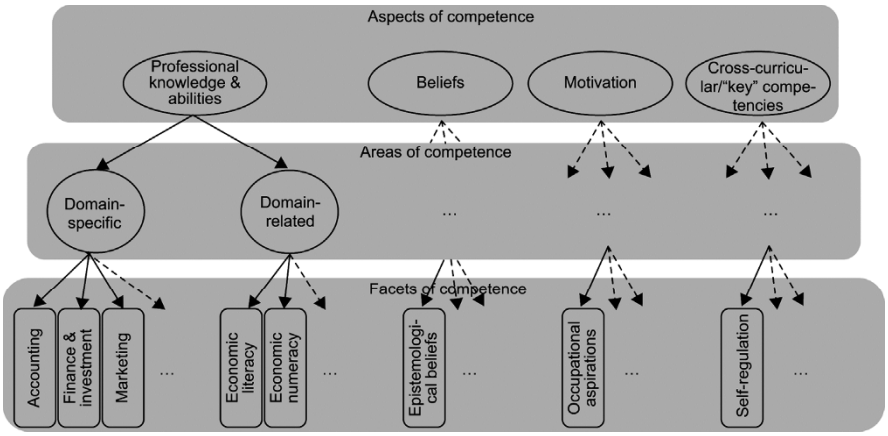


Figure 4. Professional competence of higher education graduates in business administration (adapted from Bauer et al., 2010; Baumert & Kunter, 2006; Winther, 2010).

A similar competence model is used to conceptualize the professional competencies of higher education graduates in business administration (see Figure 4). While in the

competence model for business administration the highest level of abstraction is identical to the competence model for the teaching profession, the lower levels had to be modified in order to be applicable to the particular subject area.

As regards professional knowledge and abilities, we distinguish between a domain-specific area of competence that refers to business administration in a narrow sense and domain-related fields. In differentiating between domain-specific and domain-linked competencies, we follow Winther and Achtenhagen (2008) and Winther (2010). While domain-specific competence relates to the accomplishment of tasks within the domain, domain-related competence may facilitate coping with domain-specific requirements. In the NEPS however, domain-related competencies such as economic literacy and economic numeracy are not assessed directly, but approximated by measuring literacy and mathematical competencies in general.

As regards the domain-specific knowledge of business administration students, the NEPS is unable to cover all of the sub-domains. Due to the restricted testing time, we had to select the most important ones. In order to identify those sub-domains that should and could be addressed, we chose a curriculum-oriented approach, analyzed module descriptions of 26 bachelor degree courses at universities and universities of applied sciences and categorized the curricular information according to the classification of business administration into six functional areas proposed by Haunerdinger and Probst (2006) (cf. Aschinger et al., 2011). This analysis led to the result that the major sub-domains of the core curriculum for business administration, i.e., compulsory courses, are accounting, management and organization, finance and investment and marketing. Due to the quantitative significance of accounting, we decided to include this sub-domain in the test. In addition, we prefer to include the areas of finance and investment and marketing in the test. The sub-domain of management and organization makes up a slightly larger proportion of the compulsory study program than finance and investment, but it is relatively heterogeneous with regard to the topics addressed. The reason for opting for marketing lies in the fact that this sub-domain is preferred by women. Whether the proposed structural model of domain-specific competencies in business administration holds true is an open question that has to be empirically examined.

CONCLUSION: POTENTIAL AND LIMITATIONS OF THE NEPS FOR MEASURING COMPETENCIES IN HIGHER EDUCATION

One major advantage of the NEPS is that a broad range of different competencies are measured. The NEPS therefore can address a variety of important research questions which have not yet been answered satisfactorily. The NEPS, for example, is the first study that will shed light on how basic domain-specific competencies, such as German language competencies and mathematical literacy, develop over the entire life course – from the first years of school and adolescence to adulthood and retirement – and why they develop differently. As the longitudinal and lifelong measurement of basic domain-specific competencies is a central issue for the NEPS, these competence areas are assessed in the higher

education stage as well. The inclusion of basic domain-specific competencies in the higher education stage also opens up a unique opportunity to answer the question: In what way do these competencies contribute to the acquisition of competencies which are specific to tertiary education? In a similar vein, measuring domain-general cognitive functions makes it possible to analyze the relationship between these generalized abilities on the one hand and the acquisition of basic domain-specific competencies or subject-specific competencies in higher education on the other.

With a few exceptions, the issue of competence assessment in German higher education has thus far been neglected. It was not until recently that new initiatives began to advance the construction and implementation of instruments which are suitable for competence measurement in higher education (see, for example, the funding initiative entitled "Modeling and Measurement of Competencies in Higher Education" by the German Federal Ministry for Education and Research). As a consequence, valid and reliable instruments are rare or have yet to be developed. As the test construction process is time-consuming, the NEPS decided to restrict the measurement of subject-specific competencies in higher education in the first step of the research project to two fields of study – namely teacher education where some tests already exist, and business administration – and to focus on a single measurement at the end of the study program. It is, however, the aim to include additional subject areas such as engineering and medicine in future funding periods of the NEPS.

NOTES

[1] COACTIV: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz [Professional Competence of Teachers, Cognitively Activating Instruction, and Development of Students' Mathematical Literacy].

[2] PaLea: Panel zum Lehramtsstudium [Panel for Teacher Certification Courses].

[3] TEDS-M: Teacher Education and Development Study in Mathematics.

REFERENCES

Aschinger, F., Epstein, H., Müller, S., Schaeper, H., Vöttiner, A., & Weiß, T. (2011). Higher education and the transition to work. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 267–282). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 87–101). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bauer, J., Drechsel, B., Retelsdorf, J., Sporer, T., Rösler, L., Prenzel, M., & Müller, J. (2010). Panel zum Lehramtsstudium – PaLea: Entwicklungsverläufe zukünftiger Lehrkräfte im Kontext der Reform der Lehrerbildung. *Beiträge zur Hochschulforschung, 32*(2), 34–55.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520.

Blömeke, S., Kaiser, G., & Lehmann, R.. (2010a). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.

Blömeke, S., Kaiser, G., & Lehmann, R.. (2010b). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich*. Münster: Waxmann.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The national educational panel study: Need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften.

Braun, E., Gusy, B., Leidner, B., & Hannover, B. (2008). Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). *Diagnostica, 54*(1), 30–42.

Gross, C., Jobst, A., Jungbauer-Gans, M., & Schwarze, J. (2011). Educational returns over the life course. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 139–153). Wiesbaden: VS Verlag für Sozialwissenschaften.

Haunerdinger, M., & Probst, H.-J. (2006). *BWL visuell: Basiswissen Betriebswirtschaft für Fortbildung und Praxis*. Berlin: Cornelsen.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie, 216*(2), 61–73.

Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., Brunner, M., Jordan, A., Krauss, S., Löwen, K., Neubrand, M., & Tsai, Y.-M. (2007). Linking aspects of teacher competence to their instruction: Results from the COACTIV project. In M. Prenzel (Ed.), *Studies on the educational quality of schools: The final report on the DFG Priority Programme* (pp. 32–52). Münster: Waxmann.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M., & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology, 100*(3), 716–725.

Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 121–137). Wiesbaden: VS Verlag für Sozialwissenschaften.

Rychen, D. S. (2003). Key competencies: Meeting important challenges in life. In D. S. Rychen & L. H. Salganik (Eds.), *Key competencies for a successful life and a well-functioning society* (pp. 63–107). Cambridge, MA: Hogrefe & Huber.

Schmidt, W. H., Blömeke, S., & Tatto, M. T. (2011). *Teacher education matters. A study of middle school mathematics teacher preparation in six countries*. New York: Teacher College Press.

Schuetze, H. G., & Wolter, A. (2003). Higher education, non-traditional students and life-long learning in industrialized countries. *Das Hochschulwesen, 51*(5), 183–189.

Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 103–119). Wiesbaden: VS Verlag für Sozialwissenschaften.

Weinert, F. E. (1998). Vermittlung von Schlüsselqualifikationen. In S. Matalik & D. Schade (Eds.), *Entwicklungen in Aus- und Weiterbildung: Anforderungen, Ziele, Konzepte* (pp. 23–43). Baden-Baden: Nomos.

Weinert, F. E. (2001). Competencies and key competencies: Educational perspective. In N. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 4, pp. 2433–2436). Amsterdam: Elsevier.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: Bertelsmann.

Winther, E., & Achtenhagen, F. (2008). Kompetenzstrukturmodell für die kaufmännische Bildung. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 104*(4), 511–538.

Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M., & Blossfeld, H.-P. (2011). Motivational concepts and personality aspects across the life course. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study (NEPS)* (Zeitschrift für Erziehungswissenschaft, Special Issue 14) (pp. 155–168). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hildegard Schaeper
HIS-Institute for Research on Higher Education
Hannover, Germany

OLGA ZLATKIN-TROITSCHANSKAIA, MANUEL FÖRSTER AND CHRISTIANE KUHN

MODELING AND MEASURING UNIVERSITY STUDENTS' SUBJECT-SPECIFIC COMPETENCIES IN THE DOMAIN OF BUSINESS AND ECONOMICS – THE ILLEV PROJECT

Current research provides very little empirical evidence regarding the influence of academic higher education on the development of professionalism among students. Over the course of the reform of the higher education systems in Europe (Bologna declaration) this issue has become increasingly important. So far, on the European level the effects of the discontinuation of the Diplom program and the implementation of the new bachelor/master (BA/MA) program are not sufficiently known. Therefore, in the ILLEV research project financed by the German Federal Ministry of Education and Research (BMBF), the effects of the different programs of study on professionalism and its development in the domain of business and economics are compared.

The project is focused on modeling, measuring and assessing business and economics competence while controlling other relevant parameters such as dimensions of intelligence, motivation, epistemological beliefs and socio-demographic variables. In the context of a longitudinal analysis, Diplom and BA/MA students of business and economics and of business and economics education (n = approx. 1000) were accompanied and questioned over four years (fall 2008, 2009, 2010, 2011) and their development of competence was analyzed. Existing and validated tests were used for the measurement of economics and business competence. Another part of the project is the assessment of didactical competence in the domain of business and economics. Therefore, a test is being developed and will be validated by the end of the project.

INTRODUCTION

The importance of professionalization in a knowledge-based society is undisputed. Current research, however, provides very little empirical evidence regarding the influence of academic higher education on the development of professionalism among students. The many actions taken in Germany in response to the Programme for International Student Assessment (PISA) debacle in school education gave cause to hope for similar empirical research in higher education. Because of the Bologna declaration these questions have become increasingly important. So far, on the German and European levels the effects of the

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 159–170.

discontinuation of the Diplom program and of the newly implemented BA/MA program are not sufficiently known. One of the main reasons is the lack of adequate theoretical models and reliable processes for measuring academic competencies among students (cf. Kuhn & Zlatkin-Troitschanskaia, 2011).

In the ILLEV[1] research project, the effects of the different courses of study on professionalization and its development in the domains of economics and business are compared. It is one of the few projects in the BMBF funding program "University Research as a Contribution to Professionalizing Higher Education" that also focuses on modeling and measuring subject and subject-didactical competence, especially among students of business and economics and business and economics education.

In this paper, a discussion of the basic *aims and research questions* of the ILLEV project, its *research design* and the *survey instruments employed* will be presented first, followed by descriptions of *the sample* and the first *content and measuring results* from the fall 2008 survey on modeling and the measurement of the degree of professional competence in business and economics students (for further comparative analysis and information on the test validity cf. Förster & Zlatkin-Troitschanskaia, 2010; for longitudinal analysis of the four surveys cf. Happ, in preparation). The paper concludes with a *discussion* which includes an overview of further approaches.

AIMS AND RESEARCH QUESTIONS OF THE PROJECT

The ILLEV (innovative teach-study-network in academic higher education, cf. www.wipaed.uni-mainz.de/illev/) research project makes use of the historically unique situation of an almost natural experiment in which the traditional degree program and the new consecutive program coexist. It compares the old Diplom model with the new BA/MA model in business and economics (with or without the teaching perspective) regarding their effects on the scope and development of professionalism over the course of the four-year programs. The reference analyses allow empirically-based statements to be made regarding the question of whether the new consecutive BA/MA model contributes to a higher level of professionalism. In this context professionalism is seen as a result of individual personal traits and structural factors in the degree program. The assessed structural factors can be divided into the following three levels: (1) the macro level of formal and organizational parameters (i.e., lecture times, spatial and personal environment); (2) the meso level with elements of the development of the teach-study process (i.e., e-learning, examination procedures); and (3) the micro level of factorially-implemented curricular and didactical/methodological parameters.

The project focuses on modeling and measuring the cognitive dimensions of professionalism, that is, subject-specific and didactical competence in the field of business and economics. It is based on Klieme and Leutner's (2006) definition of competence as a context-specific performance disposition. This definition grants high importance to subject-specific knowledge and thinking when it comes to conceptualizing competence (for further definitions of competence cf. Baumert & Kunter, 2006; Bromme, 1997; Shulman, 1986, 1987; Weinert, 2001).

The central research questions of the ILLEV project concentrate on two topics: (1) the influence of the traditional and the new consecutive study models on the development and scope of the professionalism of students in the domain of business and economics – for the project, professional competence was *modeled by means of, for example,* IRT- and MIMIC-analysis and was assessed at four different points in time; and (2) the identification and quantification of individual features (i.e., previous knowledge, motivation) and structural factors in the degree program that have predictor or mediator characteristics.

RESEARCH DESIGN

To assess the development of competence among students of the new consecutive BA/MA model of business and economics ("intervention group") over the course of time, *longitudinal surveys* were conducted over three years: fall 2008, 2009 and 2010. The fourth and last survey takes place in fall 2011. According to cohort design, two identified cohorts can be divided into the individual study phases as follows: cohort 1 is surveyed at the beginning of university studies (t1), after the BA orientation phase (t2), after the BA specialization phase (t3) and at the end of the BA degree program (t4); and cohort 2 is assessed after the BA orientation phase (t1), after the BA specialization phase (t2), at the end of the BA degree program (t3) and after the first phase of the Master program (t4).

For a systematic comparison of the traditional and new consecutive degree programs, Diplom students (length of study: eight semesters) have been and are being surveyed ("control group") at the four test dates with the same survey instruments. Apart from students of business and economics (BA/MA and Diplom), students of business and economics education[2] also were surveyed (BA/MA and Diplom). This enables inter alia statements to be made regarding whether the development of professional competence among students without a teaching perspective is identical to its development among students with a teaching perspective or whether the development of professional competence is influenced by other structural or personal factors. A comparison of these sub-groups is possible because the contents of the programs are the same during the basic study period (Diplom) and the orientation and consolidation phase (BA) for both business and economics programs and business and economics education programs.

The total sample consisted of 901 students in fall 2008 (t1), 800 students in fall 2009 (t2) and 1243 students in fall 2010 (t3). Students not only were assessed regarding their business and economics competence, but also regarding personal traits (age, job experience, etc.) and structural factors which can influence the development of professionalism positively or negatively.

INSTRUMENTS EMPLOYED

As well as questions taken from the economics education test (WBT), the questionnaire in the first survey (fall 2008) contained questions taken from the intelligence-structure test (IST) (analogies and numeric tasks, Liepmann, Beauducel,

Brocke & Amthauer, 2007; Amthauer, 1970), questions on attitudes towards economic circumstances (Beck, 1993), questions on the study interest or choice of degree course (inter alia items from the questionnaire on study interest (FSI), Schiefele, Krapp, Wild & Winteler, 1993) and questions on socio-demographic data.

The dimensions of economic knowledge and thinking were assessed with the help of the economics education test (WBT). The WBT is the German version of the English Test of Economic Literacy (TEL) by Beck and Krumm (Beck, 1993; Beck, Krumm & Dubs, 1998). Soper and Walstad (1987) developed the TEL which permits differentiation between relatively low and relatively high levels of development of economic knowledge and thinking. The translated test was adopted in a number of German-speaking countries, facilitating international comparison (Lüdecke-Plümer & Sczesny, 1998). According to classical test theory, the measurement features and quality factors of the WBT have been researched and validated for both the English and German versions (Beck & Krumm, 1990; Beck, Krumm & Dubs, 2001; Soper & Brenneke, 1981; Soper & Walstad, 1987). Thus the WBT is an adequate tool to use to assess economics knowledge and thinking.

Beck et al. (2001) recommend the use of the WBT, especially in the field of vocational business training. The WBT was created for target groups with an economic knowledge one level below university level, although some questions are at university level. For the ILLEV project, the test was assessed inter alia so a decision could be made regarding the extent to which it is suitable for measuring competencies in higher education or whether it is too easy for university students. Attention was paid especially to the occurrence of possible ceiling effects and selectivity indexes of the items. The curricular validity of the test is being reviewed, as it was designed according to high school and college curricula. The project investigates if and to what extent the test reflects the content of university economics studies (here: BA/MA and Diplom).

The WBT consists of two parallel versions with 46 items each; 15 tie items allow the two versions to be compared. The questions can be divided into the four economic sub-domains "basics of economics," "international relations," "microeconomics" and "macroeconomics." The questions also were arranged theoretically according to Bloom's cognitive levels[3] (Beck, 1993; Soper, 1979).

The survey used a version of the WBT containing 33 multiple choice items with one correct answer out of four options for each item. Thus, the original version was cut back by 13 items because the processing time had to be reduced for organizational and motivational reasons. To guarantee the curricular validity of the test, not only were the curricula analyzed, but the lecturers of the relevant classes also were surveyed. They had to assess the various items of the WBT according to their curricular relevance and difficulty.[4] The domain "international relations" was erased for curricular reasons (eight questions). It is not a substantial part of the BA or Diplom basic studies in economic science (or business administration and economics) at the University of Mainz. The items used in the survey which cover the three other domains correspond well with the curriculum. Two further items in the survey had complex graphics and were left out as they were not practical. The other three questions were left out because of their theoretical allocation to Bloom's taxonomy and the item-specific results (Beck, 1993).[5]

The second survey (November 2009) included the Business Administration Knowledge Test (BAKT) (Bothe, Wilhelm & Beck, 2005) that was validated curricularly as a further instrument. Together with the WBT it was used to assess business and economics competence. The BAKT implemented during the third survey in November 2010 focuses on declarative business competence although it was created for the university level. Before this test could be implemented as part of the project it had to be established that the BAKT meets the requirements for measuring business competence in higher education (i.e., assessing declarative, procedural and conative knowledge dimensions).

SAMPLE, ANALYSIS OF REPRESENTATIVENESS AND WEIGHTING

The first survey was part of the longitudinal project and took place during the winter semester 2008/09. The students were surveyed during classes.[6] Project assistants controlled the questions and presented the survey to the subjects. In total, 901 students of business and economics and business and economics education were surveyed and the responses of 743 of them were analyzed after the content and subject relationship had been checked.[7] Some 44.5% male and 55.5% female students (three missing values) were surveyed; 84% had German as their mother tongue (two missing answers). About one-fifth of the students had completed vocational training before their university studies. The average grade at school leaving was 2.34[8] with a standard deviation of 0.56 (61 missing values).

The representative analysis revealed some differences between the sample and the population regarding the distribution of certain features,[9] for example, students in the Diplom degree were underrepresented in the sample (cf. Chart 1).

Table 1. Quota of degree courses of the population, the unweighted and the weighted samples

		Quota of the population	Quota of the unweighted sample	Quota of the weighted sample
Degree program	Diplom business administration	29.34%	16.99%	30.37%
	Diplom economics	12.72%	5.75%	10.44%
	Diplom business and economics education	10.88%	13.01%	11.74%
	BA business and economics education	6.23%	10.00%	6.45%
	BA business and economics	40.82%	54.25%	41.00%
Total		100.00%	100.00%	100.00%

For that reason the sample cases were weighted to align the distribution of known parameters with the population. Several matching processes (classification tree

method, propensity scores, distance weighting; cf. Rässler, 2002) were used to match the subjects from the sample with people from the population. The weighting reflected how many people in the population were represented by one subject in the sample. The matching processes were supported by bootstrapping analyses to stabilize the estimation parameters. In the next step, the extent to which the weighted sample approximated the individual distribution and the relationship structure of the variables within the population were analyzed. The best approach was offered by the weighting that averaged the classification tree and distance weighting. One of the main results of the analysis shows that even the unweighted sample was a good approximation to a relationship structure in the population, whereas the basic distribution of variables in the population is represented only moderately. Hence, all analyses have been conducted with weighted and unweighted samples. Different analysis results can be an indicator of sample effects.

COMPARING ANALYSES OF PROFESSIONALISM BASED ON THE WBT SUM SCORES

The distribution of the sum score added up to an average (MW) of 20.91 correct answers to 33 items and a standard deviation of 4.77. The distribution's skewness is-0.383, the kurtosis is -0.122 (cf. Chart 1). It is obvious that only a small number of students gave correct answers to all 33 questions. Clearly, most students can improve. The results show that the test is not too easy and there are no obvious ceiling effects.

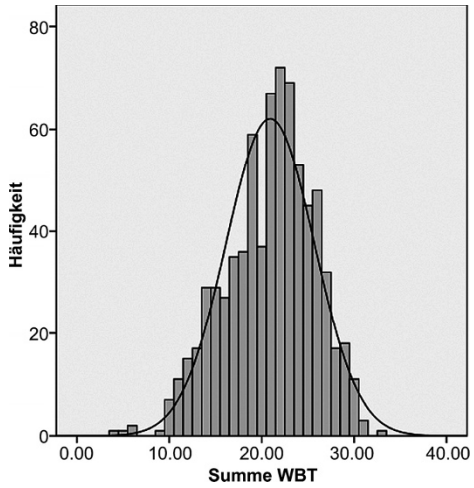


Chart 1. Distribution of the WBT's sum score within the sample

First, a simple t-test and a Mann-Whitney-U test were conducted to compare the scores of the BA and Diplom groups. In the unweighted sample the Diplom students reached an average of 22.09 correct answers; the BA students only reached an average of 20.28. The differences were even greater in the weighted

sample where the Diplom students gave correct answers to 21.96 questions and the BA students only to 19.52. In both samples the differences between the two groups in the t- and Mann-Whitney-U tests were significant (p = 0.00).

The average score of students of business and economics education (prospective teachers) was slightly higher than that of students of business and economics (no teaching ambitions). The Diplom business and economics education students scored an average of 22.64 whereas the Diplom economics students scored an average of 21.77. Furthermore, BA business and economics education students on average answered 21.41 questions correctly and BA business and economics students only correctly answered 20.06. Significant differences in the results can be found when we compare the business and economics students with the business and economics education students with the help of a simple t-test or a Mann-Whitney-U test. BA students of economics education had significantly higher scores than BA students of business and economics (p- = 0.016 in the t-test). A comparison of Diplom students of business and economics education and Diplom students of business and economics shows a p-value of 0.093 in the t-test and, thus, the average difference between the two groups cannot be regarded as significant.

The Diplom students usually were further on in terms of semesters and had attended more business and economics classes; for this reason other relevant influences (i.e., duration of studies, number and character of classes attended) had to be controlled to make an objective comparison. Because of this, a linear regression was conducted that included further declarative variables in the model (Table 2).

Table 2. Weighted and unweighted regression on the sum score of the WBT (beta coefficient and level of significance)

WBT (included)				
R = 0.548 **R = 0.574**	**R^2 = 0.300** **R^2 = 0.330**	**Korr R^2 = 0.292** **Korr R^2 = 0.321**	**Unweighted Weighted**	
	Coefficient (unweighted)		Coefficient (weighted)	
	B	Significance	B	Significance
(Absolute term)	20.62	0.00	21.37	0.00
BA	0.55	0.30	–0.31	0.59
Female	–1.65	0.00	–1.63	0.00
Other language	–1.89	0.00	–1.12	0.00
Semester	0.79	0.00	0.56	0.00
No job training	–1.31	0.00	–1.32	0.00
Analogy score	0.30	0.00	0.30	0.00
Mathematics score	0.27	0.00	0.32	0.00
School leaving grade	-0.90	0.00	-0.25	0.04

The regression analysis shows that 30% (weighted 33%) of the variance in economics knowledge and thinking can be explained by the variables study model, gender, mother tongue (German or other), number of semesters, completed vocational training, school leaving grade and the intelligence sub-scale analogies and numeric tasks. As the beta coefficient of the BA degree program is not significant when the other variables are controlled, a systematic study model effect cannot be assumed. The male subjects on average correctly answered 1.5 more questions than the female subjects. The following surveys will concentrate in particular on the question of whether this can be attributed to the questions or answering format effects (cf. Spiel, Schober & Litzenberger, 2008). A higher school leaving grade,[10] higher marks on the intelligence tests (analogies and numeric tasks), completed vocational training and German as mother tongue also led to higher test results in economic knowledge and thinking. As expected, the number of years of study at the university level also influenced the results because with every semester completed an average of 0.56 (weighted) and 0.79 (unweighted) more items can be answered correctly. The estimations of the beta coefficient of most cause variables do not differ greatly between the weighted and unweighted samples. The results of the two samples only differ substantially regarding the independent variables BA students (0.55 vs. -0.31), other mother tongues (-1.89 vs. -1.12) and school leaving grade (-0.90 vs. -0.25).

The sum score was implemented as it was assumed that all items used assessed economics knowledge and thinking to the same extent. Hence every item has the same selectivity and the individual items only differ regarding their difficulty. The basic underlying measurement model is a simple dichotomous Rasch model (1PL model). However, this assumption is very restrictive and it will be reviewed to see if a 2PL model (Birnbaum model) can explain the data more accurately (cf. i.e. Hambleton & Swaminathan, 1990; Hartig, 2009).[11] If the 2PL measurement model clearly provides better adjustment, the next step will be to calculate a MIMIC model based on this alternative to identify the influence of various factors on expertise as well as possible item functions (for further analysis cf. Förster & Zlatkin-Troitschanskaia, 2010).

CONCLUSION AND DISCUSSION

The regressions of the sum score of the WBT indicate that BA students have the same level of expertise as Diplom students. The weighted and the unweighted models both showed no significant effects between the two study models.[12]

The restructuring of degree programs at the University of Mainz began in 2007 and, thus, still is in its preliminary phase. It will be very interesting to see if the results are repeated in the longitudinal surveys that follow and especially at the end of the BA program. In addition, only economics expertise and thinking have been compared. Further research needs to be conducted with regard to the question of the extent to which other competence dimensions are influenced by the study models. The results of the second and third surveys will be used to assess the

differences in development between the students in the two study models and the types of development that can be identified within the programs.

When the degree programs in business and economics education and business and economics were compared, the students in business and economics education had higher scores. This effect was only significant among the BA students; there was no systematic difference among the Diplom students. The responses to surveys which will be conducted later in the project will be used to evaluate whether the positive results from the students in business and economics education (prospective teachers) can be confirmed and if they are, what the possible reasons for this could be.

Regarding the most important variables explaining economic competence, the regression confirms the high explanatory potential of gender and mother tongue. Other German and international studies on general/specific knowledge had similar results (i.e., Walstad & Robson, 1997; Lüdecke-Plümer & Sczesny, 1998). This is especially significant because of the increasing relevance of performance surveys in higher education and the increasing number of admission tests owing to the change in program models (especially for MA programs). The reasons for the better test results among male subjects will be scrutinized over the course of the project. Among other things, whether the BAKT shows comparable gender-specific differences from the WBT will be assessed. Influences other than gender differences, such as the question/answer format and the use of study strategies, also will be assessed.

As expected, German native speakers have an advantage in the WBT. The test consists of full, grammatically-complex sentences and, thus, requires a certain level of language skills. There also is evidence to indicate that vocational training and a good school leaving certificate positively influence economics knowledge and thinking (cf. Beck & Wuttke, 2004). Vocational training obviously teaches test-relevant contents. The grade of the school leaving certificate could indicate ambition, study strategies and general test performance, which are also good predictors of knowledge and university performance. Further models include other latent variables (i.e., dimensions of intelligence, motivational and attitude features) as factors that explain economic competence.

Our analyses further show that the results of the weighted and unweighted samples do not differ significantly; however, in some cases these small differences could be first indicators of potential sample effects in the first survey. Whether these differences will be visible in further samples and the necessary weightings in the second and third surveys will be scrutinized.

The survey among lecturers at the University of Mainz also showed that the contents of the items are anchored in the university curriculum. On the basis of the project results so far, the WBT can be seen as a tool sufficiently applicable and valid for measuring economics competence during the *first phase* of the BA and Diplom degree programs.

During the remainder of the project, the extent to which the WBT can be used to show individual development regarding the professionalism of students will be assessed. Comparing the results of the first and second surveys will provide further

clues. During the next survey the BAKT will be assessed regarding its psychometric suitability for measuring business competence. The performances of the students of the different study models and degree programs will be compared again (lateral and longitudinal). It will be of great interest if and to what extent the two test methods, WBT and BAKT, illustrate one or several easily definable dimensions of subject-specific competence. In addition, the self-developed test for measuring didactical competence in business and economics will be validated as part of the next survey (cf. Kuhn, 2011). It will be interesting to see how closely the subject-related (content) competence and the didactical competence correlate.

NOTES

[1] The project is led by Prof. Dr Olga Zlatkin-Troitschanskaia (University of Mainz). Co-operation partners are the University of Tübingen (Prof. Dr Martin Biewen), the University of Berlin (Dr Sigbert Klinke), the study seminar for vocational schools in Mainz (Prof. Dr Markus Böhner) and the University of Applied Sciences, Mainz (Prof. Dr Daniel Porath).

[2] Because only students of business and economics education were assessed regarding their subject-didactical competence, a test was developed and validated during the project (cf. Kuhn, 2011). The research question focuses on the influence of various phases of teacher training (theoretical, student teaching, reflection) on the development of subject-didactical knowledge. A major comparison between the old study model and new study model, which is defined by a more practical orientation, is possible here as well.

[3] Bloom's cognitive taxonomy divides study goals or requirements for completing tasks into six consecutive categories of increasing complexity (Bloom, 1971).

[4] The survey was conducted online.

[5] The complete processing time was 75 minutes. The greater part of the test focuses on assessing cognitive ability dimensions so the processing time, which calls for high concentration by subjects, already is very long. Hence, the WBT had to be shortened.

[6] The classes were selected so that as many BA and Diplom students as possible studying in different semesters could be surveyed without students being assessed twice in different classes. Therefore, classes with mostly different students were selected.

[7] For example, only students with business and economics or business and economics education as their majors were included in the sample.

[8] The German grading system starts with one, very good (equivalent to an A), and goes up to six, insufficient.

[9] Comparing the samples with the population is possible thanks to anonymous data for the criteria of gender, degree program, age, university and semester, school leaving grade and federal state in which the school leaving certificate was granted.

[10] According to the German grading system in schools, a very good performance is graded with "1," a good performance with "2," a satisfactory performance with "3," a sufficient performance with "4," a deficient performance with "5" and an insufficient performance with "6."

[11] The 3PL model was not used as it calls for 1000 subjects per item to guarantee a precise estimation of its parameters, according to Eggen (2008).

[12] Further analysis based on a MIMIC Model also showed no differences in the WBT scores or differential item functions between the two groups.

REFERENCES

Amthauer, R. (1970). *Intelligenz-Struktur-Test IST 70*. Göttingen: Hogrefe.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520.

Beck, K. (1993). *Dimensionen der ökonomischen Bildung: Meßinstrumente und Befunde*. Nürnberg: Universität Erlangen-Nürnberg.

Beck, K., & Krumm, V. (1990). *William B. Walstad/ John C. Soper: Test zur wirtschaftskundlichen Bildung. Manual*. Zweite Ausgabe. Auszugsweise ins Deutsche übertragen, ergänzt und kommentiert. Nürnberg: Salzburg.

Beck, K., Krumm, V., & Dubs, R. (1998). *Wirtschaftskundlicher Bildungs-Test (WBT)*. Göttingen: Hogrefe.

Beck, K., Krumm, V., & Dubs, R. (2001). WBT - Wirtschaftskundlicher Bildungstest. In W. Sarges & H. Wottawa (Eds.), *Handbuch wirtschaftspsychologischer Testverfahren* (S. 559–562). Lengerich: Pabst.

Beck, K., & Wuttke, E. (2004). Eingangsbedingungen von Studienanfängern – Die Prognostische Validität wirtschaftskundlichen Wissens für das Vordiplom bei Studierenden der Wirtschaftswissenschaften. *ZBW, 100*, 116–124.

Bloom, B. S. (1971). *Cognitive domain. Taxonomy of educational objectives: Bd. handbook 1*. New York, NY: McKay.

Bothe, T., Wilhelm, O., & Beck, K. (2005). *Business administration knowledge: Assessment of declarative business administration knowledge: Measurement development and validation.* Humboldt-Universität zu Berlin. Berlin: Institut zur Qualitätsentwicklung im Bildungswesen (IQB).

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Ed.), *Enzyklopädie der Psychologie. Themenbereich D: Praxisgebiete. Serie 1: Pädagogische Psychologie. Bd. 3: Psychologie des Unterrichts und der Schule* (S. 177–212). Göttingen u. a.: Verl. f. Psychologie Hogrefe.

Eggen, T. J. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (S. 215–234). Göttingen: Hogrefe.

Förster, M., & Zlatkin-Troitschanskaia, O. (2010). Wirtschaftliche Fachkompetenz bei Studierenden mit und ohne Lehramtsperspektive in den Diplom- und Bachelorstudiengängen – Messverfahren und erste Befunde [Competencies in Business & Economics among Students of the "Diplom" and BA Degree Course with or without the Ambition to Become Teachers – Methods and First Results]. *Lehrerbildung auf dem Prüfstand, 3*, 106–125.

Hambleton, R. K., & Swaminathan, H. (1990). *Item response theory: Principles and applications* (5th ed.). Evaluation in education and human services series. Boston, MA.: Kluwer-Nijhoff.

Happ, R. (in preparation). The development of content knowledge in business and economics – A longitudinal approach in the ILLEV project.

Hartig, J. (2009). Messung der Kompetenzen von Lehrpersonen mit Modellen der Item-Response-Theorie. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Eds.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 295–310). Weinheim: Beltz.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik, 52*, 876–903.

Kuhn, C. (2011). *Assessing didactic competence: Developing a measuring instrument in the domain of business and economics*. Paper Presentation within the Session "Competence measurement and modeling" at the EARLI JURE Conference, August 30, Exeter, UK.

Kuhn, C., & Zlatkin-Troitschanskaia, O. (2011). *Assessment of competencies among university students and graduates – Analyzing the state of research and perspectives*. Johannes Gutenberg University Mainz: Arbeitspapiere Wirtschaftspädagogik [Working Paper: Business Education], *59*.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R: Intelligenz-Struktur-Test 2000R*. Göttingen: Hogrefe.

Lüdecke-Plümer, S., & Sczesny, C. (1998). Ökonomische Bildung im internationalen Vergleich. In K. Beck & K. Breuer (Eds.), *Arbeitspapiere Wirtschaftspädagogik Nr. 11*. Mainz: Lehrstuhl für Wirtschaftspädagogik, Johannes Gutenberg-Universität Mainz.

Rässler, S. (2002). *Statistical matching*. New York, NY: Springer.

Schiefele, U., Krapp, A., Wild, K.-P., & Winteler, A. (1993). Der "Fragebogen zum Studieninteresse" (FSI). *Diagnostica, 39*(4), 335–351.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *The Elementary School, 57*(1), 1–22.

Soper, J. C. (1979). *Test of economic literacy: Discussion guide and rationale*. New York: Joint Council on Economic Education. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/3a/1d/6d.pdf [7.12.2009].

Soper, J. C., & Brenneke, J. S. (1981). The test of economic literacy and an evaluation of the DEEP system. *The Journal of Economic Education, 12*(2), 1–14.

Soper, J. C., & Walstad, W. B. (1987). *Test of Economic Literacy.: Second Edition. Examiner's Manual*. Joint Council on Economic Education. (Ed.). New York, NY.

Spiel, C., Schober, B., & Litzenberger, M. (2008). *Evaluation der Eignungstests für das Medizinstudium in Österreich*. Vienna (Projektbericht).

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *The Journal of Economic Education, 28*(2), 155–171.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Olga Zlatkin-Troitschanskaia, Manuel Förster & Christiane Kuhn
Chair of Business Education
Johannes Gutenberg University Mainz

KAROLINE KOEPPEN[1], JOHANNES HARTIG[i], ECKHARD KLIEME[i]
AND DETLEV LEUTNER[2]

COMPETENCE MODELS FOR ASSESSING INDIVIDUAL LEARNING OUTCOMES AND EVALUATING EDUCATIONAL PROCESSES – A PRIORITY PROGRAM OF THE GERMAN RESEARCH FOUNDATION (DFG)[3]

INTRODUCTION

Social change, social cohesion, and opportunities for societal development are all dependent on the educational level of the members of a society. Current discussion in educational research emphasizes the importance of the products of educational processes, often referred to as educational *output* or *outcomes*, for human resources (Klieme & Leutner, 2006). The outcomes of education are the knowledge acquired, the abilities, skills, attitudes, and dispositions developed, and the qualifications attained.

Several large-scale international assessments of domain-specific competencies (e.g., reading literacy, science competencies) at the end of compulsory schooling (e.g., TIMMS, PISA) and in adulthood (e.g., IALS, ALL) have recently drawn increased public and scientific attention to educational outcomes and their assessment. The studies identified huge gaps between the competencies attained, on the one hand, and the goals of the education system, on the other. Clearly, effective quality development of educational processes is facilitated when the productivity of educational systems, the quality of educational institutions, and the learning gains of individuals are measurable. Thus, there has been an increasing focus within educational systems on defining and evaluating the goals to be attained by schools. In many cases, however, adequate assessment procedures are still lacking, as are procedures for analyzing and reporting the results.

The concept of competence is increasingly considered as an anchor point in this discussion. In this article, we focus on two sets of research questions that are central to the debate on the concept of competence. In the first part of the article, we discuss how competencies can be defined in general and in specific contexts. The new focus on competence has shifted attention from the measurement of general cognitive abilities to more complex ability constructs related to real world contexts. Sophisticated models of the structure and levels of these complex constructs need to be developed. Typical examples are different levels of reading literacy, mathematical modeling of real-world situations, planning and analyzing scientific experiments, or self-regulation and metacognition in domain-specific

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 171–192.

problem solving. Given the complexity of competencies, it is important that they be precisely defined in specific domains. The development of cognitive models of domain-specific performance is a central issue in this context.

The second set of research questions relates to the design and practical implementation of competence assessments. School policy and practice are moving toward *evidence-based policy and practice* (Slavin, 2002), where "evidence" often implies empirical assessments of competencies. Obviously, the assessment of competencies plays a key role in optimizing educational processes and advancing educational systems. At the same time, it is evident that assessments pursue various goals (i.e., the focus may be on individual learning outcomes, on program evaluation, or on system monitoring). Unfortunately, the difficulties and complexities of assessing learners' baseline competencies and learning gains are often underestimated in educational policy and practice. Developing appropriate measurement instruments that can be used for different purposes is a time- and resource-intensive undertaking that can only be achieved on the basis of theoretically and empirically founded cognitive models of competencies.

In this article, we give an overview of current issues in cognitive modeling and competence assessment. We first provide a working definition of the term competence and describe different goals of competence assessment. In the main part of the article, we outline the central research questions and the current state of research, identifying four main research areas. Finally, we present an interdisciplinary research program funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) to integrate, structure, and coordinate research activities relating to competence modeling and assessment.

The article focuses on current assessment practices in Germany, where the standardized assessment of student achievement does not have the same tradition as in the United States or Great Britain. In the past decade, however, the results of large-scale international assessments, particularly the PISA 2000 study (e.g. Baumert, Stanat, & Demmrich, 2001), have sparked intense discussion among both the public and educational policy makers. Extensive educational reforms have been initiated in response to the PISA findings (e.g., changes in mathematics and science instruction), and the results of the PISA 2006 assessment seem to indicate that these reforms have had positive effects on students' performance, especially in science (OECD, 2007b; Prenzel et al., 2007).

THE COMPETENCE CONCEPT AND THE CHALLENGES OF ITS ASSESSMENT

The competence concept is central to empirical studies dealing with the development of human resources and the productivity of education. Although it has been in use for decades, the term "competence" has enjoyed increasing currency in psychology and its neighboring disciplines in the last few years (e.g., Csapó, 2004; Klieme, Funke, Leutner, Reimann, & Wirth, 2001; Rychen & Salganik, 2001, 2003; Sternberg & Grigorenko, 2003; Weinert, 2001). Research uses the concept to characterize the changing demands of modern life and the working world, as well as the educational goals involved. However, its definition

remains fuzzy in educational research. It seems essential to narrow it down to specific contexts of abilities.

Drawing on Klieme and Leutner (2006; see also Klieme, Maag-Merki, & Hartig, 2007), we define competencies as context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains.

An essential element of competencies is their *context-specificity*. The concept of competence was introduced to psychology as an alternative to the focus in classical intelligence research on generalized, context-independent cognitive dispositions that are learnable only to a limited extent (e.g., McClelland, 1973; "Testing for competence rather than for 'intelligence'"). In contrast, competencies reflect a person's potential to meet cognitive demands in specific areas of learning and behavior. Competencies are, thus, more closely related to "real life". Connell, Sheridan, and Gardner (2003, p. 142) concisely characterize competencies as "realized abilities."

Having considered different theoretical and pragmatic arguments, Weinert (1999, 2001) proposed that the term competence be restricted to *cognitive* context-specific aspects, and that it should exclude motivational orientations or affective requirements for successful learning. Given that Weinert himself also discussed so-called *action competencies*, including motivation, attitudes, tendencies, and expectations in the context of competencies, this distinction was not self-evident. Nonetheless, Weinert proposed that cognitive and motivational aspects be assessed as separate constructs to allow the empirical analysis of their interaction. In this article, we focus on the cognitive aspects of competencies.

In psychological and educational practice and research, competencies often relate to specific content areas (e.g., Hartig & Klieme, 2006; Weinert, 2001). In the tradition of research on psychological expertise, we refer to these areas as *domains*. Typical domain-specific competencies in primary and secondary education include reading literacy, mathematical competence, and scientific competence.

Given their context-specificity, competencies have to be acquired by learning and experience in relevant, domain-specific situations. Consequently, they are amenable to external interventions (e.g., Baumert et al, 2001; Hartig & Klieme, 2006; Simonton, 2003). Basic cognitive abilities, in contrast, are much more difficult to train or learn (Weinert, 2001). In the construction of competence models, it is therefore important to consider and empirically examine the connections between specific competencies and basic cognitive abilities.

Valid measures of competence need to be based on theoretically sound and empirically tested competence models. These models have to (a) represent the internal structure of competencies in terms of specific basic skills and abilities, (b) describe different levels of competencies with reference to domain-specific performance, and (c) take into account changes occurring in learning and developmental processes.

In addition, measurements of competence should build on psychometric models that link the empirical measurement operations with theoretical (cognitive) models of competencies. In short, the measurement of competencies should be based on a

solid theoretical and psychometric basis that allows the measurement result (e.g., quantity and quality of solved tasks) to be interpreted with reference to an underlying theoretical model of competencies.

Valid, model-based measures of competence can be used for different purposes. First, model-based measurement instruments can inform individual educational decisions (e.g., assignment to a certain track, conferral of qualifications, provision of educational interventions). In this context, assessment focuses on individual learning outcomes; it is "a process by which educators use students' responses to specially created or naturally occurring stimuli to draw inferences about the students' knowledge and skills" (Pellegrino, Chudowsky, & Glaser, 2001, p. 20). A second goal of competence assessment is to evaluate learning outcomes on the aggregated class, school, or even system levels, rather than the individual level (Leutner, Fleischer, Spoden, & Wirth, 2007). This ranges from classroom-based assessment to large-scale standardized assessment of competence levels across whole education systems (system monitoring; for example, the National Assessment of Educational Progress [NAEP] in the United States or the OECD PISA studies).

Assessments with a focus on individual learning outcomes, on the one hand, and different aggregated levels, on the other, make distinct demands on measurement instruments (e.g., in terms of reliability and testing time). Depending on the focus of the assessment and the level of aggregation, different measurement techniques and research designs may be more or less suitable. In many cases, however, different goals have to be accomplished within the same study (e.g., system monitoring and feedback on classroom level). Thus, another challenge in the research area of competencies is to develop competence measures and research designs that simultaneously satisfy different assessment goals.

CENTRAL RESEARCH QUESTIONS AND THE CURRENT STATE OF RESEARCH ON COMPETENCE MODELING AND ASSESSMENT

The development of adequate cognitive models for contextualized competence constructs is a challenging and multifaceted task. Theoretical models must provide a basis for describing the interaction between individual abilities and the environment, different levels of competence, and developmental processes. Furthermore, they must be related to advanced psychometric techniques and translated into appropriate empirical measurement procedures. As yet, neither cognitive research nor psychometrics meets these requirements; adequate measurement concepts and models are still lacking (Prenzel & Allolio-Näcke, 2006). Both disciplines need to contextualize their models in cooperation with representatives of other disciplines, such as educational researchers and domain experts. Indeed, the Committee on the Foundations of Assessment (Pellegrino at al., 2001), founded by the US National Research Council, has called for multidisciplinary research activities focusing on three different facets:

> "(1) development of cognitive models of learning that can serve as the basis for assessment design, (2) research on new statistical measurement models and their applicability, (3) research on assessment design" (p. 284).

As Pellegrino et al. (2001) accurately summarize: "Much work remains to focus psychometric model building on the critical features of models of cognition and learning and on observations that reveal meaningful cognitive processes in a particular domain (…). Therefore, having a broad array of models available does not mean that the measurement model problem has been solved. The long-standing tradition of leaving scientists, educators, task designers, and psychometricians each to their own realms represents perhaps the most serious barrier to progress" (p.6).

We identify four key areas in this research field: first and foremost, the development of theoretical models of competence (Area 1), complemented by the construction of psychometric models (Area 2). This leads onto the construction of measurement instruments for the empirical assessment of competencies (Area 3). Research on the use of diagnostic information (Area 4) rounds off the research field. In the following, we explicate the concrete questions and problems addressed within each of the four areas and outline the current state of research.

Area 1: Development of Cognitive Models of Competencies

As mentioned above, the shift toward the competence construct has prompted efforts to improve the assessment of these complex and contextualized constructs. The first question to arise here is which models provide a basis for developing measurement instruments and interpreting their results. In current educational research, only a limited number of competence models exist. Therefore, it is important to develop cognitive models that explain interindividual differences in domain-specific performance.

A first challenge in model development is the *contextualized character of competencies*, which means that both person- and situation-specific factors have to be taken into account. For example, when describing foreign language skills with reference to situational demands, the competencies required to read a text can be distinguished from those required to engage in conversation (e.g., by distinguishing written vs. spoken text, or text comprehension vs. text production). For individuals, knowledge structures relevant to different situations must be taken into account; for example, the available vocabulary, grammatical knowledge, and mastery of socio-pragmatic rules (Chen, 2004; Kobayashi, 2002). This simultaneous consideration of individual- and situation-specific components has consequences for the structure of competencies as well as for the description of competence levels. Hence, two groups of theoretical models devised to describe and explain competencies can be distinguished: *models of competence levels* and *models of competence structures* (Hartig & Klieme, 2006; Klieme et al., 2007). Models of competence levels define the specific situational demands that can be mastered by individuals with certain levels or profiles of competencies; levels of competencies are used to provide a criterion-referenced interpretation of measurement results. These models are particularly useful for assessing and evaluating educational

outcomes on an aggregated level. Models of competence structures deal with the relations between performances in different contexts and seek to identify common underlying dimensions. These models are especially interesting for explaining performance in specific domains in terms of underlying basic abilities, and can provide a basis for more differentiated measurement results of individual-centered assessments. The two kinds of models relate to different aspects of competence constructs. They are not mutually exclusive, but ideally complementary.

The aspect of *development* is also very relevant in the context of theoretical competence models. To date, only a few competence models have addressed the issue of competence development (primarily in the domain of science; e.g., Bybee, 1997; Prenzel et al., 2004, 2005). For the most part, these models have no empirical foundation, and their conceptualizations of competence development differ. Some models see competence development as a continuous progression, shifting successively from the lowest to the highest competence level (e.g. Prenzel et al., 2004, 2005). The level of elaboration and systematization increases with the competence level (as described by Bybee, 1997, for scientific literacy). Other models conceptualize competence development as a noncontinuous process characterized by qualitative leaps (e.g., conceptual change in science; Schnotz et al., 2004; Schnotz, Vosniadou, & Carretero, 1999). This process involves a fundamental reorganization of concepts and structures from everyday life to correspond with new science-based ideas (e.g., DiSessa, 2006; Vosniadou, Ioannides, Dimitrakopoulou, & Papademetriou, 2001; Wilson, 2008).

In addition, the design of cognitive models of competencies depends on the questions addressed or the decisions to be informed. A model fitting for some purposes (e.g., giving immediate feedback) may be totally ineffective for other purposes (e.g., comparative evaluation of educational institutions). A more detailed model of competencies is needed in the first case than in the second. In one case, precise estimates might be required on an individual level, in another case on an aggregated level. Switching between two purposes can cause a whole host of problems, as recent experiences in the United States have shown (Cheng, Wanatabe, & Curtis, 2004; Fuhrmann & Elmore, 2004).

In a next step, these theoretically founded models of competencies must be used as a basis for constructing psychometric models and measurement instruments (e.g., Hartig & Höhler, 2008; Wirth & Leutner, 2008). To date, however, these efforts have rarely proved successful, primarily because many of the existing cognitive models are insufficiently elaborated. Nevertheless, some research progress has been made in specific domains, such as foreign languages in the context of the Common European Framework of Reference (Alderson, 2005; Alderson et al., 2005; Gogolin, 2002), or in the area of mathematical modeling (e.g., Blum et al., 2004). Likewise, important contributions to the theory-based formulation of competence models have been made in specific areas of cognitive psychology (e.g., Frensch et al., 2003; Haider & Frensch, 1996, 1997, 2002; Hasselhorn & Grube, 2003; Oberauer, Schulze, Wilhelm, & Süß, 2005; Schneider, Lockl, & Fernandez, 2005; Spiel & Glück, 2008; Weinert & Schneider, 1995) and in research on personality and individual differences (e.g., Kröner, Plass, &

Leutner, 2005; Leutner, 2002; Leutner & Plass, 1998; Plass, Chun, Mayer, & Leutner, 1998; Wilhelm & Engle, 2005).

However, when it comes to developing actual test items, the level of abstraction of the competence models often turns out to be too high. As a consequence, test developers have to develop huge numbers of tasks that then have to be tested empirically. Those tasks that correspond to a (usually) relatively simple psychometric model (e.g., a unidimensional Rasch model) are retained. The scales and competence levels reported in the PISA study are an example for this procedure (e.g., in reading: Artelt, Schiefele, & Schneider, 2001; in mathematics: Klieme, Neubrand, & Lüdtke, 2001; in science: Prenzel, Rost, Senkbeil, Häußler, & Klopp, 2001; in cross-curricular problem solving: Dossey, Hartig, Klieme, & Wu, 2004). In these examples, levels of competence are not theoretically specified a priori, but defined post hoc after the inspection of model-conform leftover items. From a theoretical perspective, this procedure is less than satisfactory.

To summarize, in many domains where the need for well-founded competence assessments is evident, basic research concerning theoretically as well as empirically sound models of competence structures, competence levels, and competence development is still required. Although attempts have been made to interconnect cognitive competence models with psychometric models and measurement instruments, they have often failed to meet the demands of the current, more complex definition of competencies. There is a clear need for more integrative, interdisciplinary research activities.

Area 2: Psychometric Models

As Embretson (1983, p. 184) put it, psychometric models are about "modeling the encounter of a person with an item". Psychometric models are the link between theoretical constructs and the results of empirical assessments; they provide the measurement rules by which test scores are assigned based on performance in test situations. Given the contextualized nature and complexity of competence constructs, psychometric models have to meet certain requirements (Hartig & Klieme, 2007). On the one hand, they have to incorporate all relevant characteristics of the individuals whose competencies are to be evaluated. Because competencies refer to performance in complex domains, the models should take into account that multiple abilities may be required. At the same time, they have to take into account domain-specific situational demands. Because competencies are conceptualized as context-specific constructs, the results of competence assessments should be related to the mastery of specific, domain-relevant situations. Item response theory (IRT) has a long tradition in educational assessment, and many of its past and recent developments were made for specific needs in this area. IRT allows ability estimates and item difficulties to be compared (Embretson, 2006), thus providing a basis for models incorporating individual and situational characteristics. Several recent developments in IRT hold considerable promise for the modeling of competencies, namely, *explanatory IRT models,*

multidimensional IRT models (e.g., Hartig & Höhler, 2008), and *models for cognitive diagnosis.*

Explanatory IRT models (Wilson & De Boeck, 2004; Wilson, De Boeck, & Carstensen, 2008) incorporate predictors for successful interactions of a person with an item. These predictors can be either attributes of the person or features of the item ("person predictors" or "item predictors"). Specific item features can be used to represent certain situational demands. Incorporating effects of item features into the psychometric model is a highly suitable way of constructing a psychometric model of competence that takes the corresponding demands into account. Although models including item features have been in use for some time (e.g., the linear-logistic test model, LLTM; Fischer, 1973), recent developments such as the inclusion of random effects on the items side (e.g., Janssen, Schepers, & Peres, 2004; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000) make them more flexible to model empirical data from complex performance situations. Applications of models with item predictors have been presented by Janssen et al. (2000, 2004), Hartig and Frey (2005), and Wilson et al. (2008). Whereas the analysis of item predictors is highly promising for psychometric models of competence, the use of observed person predictors for latent abilities ("latent regression," e.g. Adams, Wilson, & Wu, 1997; van den Noortgate & Paak, 2004) is of less interest in this context. Observed person predictors may, however, be incorporated into models of competence in order to model interactions between certain person characteristics, abilities, and task demands (e.g., in differential item functioning).

Models with item features that allow situational demands to be incorporated (e.g., the LLTM) are typically unidimensional. To model performance in complex situations, it may be necessary to include more than one ability dimension in the model. A straightforward way of doing so is to apply *multidimensional IRT* (MIRT) models. These models are generalizations of unidimensional models such as the Rasch model, the two-parameter logistic model, and the normal-ogive model. Instead of one ability dimension, the probability of mastering a test item is modeled as a function of multiple basic abilities (e.g., McDonald, 2000; Reckase, 1997). The most frequently applied models are compensatory models, which model the probability of success on an item as a function of the sum of all abilities relevant for an item. In these models, low ability in one dimension can be compensated by high ability in a second dimension, and vice versa. However, non-compensatory models have also been described (e.g., Whitely, 1981). MIRT models differ in their complexity in terms of how many different abilities are required to solve each item. Models in which each item draws on a single ability are said to have between-item multidimensionality (Adams, Wilson, & Wang, 1997). In factoranalytic terms, these models have a simple-structure loading pattern. MIRT models with between-item multidimensionality have been applied to data from educational assessments to take into account relations between performance in different domains (e.g., in the PISA studies; Adams, 2005; Adams & Wu, 2002). Models with between-item multidimensionality consist of multiple scales that are, within themselves, unidimensional. In contrast, models that incorporate multiple abilities for each item

are said to have within-item multidimensionality. These models are more appealing for psychometric models of competencies, because within-item multidimensionality makes it possible to model successful performance as the result of a mixture of different abilities. Models with within-item multidimensionality have been applied relatively rarely, typically to account for "nuisance" dimensions (e.g., local item dependencies within testlets; Wang & Wilson, 2005) rather than for theoretically defined ability dimensions. Examples of simple MIRT models with meaningful within-item multidimensionality are given by Walker and Beratvas (2003), Stout (2007), and Hartig and Höhler (2008).

A third development of great relevance to psychometric models of competencies has been the emergence of *cognitive diagnostic models* or *multiple classification models* (DiBello & Stout, 2007; Maris, 1999). These models assume multiple latent ability variables that are modeled as latent categories instead of latent dimensions. They can be characterized as multidimensional latent class models with specific restrictions defining which items require which abilities, and how the abilities are combined for successful performance. A conceptual strength of cognitive diagnostic models is that they are explicitly designed to model mixtures of abilities within items, and that models and estimation methods for non-compensatory mixtures are available. However, the categorical nature of the ability variables (e.g., "does know" vs. "does not know") seems more appropriate for relatively fine-grained ability constructs than for cases in which latent variables represent a broader set of required abilities. Examples of empirical applications of cognitive diagnostic models are presented by von Davier (2005) and Gierl, Leighton, and Hunka (2007).

To summarize, recent years have seen a number of significant developments in psychometrics that hold great promise for the translation of theoretical models of competencies into measurement models. Models that succeed in taking both situational characteristics and individual abilities into account can do more than provide rules for measurement (i.e., generate test scores). They can also serve as empirically testable models of the interaction between individual abilities and situational demands. However, to realize the potential of the advanced psychometric methods recently developed, these techniques need to be combined with strong theoretical models.

Area 3: Measurement Concepts and Instruments

This section examines how competence models and psychometric models can be translated into concrete empirical measurement procedures, with a focus on computer-based assessment.

Competencies are assessed in different educational contexts: in large-scale assessments (e.g., TIMSS and PISA), in evaluations of specific programs or institutions, in basic research, and in the assessment of individual qualifications or learning outcomes. Researchers and stakeholders in educational processes assess student competencies for purposes of system monitoring, to test the effectiveness of specific forms of instruction, to give feedback about individual learning progress, or to describe developments in competencies. For the most part,

standardized tests are applied. However, nonstandardized tests and observations of educational processes (e.g., teachers' observations in direct interaction with learners) are also common ways of assessing competencies. Given the complexity of competence constructs, it is important to adapt and advance these measurement concepts and instruments. They should be parsimonious, have a firm theoretical foundation, and allow inferences to be drawn about the mastery of demands in real-life situations.

Research interest in measurement concepts and instruments for assessing competencies first emerged in the 1960s and 1970s. In Germany, this was a time of educational reform, with the introduction of new teaching and learning goals to the curriculum, as well as the establishment and evaluation of new educational curricula (e.g., in schools combining different academic tracks). In this context, there was a surge in interest in the area of educational assessment at the individual, diagnostic level: the traditional field of activity for competence assessment (Klauer, 1978). Assessment instruments were developed, based on the concept of goal-oriented and criterion-referenced testing and, usually, using the binomial test model or one of its derivates (Klauer, 1987). At the same time, IRT became increasingly popular and widespread in educational measurements. In some countries (e.g., the Netherlands and the United States), this tradition has continued unabated (van der Linden & Hambleton, 1996).

Given the complexity of competence constructs and the need to understand the different abilities and processes that lead to success in real-life situations, it has become increasingly important that assessment procedures are based on cognitive models of competence. An excellent example of empirical assessments based on theoretical models of competence is the Berkeley Evaluation & Assessment Research (BEAR) Center, which focuses on the model-based assessment of competencies in science education (Wilson, 2008; Wilson & Draney, 2004; Wilson & Sloane, 2000). In DESI, a large-scale assessment of language competencies in Germany, measurement instruments and measurement models were developed on the basis of cognitive and linguistic models of language competence and language acquisition (e.g., Beck & Klieme, 2007; Klieme et al., 2008; Nold, 2003; Nold & Rossa, 2007a, b).

In the context of developing new measurement concepts, it is important not to overlook innovative measurement procedures, many of which capitalize on new technologies. Technology-based assessment (TBA) are widely used in educational settings in the United States and some European countries (Hartig, Kröhne, & Jurecka, 2007). In Germany, however, TBA is used primarily in psychological research, and is not yet well established in educational practice. During the 1990s, the use of TBA in psychological and educational competence assessment became increasingly widespread. This kind of assessment has numerous advantages: It allows complex stimuli and response formats, interactive testing procedures, real-time assessment of cognitive processes (Wirth, 2004; Wirth & Klieme, 2003), and automatized analysis and feedback procedures (Chung, O'Neil, & Baker, 2008; Ordinate, 2004; Reeff, 2007). In addition, TBA offers the possibility of computerized adaptive testing (CAT; e.g., van der Linden, 2005), in which the items presented are selected to fit the individual ability level of the test-taker.

Computer-adaptive testing allows for dynamic testing. The concept of dynamic testing is already well-established in the domain of intelligence assessment, but it can also be transferred to other performance domains. Dynamic testing focuses on the potential for intellectual development and is applicable in areas where the assessment of the status quo of a certain competence is unsatisfactory.

Furthermore, technology-based assessment permits the construction of complex and interactive stimuli that would be very costly or impossible to realize without the use of computers. It, thus, affords the possibility to empirically assess new competence domains that were not assessable with traditional measurement procedures. Because TBA allows the *simulation of complex and dynamic situations*, assessment designs can be more valid with respect to the demands of real-life, complex situations (Drasgow, 2002). For example, virtual patients can be used for competence assessment in medical education (Jung, Ahad, & Weber, 2005). Virtual environments can be used to examine individual navigation skills or, in networked environments, interactions between different individuals (Frey, Hartig, Ketzel, Zinkernagel & Moosbrugger, 2007). The PISA 2009 study includes a new component assessing the competence to read electronic texts (OECD, 2007a), which can be conceptually distinguished from the competence to read printed text. Thanks to technology-based procedures, the real-life situation of reading a hypertext can be simulated for these assessments. Test-takers' behavior can be recorded in log files, allowing their navigation within electronic tests to be analyzed and process indicators to be constructed that supplement responses to the test items (e.g., Wirth, 2004; Wirth & Leuter, 2008; Künsting, Thillmann, Wirth, Leutner, & Fischer, 2008).

In addition, technology-based testing allows parsimonious assessment and data administration. In particular, computers networked in local area networks or via the internet make it possible to test and provide feedback independently of time and of the test-takers' location, and to simultaneously administer tests to large samples (ETS, 2005; Groot, de Sonneville, & Stins, 2004; Jude & Wirth, 2007).

The new possibilities afforded by TBA have been used in numerous contexts. However, many of these applications are driven by the rapid development of computer technology rather than by well-founded theories. Much empirical and theoretical work is still needed to link complex computerized measurement procedures to cognitive and psychometric models.

Area 4: Reception and Usage of Assessment Results

The success of many educational decisions and interventions hinges on accurate assessments of learners' baseline competencies and learning outcomes. Assessments may have different practical goals: On an individual level, they allow educators to select appropriate interventions for individual cases (i.e., to further individual learning). The results of individual assessments may also inform decisions on the admittance to secondary tracks or to higher education. In contrast, assessment programs that are designed to report achievement on an aggregated level serve to evaluate educational programs, institutions, or systems, as well as to inform decision makers on the administrative and political levels. As Pellegrino et al. (2001)

concluded, "one size of assessment does not fit all" (p. 222). Depending on the goal of the assessment, different measurement instruments are needed and different research questions arise. Thus a continuing challenge facing researchers in this area is thus to determine which models, measurement rules, and measurement procedures provide the appropriate information for various goals of assessment.

Assessment to further individual learning can be regarded as *formative evaluation* on an individual level. It should allow precise conclusions to be drawn about individual learning processes and learners' strengths and weaknesses with respect to specific curricular units. These conclusions can help to support individual instruction and learning, and ideally offer considerable potential to enhance teaching. Teachers make observations of students' understanding and performance in a variety of ways: In classroom dialog, homework assignments, and formal tests (Pellegrino et al., 2001). These procedures should permit diagnosis on an individual level, in terms of understanding students' individual solution paths, misconceptions, etc. (Seger, Dochy, & Cascallar, 2003; Wilson, 2008). Appropriate individual feedback is crucial to support the subsequent learning process. A number of research questions arise in this context: What kind of diagnostic information is best understood by students, and what kind by teachers? How well can teachers evaluate individual learning processes? What factors influence teachers' grading decisions? What models of competence do teachers rely on – implicitly or explicitly? How well founded and how helpful is the feedback provided by the teacher to the individual students?

The assessment of individual achievement may also entail the *summative evaluation* of an individual's competencies. These evaluations help to determine whether a student has attained a certain level of competence after completing a particular phase of education (e.g., in end-of-unit tests or the letter grades assigned at the end of a course; Pellegrino et al., 2001). These performance measurements are often *high stakes*, meaning that their outcomes have significant consequences. Students who fail to attain certain standards (e.g., passing their final school exams) may be refused access to the next level. An important question for research on assessment in this field is how tests can be constructed to reflect educational goals and how results can be interpreted with reference to curricula (e.g., Cizek, Bunch, & Koons, 2004; Haertel & Lorié, 2004; Klauer & Leutner, 2007; Klieme et al., 2003). A related research question is how the content of educational assessments affects the methods and content of instruction ("washback effects"; e.g., Cheng et al, 2004; Cizek, 2001; Fuhrman & Elmore, 2004; Nichols, Glass, & Berliner, 2006; as well as Pellegrino et al. 2001, p. 212 ff). Individual data aggregated on the classroom level can be used by teachers to evaluate their own instruction and to identify their students' specific instructional needs (e.g., Leutner et al., 2007). Teachers need detailed, contextualized information about their students' learning progress to efficiently adjust their instructional focus – like the information needed to inform decisions on an individual level. However, there is a marked gap between the information teachers need and the information they are given.

Huff and Goodman (2007) report that although most teachers in the United States receive assessment results from state mandated or commercial large-scale

assessments, 20–30% of them almost never use these results to reflect on their instruction. Moreover, 30–38% of the teachers state that the diagnostic information provided by large-scale assessments is not detailed enough to use.

The results of competence assessments on an *aggregated level* provide information about classrooms and schools. Aggregated data usually serves evaluation purposes and supports the quality development of educational processes. School principals can use classroom-level data as a basis for evaluating teachers' performance and as indicators for the need for professional development. Educational administrations can use the results of competence assessments aggregated at the school level to inform budget decisions concerning individual schools. Aggregated data from educational assessments can also be used to guide and control whole education systems on the political level, that is, for purposes of *system monitoring* (Leutner et al., 2007). Policy makers can use information aggregated on a district or country level to gauge the effectiveness of educational systems and to make decisions about measures to improve their effectiveness. Of course, the information required for such decisions differs markedly from that required by students and teachers to further learning processes. Aggregated information about the overall achievement concerning curricular targets is more functional than is detailed contextualized information. The *proficiency levels or levels of competence* used to report the results of large-scale assessments (e.g., Adams, 2005; Adams & Wu, 2002) constitute one well-known technique for facilitating the understanding of assessment results among administrators, policy makers, and the public. However, there is little empirical research on which kind of information is actually best suited to guide administrative and political decisions.

In Germany, standardized assessments of students' competencies were previously a relatively rare occurrence, but this is now changing. For example, there is a marked trend towards the use of standardized achievement tests to control admittance to higher education (e.g., Amelang & Funke, 2005; Gold & Souvignier, 2005; Köller, 2004). Several of the newly introduced Bachelors and Masters programs have been designed to communicate specific competencies that are generally verifiable. More importantly, educational standards have been developed to describe the goals of primary and secondary education (Klieme et al., 2003). Based on these standards, a system has been developed to assess students' competencies. New evaluation agencies have been founded as part of these ongoing educational reforms. These agencies assess learning outcomes on both the classroom and the school level, and provide information for policy makers.

DESCRIPTION OF THE DFG PRIORITY PROGRAM ON THE ASSESSMENT OF COMPETENCIES

As outlined above, the accurate empirical assessment of competencies is essential for the enhancement of educational processes and the development of educational systems. Yet devising and implementing such assessments entails numerous theoretical and methodological challenges.

To facilitate this task, the German Research Foundation, DFG, has funded the priority program "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (Klieme & Leutner, 2006). The program, which is scheduled to run for six years, involves a network of currently 20 individual research projects covering different areas of competence assessment.

The program unites experts in different domains of study with cognitive psychologists and experts in educational measurement. Its objective is to develop theoretically and empirically grounded models of competencies as a basis for constructing valid and fair instruments for the assessment of student competencies in terms of both individual learning outcomes (thus promoting individual learning processes) and the output of educational institutions and systems on an aggregated level. Research dealing with the reception of assessment results and their application in pedagogical decisions rounds off the research program. The program extends ongoing research on existing models of competence (e.g., the competence levels used in large-scale assessments) and initiates research in qualitatively new areas (e.g., development of competence models in new content areas; application of innovative psychometric models).

Based on our working definition of competence, the DFG priority program defines competencies as domain-specific cognitive dispositions that are required to successfully cope with certain situations or tasks, and that are acquired by learning processes. Thus, a specific competence is understood as the potential to meet the cognitive demands of a certain domain of learning, or vocational demands. In line with the four areas of research presented above, the program has four specific objectives:

1. To develop cognitive competence models that reflect the contextualized and domain-specific nature of competencies and that enable the theory-based development of instruments for their assessment. These models will focus on cognitive processes, characterize different levels of competence, and describe and explain the quantitative and qualitative development of competence. Ten of the program's projects deal with questions related to this area.

2. To develop and empirically examine appropriate psychometric models on the basis of these theoretical models. The psychometric models will take into account the contextualized character of competencies and to incorporate interindividual differences in underlying abilities, as well as situational demands of performance in complex tasks. Four projects have their main focus in this research area.

3. To develop instruments for the empirical assessment of specific competencies. These instruments permit the empirical examination of the theoretical and psychometric models of competencies, and are essential for basic research on these models. They are also required for basic research that needs measures of competence as outcome variables (e.g., research on the prerequisites of competence development or on educational processes). In applied contexts, these instruments allow individual and institutional learning outcomes to be monitored and provide feedback for learners and educators. Five of the program's projects focus on the development of measurement concepts and instruments.

4. To examine how model-based assessments of competencies are used to inform educational decisions on the individual level, as well as political and administrative decisions concerning educational systems and institutions at the aggregated level. It is crucial to know how different stakeholders in educational processes and decision making understand and use the information provided by empirical assessments of competencies, depending on the underlying theoretical and psychometric models and the measurement methods employed, and how this information impacts the subsequent teaching and learning process. One project addresses these aspects.

Although the program's projects are all assigned to one of the four research areas, most of them are not restricted to a single area. As outlined in this article, research on competency assessment is interdependent, and different areas build on each other. Theoretical models are needed as a starting point; psychometric models translate theoretical models into rules of measurement. Empirical measurement procedures apply the theoretical and psychometric models of competencies and provide data that can be used to inform educational processes and decision making. It is in the nature of the field of research that the majority of the individual projects will initially focus on theoretical and psychometric models. Most projects will then move on to the development of measurement instruments and, in some cases, to research on the reception of the assessment results.

The majority of the projects relate to subjects in primary and secondary education (e.g., mathematics or reading), which is not surprising, given that there has been more research on competencies and competence assessment in these areas and that the corresponding theories are already better developed. However, some projects address competencies in vocational domains; for example, there is a particular focus on well-defined aspects of teachers' professional competence, such as diagnostic competence and specific aspects of pedagogical content knowledge (Weinert, Helmke, & Schrader, 1992). Other projects examine the competencies required in specific non-professional areas of life, such as health or attitudes to environment protection. Altogether the projects cover a wide range of competence domains.

NOTES

[1] German Institute for International Educational Research

[2] Duisburg-Essen University

[3] Reprint of Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current issues in competence modeling and assessment. Zeitschrift für Psychologie/ Journal of Psychology, 216, 61–73 (with permission of Hogrefe & Huber Publishers). The preparation of this paper was supported by grants KL 1057/9-1 and DL 645/11-1 from the German Research Foundation (DFG) in the Priority Program "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293). The paper is an extension of an article by Klieme & Leutner (2006).

REFERENCES

Adams, R.. (2005). *PISA 2003 technical report.* Paris: OECD.

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Adams, R., Wilson, M., & Wu, M.L. (1997). Multilevel item response modelling: An approach to errors in variables regression, *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Adams, R., & Wu, M.. (2002). *PISA 2000 technical report.* Paris: OECD.

Alderson, C. (2005). Diagnosing foreign language proficiency: the interface between learning and assessment. London: Continuum.

Alderson, C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2005). *The Dutch CEF grid reading/ listening (revised internet version available for test development and analysis).* http://www.lancs.ac.uk/fss/projects/grid/.

Amelang, M., & Funke, J. (2005). Entwicklung und Implementierung eines kombinierten Beratungs- und Auswahlverfahrens für die wichtigsten Studiengänge an der Universität Heidelberg [Development and implementation of a combinded instrument for counseling and selection for the most important courses at the University of Heidelberg]. *Psychologische Rundschau, 56,* 135–137.

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*, 363–383.

Baumert, J., Stanat, P., & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie [Subject, theoretical background and implementation of the study]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 15–68). Opladen: Leske & Budrich.

Beck, B., & Klieme, E.. (2007). *Sprachliche Kompetenzen – Konzepte und Messung* [Language competencies – concepts and measurement]. Weinheim: Beltz.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, H.C. (2004). Mathematische Kompetenz [Mathematical literacy]. In PISA-Konsortium Deutschland (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (pp. 47–92). Münster: Waxmann.

Bybee, R.W. (1997). Toward an understanding of scientific literacy. In W. Gräber & C. Bolte (Eds.), *Scientific Literacy, an international Symposium* (pp. 37–68). Kiel: IPN.

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: a reply to Kobayashi, 2002. *Language Testing, 21,* 228–234.

Cheng, L., Watanabe, Y., & Curtis, A.. (2004). *Washback in language testing. research contexts and methods.* Mahwah: Lawrence Erlbaum.

Chung, G.K.W.K., O´Neil, H.F., & Baker, E.L. (2008). Computer-based assessments to support distance learning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe & Huber.

Cizek, G.J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement, Issues and Practice, 20(4)*, 19–28.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). A NCME Instructional Module on Setting Performance Standards: Contemporary Methods. E*ducational Measurement: Issues and Practice, 23*, 31–50.

Connell, M.W., Sheridan, K., & Gardner, H. (2003). On abilities and domains. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 126–155). Cambridge: Cambridge University Press.

Csapó, B. (2004). Knowledge and competencies. In J. Letschert (Ed.), *The integrated person. How curriculum development relates to new competencies* (pp. 35–49). Enschede: CIDREE/SLO.

von Davier, M. (2005). A general diagnostic model applied to language testing data. ETS Research Report 0x-2005.

DiBello, L.V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44,* 285–291.

DiSessa, A. (2006). A history of conceptual change research. In K. Sawyer (Ed.), *The Cambridge Handbook of the learning Sciences* (pp. 265–281). Cambridge: Cambridge University Press.

Dossey, J., Hartig, J., Klieme, E., & Wu, M. (2004). Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003. Paris: OECD Publications.

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive testing. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Wards (Eds.), *Computer-based testing. Building the foundation for future assessments* (pp. 1–35). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S.E. (1983). Construct validity: construct representation vs. nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S.E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61,* 50–55.

ETS (2005). TOEFL iBT at a glance. Retrieved September 25, 2005, from http://www.ets.org/Media/Test/TOEFL/pdf/TOEFL_at_a_Glance.pdf.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Frensch, P. A., Haider, H., Rünger, D., Neugebauer, U., Voigt, S., & Werg, D. (2003). The route from implicit learning to awareness of what has been learned. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 335–366). New York: John Benjamins Publishing Company.

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A., & Moosbrugger, H. (2007). Usability and internal validity of a modification of the computer game Quake III Arena® for the use in psychological experiments. *Computers in Human Behavior, 23,* 2026–2039.

Fuhrmann, S.H., & Elmore, R.F.. (2004). *Redesigning accountability systems for education.* New York: Teachers College Press.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2007). Using the attribute hierarchy method to make inferences about examinees' cognitive skills. In M.J. Gierl & J.P. Leighton (Eds.), *Cognitive diagnostic assessment for education* (pp. 242–274). Cambridge: Cambridge University Press.

Gogolin, I. (2002). Linguistic and cultural diversity in Europe: a challenge for educational research and practice. *European Educational Research Journal, 1*, 123–138.

Gold, A., & Souvignier, E. (2005) Prognose der Studierfaehigkeit. Ergebnisse aus Laengsschnittanalysen [Prediction of college graduation. Results from longitudinal studies]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37*, 214–222.

Groot, A.S., de Sonneville, M.J., & Stins, J.F. (2004). Familial influences on sustained attention and inhibition in preschoolers. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 45,* 306–314.

Haertel, E.H., & Lorié, W.A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary research and perspectives, 2*, 61–103.

Haider, H., & Frensch, P.A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30*, 304–337.

Haider, H., & Frensch, P.A. (1997). Lernmechanismen des kognitiven Fertigkeitserwerbs [Learning mechanisms in cognitive skill acquisition]. *Zeitschrift für Experimentelle Psychologie, 44*, 521–560.

Haider, H., & Frensch, P.A. (2002). Why individual learning does not follow the power law of practice but aggregated learning does: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 392–406.

Hartig, J., & Frey, A. (2005). *Application of different explanatory item response models for model based proficiency scaling.* Paper presented at the 70th Annual Meeting of the Psychometric Society in Tilburg, July 5–8, 2005.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within- and between-item multidimensionality. *Zeitschrift für Psychologie / Journal of Psychology, 216,* 88–100.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik [Competence and competence diagnosis]. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 127–143). Berlin: Springer.

Hartig, J., & Klieme, E. (2007). *From theoretical notions of competence to adequate psychometric models*. Paper presented at the 12th Biennial EARLI Conference, Budapest, August 28-September 1, 2007.

Hartig, J., Kröhne, U., & Jurecka, A. (2007). Anforderungen an Computer- und Netzwerkbasiertes Assessment [Requirements for computer- and network based assessments]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 57–67). Berlin: Federal Ministry of Education and Research (available at http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Hasselhorn, M., & Grube, D. (2003). The phonological similarity effect on memory span in children: Does it depend on age, speech rate, and articulatory suppression? *International Journal of Behavioral Development*, *27*, 145–152.

Huff, K., & Goodmann, D.P. (2007). The demand for cognitive diagnostic assessment. In M.J. Gierl & J.P. Leighton (Eds.), *Cognitive diagnostic assessment for education* (pp. 19–61). Cambridge: Cambridge University Press.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck, & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Jude, N., & Wirth, J. (2007). Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen [New opportunities of technology based assessment of competencies]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 81–91). Berlin: Federal Ministry of Education and Research (available at URL: http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Jung, B., Ahad, A., & Weber, M. (2005). The Affective Virtual Patient: An e-learning tool for social interaction training within medical field (*Proceeding TESI 2005 – Training Education & Education International Conference).* Kent, UK: Nexus Media (available at http://isnm.de/aahad/Downloads/AVP_TESI.pdf).

Klauer, K.J. (1978). Perspektiven pädagogischer Diagnostik [Perspectives of educational assessment at the individual level]. In K.J. Klauer (Ed.), *Handbuch der Pädagogischen Diagnostik* (pp. 3–4). Düsseldorf: Schwann.

Klauer, K.J. (1987). *Kriteriumsorientierte Tests* [Criterion-referenced tests]. Göttingen: Hogrefe.

Klauer, K.J., & Leutner, D. (2007). *Lehren und Lernen. Einführung in die Instruktionspsychologie* [Teaching and learning. Introduction to instructional psychology]. Weinheim: Beltz-PVU.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. Reiss, K., Riquarts, K., Rost, J., Tenorth, H.E., & Vollmer, J. (2003). *The development of national educational standards. An expertise.* Berlin: Federal Ministry of Education and Research (avaliable at http://www.bmbf.de/pub/the_development_of_national_educationel_standards.pdf).

Klieme, E., Eichler, W., Helmke, A., Lehmann, R. H., Nold, G., Rolff, H.-G., Schröder, K., Thomé, G., & Willenberg, H. (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* [Instruction and competence development in German and English. Results of the DESI study]. Weinheim: Beltz.

Klieme, E., Funke, J., Leutner, D., Reimann, P., & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz? Konzeption und erste Resultate aus einer Schulleistungsstudie [Problem solving as cross-curricular competence? Concepts and first results from an educational assessment]. *Zeitschrift für Pädagogik*, *47,* 179–200.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms bei der DFG [Competence models for assessing individual learning outcomes and evaluating educational processes. Description of a new priority program of the German Research Foundation, DFG]. *Zeitschrift für Pädagogik, 52,* 876–903.

Klieme, E., Maag-Merki, K., & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen [The concept and relevance of competencies in education]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 5–16). Berlin: Federal Ministry of Education and Research (available at http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse [Mathematical literacy: assessment framework and results]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 139–190). Opladen: Leske & Budrich.

Koeller, O. (2004). *Konsequenzen von Leistungsgruppierungen* [Consequences of homogenous groups with regard to school performance]. Muenster: Waxmann.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing 19,* 193–220.

Kröner, S., Plass, J.L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33,* 347–368.

Künsting, J., Thillmann, H., Wirth, J., Fischer, H.E., & Leutner, D. (2008). Strategisches Experimentieren im naturwissenschaftlichen Unterricht [Strategic experimentation in science lessons]. *Psychologie in Erziehung und Unterricht, 55*, 1–15.

Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior, 18*, 685–697.

Leutner, D., & Plass, J. L. (1998). Measuring learning styles with questionnaires versus direct observation of preferential choice behavior: Development of the Visualizer/Verbalizer Behavior Observation Scale (VV-BOS). *Computers in Human Behavior, 14,* 543–557.

Leutner, D., Fleischer, J., Spoden, C., & Wirth, J. (2007). Landesweite Lernstandserhebung zwischen Bildungsmonitoring und Individualdiagnostik [State-wide standardized assessments of learning between educational monitoring and individual diagnostics]. *Zeitschrift für Erziehungswissenschaft, Sonderheft, 8,* 149–167.

van der Linden, W. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42,* 283–302.

van der Linden, W., & Hambleton, R.K. (1996). Item response theory: brief history, common models, and extensions. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item-response theory* (pp. 1–28). Berlin: Springer.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187–212.

McClelland, D.C. (1973). Testing for competence rather than for „intelligence". *American Psychologist, 28*, 1–14.

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99–114.

Nichols, S.L., Glass, G.V., & Berliner, D.C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14* (available at http://epaa.asu.edu/epaa/v14n1/).

Nold, G. (2003). DESI – a language assessment project in Germany and the pros and cons of large-scale testing. *Empirische Pädagogik, 17,* 368–379.

Nold, G., & Rossa, H. (2007a). Hörverstehen [Listening comprehension]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 178–196). Weinheim: Beltz.

Nold, G., & Rossa, H. (2007b). Leseverstehen [Reading comprehension]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 197–211). Weinheim: Beltz.

Van Den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (pp. 167–187). New York: Springer.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence – their correlation and their relation: A comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 61–65.

OECD (2007a). *PISA – Programme for International Student Assessment* (available at http://www.oecd.org/dataoecd/51/27/37474503.pdf).

OECD (2007b). *PISA 2006 Science competencies for tomorrow's world* (volume 1: analysis). Paris: OECD.

Ordinate (2004). *SET-10 test description & validation summary*. Menlo Park, CA: Ordinate.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment*. Washington, DC: National Academic Press.

Plass, J.L., Chun, D., Mayer, R.E., & Leutner, D. (1998). Supporting visualizer and verbalizer learning preferences in a second language multimedia learning environment. *Journal of Educational Psychology, 90*, 25–36.

Prenzel, M., & Allolio-Näcke, L.. (2006). *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* [Research on educational quality of schools. Final report of the DFG priority program]. Münster: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R.. (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006. Results of the third international study]. Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U.. (2004). *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* [PISA 2003: Educational outcomes of German students – results of the second international study]. Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U.. (2005). *PISA 2003: Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?* [The second comparison of the German laender – What do students know?]. Münster: Waxmann.

Prenzel, M., Rost, J., Senkbeil, M., Häußler, P., & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse [Scientific literacy: assessment framework and results]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 192–248). Opladen: Leske & Budrich.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C.H., & Hammann, M. (2007). Naturwissenschaftliche Kompetenzen im internationalen Vergleich [Science competencies in international comparison]. In PISA-Konsortium Deutschland (Eds.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 63–105). Münster: Waxmann.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.

Reeff, J.-P. (2007). Technische Lösungen für ein computer- und internetbasiertes Assessment-System [Technical solutions for computer and internet based assessment systems]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 81–91). Berlin: Federal Ministry of Education and Research (available at http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Rychen, D.S., & Salganik, L.H.. (2001). *Defining and selecting key competencies.* Seattle: Hogrefe & Huber Publishers.

Rychen, D.S., & Salganik, L.H.. (2003). *Key competencies for a successful life and a well-functioning society.* Washington: Hogrefe & Huber Publishers.

Schneider, W., Lockl, K., & Fernandez, O. (2005). Interrelationships among theory of mind, executive control, language development, and working memory in young children: A longitudinal analysis. In W. Schneider, R. Schumann-Hengsteler & B. Sodian (Eds.), *Young children's cognitive development: Interrelationships among executive functioning, working memory, verbal ability, and theory of mind* (pp. 259–284). Mahwah, NJ: Lawrence Erlbaum.

Schnotz, W., Eckhardt, A., Molz, M., Niegemann, H., Hochscheid-Mauel, D., & Hessel, S. (2004). Deconstructing instructional design models towards an integrative conceptual framework for

instructional design research. In H. Niegemann, R. Brünken & D. Leutner (Eds.), *Instructional design and multimedia learning* (pp. 71–89). Münster: Waxmann.

Schnotz, W., Vosniadou, S., & Carretero, M.. (1999). *New perspectives on conceptual change*. Oxford: Elsevier.

Segers, M., Dochy, F., & Cascallar, E.. (2003). *Optimising new modes of assessment: in search of quality and standards*. Dordrecht: Kluwer.

Simonton, K. (2003). Expertise, competence, and creative ability: The perplexing complexities. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 213–239). Cambridge: Cambridge University Press.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31,* 15–21.

Spiel, C., & Glück, J. (in press). A model based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe.

Sternberg, R.J., & Grigorenko, E.. (2003). *The psychology of abilities, competencies, and expertise.* New York: Cambridge University Press.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*, 313–324.

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction, 15,* 317–419.

Walker, C.M., & Beratvas, S.N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255–275.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29,* 126–149.

Weinert, F.E. (1999). *Konzepte der Kompetenz* [Concepts of competence]. Paris: OECD.

Weinert, F.E. (2001). Concept of competence: a conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key compe-tencies* (pp. 45–65). Seattle: Hogrefe & Huber Publishers.

Weinert, F. E., Helmke, A., & Schrader, F.-W. (1992). Research on the model teacher and the teaching model: Theoretical contradiction or conglutination? In F. Oser, A. Dick & J.L. Patry (Eds.), *Effective and responsible teaching: The new synthesis* (pp. 249–260). San Francisco: Jossey-Bass Publishers.

Weinert, F.E., & Schneider, W.. (1995). *Memory performance and competencies: Issues in growth and development.* Hillsdale, NJ: Lawrence Erlbaum.

Whitely, S.E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement, 18,* 67–84.

Wilhelm, O., & Engle, R.. (2005). *Understanding and measuring intelligence*. London: Sage.

Wilson, M. (2008). Cognitive Diagnosis using Item Response Models. *Zeitschrift für Psychologie/ Journal of Psychology, 216, 73–87.*

Wilson, M., de Boeck, P., & Carstensen, C. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe & Huber.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson, (Ed.), *Towards coherence between classroom assessment and accountability (*103rd Yearbook of the National Society for the Study of Education, Part II). Chicago: University of Chicago Press.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181–208.

Wilson, M., & DeBoek, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models. A generalized linear and nonlinear approach* (pp. 43–74). New York: Springer.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: state of the art and future prospects.* Göttingen: Hogrefe & Huber.

Wirth, J. (2004). *Selbstregulation von Lernprozessen* [Self-regulation in learning processes]. Münster: Waxmann.

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem-solving competence. *Assessment in Education: Principles, Policy & Practice, 10,* 329–345.

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence. Implications of theoretical models for assessment methods. *Zeitschrift für Psychologie / Journal of Psychology, 216,* 101–109.

Detlev Leutner
Department of Instructional Psychology,
Duisburg-Essen University, Germany

Karoline Koeppen, Johannes Hartig & Eckhard Klieme
German Institute for International Educational Research

PART 3

LONG-TERM OUTCOMES

CHRISTIANE SPIEL, BARBARA SCHOBER AND RALPH REIMANN

MODELING AND MEASUREMENT OF COMPETENCIES IN HIGHER EDUCATION – THE CONTRIBUTION OF SCIENTIFIC EVALUATION

This chapter focuses on the contribution scientific evaluation can make to the modeling and measurement of competencies in higher education. The chapter is divided into three parts. In the first section, the basics of scientific evaluation are briefly described with a specific focus on relevant aspects of program evaluation. The second part of the chapter presents the evaluation of a curriculum as an example of measuring competencies in higher education – the profile of measured competencies is strongly based on specific literature and empirical data. Typical challenges evaluators face when investigating the effectiveness of a curriculum are described as well as strategies to overcome them. In the third part we present recommendations for modeling and measurement of competencies in higher education based on the insights of scientific evaluation.

THE BASICS OF SCIENTIFIC EVALUATION

In a broad sense, scientific evaluation (or evaluation research) is concerned with the analysis of the effectiveness (goal achievement) and the efficiency (relation between costs and benefits) of programs, projects, and organizations, etc. (see e.g. Spiel, 2001; Spiel, Gradinger & Lüftenegger, 2010). Whereas commonsense evaluation has a very long history, evaluation which relies on scientific methods is a young discipline but has grown massively in recent years (for introductory texts see e.g. Berk & Rossi, 1998; Fink, 1995; Pawson & Tilley, 1997; Rossi & Freeman, 1993). In this chapter, we focus on program evaluation, as curricula in higher education can be seen as programs. The statements are partly based on Spiel et al. (2010) and on Spiel, Lüftenegger, Gradinger and Reimann (2010). In the following, the terms evaluation, program evaluation, and scientific evaluation are used synonymously.

Introductory texts on evaluation describe programs as systematic activities that are provided on a continuing basis to achieve pre-planned purposes (e.g. Fink, 1995; Joint Committee on Standards for Educational Evaluation, JCSEE, 1994).

> *Program evaluation* is the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future programming (Patton, 1996, p. 23).

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 195–206.

This implies that program evaluation capitalizes on existing theory and empirical generalizations from the social sciences, and is, therefore, after all, applied research (e.g. Pawson & Tilley, 1997; Berk & Rossi, 1998).

Evaluation goals are defined by the client, that is, the individual, group, or organization that commissions the evaluator(s), e.g. policy-makers. Often, however, programs and clients do not have clear and consistent goals either for their programs or for the evaluation of effectiveness. Therefore, to meet evaluation standards program evaluation should be conducted by professional evaluators who are experts in social sciences, use a broad repertoire of concepts and methods, and also have expertise in project and staff management, and at least basic knowledge of the evaluated domain.

In any case, for all evaluations the following four central questions have to be put (Spiel et al., 2010):

(1) What are the goals of the program?
(2) How should these goals be achieved?
(3) How could program effectiveness (goal attainment) be recognized?
(4) Are the needed resources available?

To answer these four questions a workshop on goal explication is usually organized (see Spiel et al., 2010). Using a wide range of moderation techniques the clients of evaluations are supported to describe their evaluation goals exactly and to prioritize them (Atria, Reimann & Spiel, 2006; Wottawa & Thierau, 1998; Question 1). Very often, the congruence between goals and activities is rather weak. Evaluators help the clients to identify how the intended activities might work and how they can contribute to reach the intended goals (Question 2). This also includes systematic discussion whether the resources needed (money, staff, but also the agreement of the stakeholders to participate in the program) are available (Question 4). Answering the third question is usually very difficult for the clients of evaluation. They have to identify measurable indicators for effectiveness that are labeled "operationalization" in scientific research (Atria et al., 2006; Bortz & Döring, 2006). Consequently, answering the third question also includes the definition of the goals of the evaluation. Again, evaluators use moderation techniques to support the process of identifying the indicators of effectiveness. If possible, criteria for effectiveness are defined (Atria et al., 2006; Fink, 1995).

Evaluation can and should be done in all phases of a program. In the following the different concepts and assignment criteria are briefly described using the implementation of the Bologna concept (bachelor and master program) as an example.

Baseline data collected before the start of the program are used to describe the current situation and, e.g., the need for an intervention. In addition, baseline data are used to monitor and explain changes. For example, in what way and to what degree does the current curriculum (before the implementation of the Bologna concept, usually diploma studies) achieve its goals? What are the current curriculum's concrete goals and how are standards of effectiveness defined? What are the weaknesses and strengths of the current curriculum?

Before the start of the program a *prospective evaluation* can be applied to determine the program's potential for realization and impact, that is, the scope of its effects. In the case of the Bologna concept, the prospective evaluation assesses the new curricula's (bachelor and master) potential for realization.

In *formative evaluation,* interim data are collected after the start of a program but before its conclusion. It is the purpose of formative evaluation to describe the progress of the program and, if necessary, to modify and optimize the program design. In the case of a bachelor or master program the curriculum concept might be revised.

A *process evaluation* is concerned with the extent to which planned activities are executed and therefore is nearly always useful. In the case of a new curriculum the quality of its implementation is evaluated.

Outcome evaluation deals with the question whether programs achieve their goals. Consequently, comparison with baseline data is highly relevant. In the case of a new curriculum its effectiveness (goal attainment) is investigated in comparison with the effectiveness of the old (diploma study) curriculum.

The entire impact of a program, e.g. the extent of its influence in other settings, is very difficult, even impossible, to assess. Here, *impact evaluations* are applied, that is, historical reviews of programs that are performed after the programs have been in operation for some period of time. In the case of the implementation of the Bologna concept the impact of the curricula depends on the specific discipline. In the case of medical education an improvement in the quality of the health system might be expected. The evaluation of such an improvement might be extremely difficult, however.

Programs can be evaluated not only in different phases but also on different levels. Kirkpatrick (1998; see also Kirkpatrick & Kirkpatrick, 2005) defined four sequential levels. Each level is important and has an impact on the next levels. The evaluation process becomes more difficult and time-consuming the higher the evaluation level, but it also provides more valuable information. In the following the four levels are briefly described (Kirkpatrick, 1998; Kirkpatrick & Kirkpatrick, 2005; see also Spiel et al., 2010).

Level 1 – reactions: Evaluation on this level measures how the participants in the program react to it. It is therefore a measure of acceptance and satisfaction. Students' ratings of courses are a typical example of the reaction level in higher education.

Level 2 – learning: Learning can be defined as the extent to which participants change attitudes, improve knowledge, and/or increase skills and competencies as a result of attending the program. That means, for example, how much students have learned in a course which could be measured by e.g. grades or tests applied before and after the course.

Level 3 – behavior: Evaluation on this level focuses on the extent to which change in behavior has occurred because the participant attended the program, i.e. how the learning can be transferred into everyday life, and, in the case of university courses, how students apply what they have learned in their working and later business behavior.

Level 4 – results: At this stage the final results are examined. These final results can include e.g. increased production, improved quality, increased sales, decreased costs, or higher profit. In the case of university courses this level is very difficult to evaluate. On the one hand, courses are part of a curriculum and therefore their specific impact might be impossible to identify. On the other hand, students usually disperse after graduating and therefore the unit of measurement might be very difficult or maybe impossible to identify. Therefore, final results are very seldom evaluated. If they are, it is when programs are applied in organizations; for example, how the implementation of the Bologna concept improves the teaching quality and/or the student orientation at a specific university.

Because of the increasing number of evaluations and the high impact that decisions based on evaluations may have, guidelines for effective evaluations have been discussed since the early 1970s. The *Joint Committee on Standards for Educational Evaluation* (JCSEE, 1994) in cooperation with evaluation and research specialists compiled knowledge about program evaluation and established a systematic progress for developing, testing, and publishing evaluation standards. The Joint Committee defined an *evaluation standard* as a principle mutually agreed on by people engaged in the professional practice of evaluation, that, if met, will enhance the quality and fairness of an evaluation (JCSEE, 1994). The standards provide advice on how to judge the adequacy of evaluation activities, but they do not present specific criteria for such judgments. The reason is that there is no such thing as a routine evaluation and the standards encourage the use of a variety of evaluation methods.

Since 1990, in many European countries evaluation societies have been established which also define guidelines and standards of evaluation mostly based on the JCSEE standards, e.g. the Guiding Principles for Evaluators published by the American Evaluation Association (Shadish, Newmann, Scheirer & Wye, 1995), the evaluation standards of the Swiss evaluation society (Widmer, Landert, & Bachmann, 2000) and the standards published by the DeGEval – Gesellschaft für Evaluation which represents scientific evaluation in Germany and Austria (DeGEval, 2002). The 25 DeGEval standards are arranged around the four important attributes of sound and fair program evaluation: utility, feasibility, propriety, and accuracy (see also JCSEE, 1994; Spiel et al., 2010).

Utility Standards guide evaluations to be informative, timely, and influential. They require the stakeholders' needs to be borne in mind in every phase of the evaluation. *Feasibility Standards* call for realistic, prudent, diplomatic, and economic evaluations. They recognize that evaluations are usually conducted in natural settings. *Propriety Standards* are intended to facilitate protection of the rights of individuals affected by an evaluation. They urge evaluators to act lawfully, scrupulously, and ethically. *Accuracy Standards* are intended to insure that an evaluation will reveal and transmit accurate information about the program's merits. They determine whether an evaluation has produced sound information.

Sound evaluations providing significant results have to consider the different forms, levels and standards of evaluation, which is obviously a challenge in the context of measuring the imparting of competencies in higher education. The

following example describes typical problems with which evaluators in higher education are faced and the solution strategies applied.

EVALUATION OF CURRICULA AS AN EXAMPLE OF SCIENTIFIC EVALUATION IN HIGHER EDUCATION

It was the intention of the presented evaluation study to provide baseline data for the development of a new curriculum for medical education at the University of Graz, Austria (Schober, Spiel & Reimann, 2004). The evaluation study was conducted in cooperation with the Faculty of Medicine. The evaluation is used to illustrate typical challenges evaluators face when investigating a curriculum's effectiveness. We present the strategies applied to overcome these challenges and selected empirical results to illustrate application of the strategies. Furthermore, we discuss open problems (for details see Spiel, Schober & Reimann, 2006).

The evaluation of a curriculum's effectiveness has to analyze concrete educational goals and their realization (Gerrity & Mahaffy, 1998). The acquired competencies of the graduates are certainly central indicators for the quality of the curriculum. The definition and measurement of these competencies are challenges to evaluators, however. So far, there is not very much research in tertiary education which systematically combines theoretical models of competencies with adequate measurement models and respective assessment instruments (see e.g. Klieme & Leutner, 2006). In the field of medical education there is a particular lack of knowledge. Therefore, in the evaluation of medical education we proceeded as follows. We reviewed the relevant literature and questionnaires applied in previous studies (Merl, Csanyi, Petta, Lischka & Marz, 2000). Then we conducted 25 interviews with representatives of the four key stakeholder groups of Austrian medical education: students, graduates, university teachers, and clinical supervisors of graduates. On the basis of these interviews we defined seven different general areas of competencies: factual knowledge in natural sciences (e.g. the pathophysiology of metabolic disorders), factual knowledge in social sciences (e.g. the influence of economic factors such as poverty on pathogenesis), communication skills (e.g. conducting a medical counseling interview), skills required to perform routine medical jobs (e.g. applying initial diagnostic measures for a newly admitted patient), socio-ethical values (e.g. regarding human dignity in work with patients), coping skills (e.g. coping with psychological strains arising as a result of the medical job), and self-regulated information management (e.g. knowledge updated according to recent medical standards). According to the literature on the definition and measurement of competencies a competent learner (advanced student or graduate) should be able to combine and use factual knowledge, procedural knowledge, skills, and attitudes as well as self-regulatory abilities, e.g. meta-cognitive strategies (Klieme, 2004). To measure these competencies we adapted existing questionnaires for our specific concerns (Clack, 1994; Hill, Rolfe, Pearson & Heathcote, 1998) and developed new items in accordance with Sonneck's (1994) recommendations on the goals of medical education and with the findings of our interviews.

A further challenge evaluators of curricula in higher education have to cope with is the definition of the frame of reference for the evaluation. To obtain information regarding teaching effectiveness student ratings of courses are predominantly used as parameters for good teaching (Marsh, 1982; Schweer, 2001). Admittedly, different data sources and therefore different views (multiple ratings) are taken into account (cf. Hewson, Copeland & Fishleder, 2001; Rolfe, Andren, Pearson, Hensley & Gordon, 1995). Yet often different views are not identical (see e.g. Calhoun, ten Haken & Woolliscroft, 1990). This obviously flags up that different groups partially refer to different criteria. To obtain comprehensive information about the effectiveness of curricula we argue for the systematic consideration of at least two reference bases: the *learners'* and the *teachers'* views. Whereas learners' *self-ratings* provide information about the subjective assessments of acquired competencies, *external ratings* by teachers provide information on how learners use and apply these competencies and how they are reflected in their behavior. To obtain information on self-perception and the external-expert view in the present evaluation both groups – learners and teachers – were asked to evaluate the seven areas of expertise described above. Whereas the teachers were asked to assess to what degree the learners actually possessed the competencies (external ratings), learners were asked to assess to what extent they believed they possessed them (self-ratings).

As stated before, effective curriculum evaluation has to analyze explicitly both the concrete educational goals and their realization. Therefore two questions have to be asked: "What should be imparted?" and "What is really imparted?" Whereas the realization corresponds to the outcome, or what is imparted by the curriculum, the educational goals describe the ideal situation or what students should learn at university. In the evaluation study of medical education we wanted teachers (experts) to evaluate to what degree medical education should impart the seven relevant areas of expertise (ideal). In addition, we asked the learners to evaluate to what extent the present educational program did impart these areas of expertise (real).

The definition of the relevant sample for the evaluation of a curriculum is an additional challenge for evaluators of curricula in higher education. In particular, in defining the sample it has to be taken into account that not only the short term-effects of a curriculum are relevant. In the present evaluation we defined the participants according to their position in medical education: students and university teachers from different teaching fields. In addition, to investigate not only the short-term effectiveness of the curriculum we collected data from graduates and their supervisors (cf. Morrison, 2003). In sum, we investigated four subsamples: (1) advanced students in medical education at university, (2) graduates who are presently completing the obligatory full-time internship in various hospitals, (3) university teachers of the students, (4) clinical supervisors of the graduates. That means two learner groups – the students just before the final exams and the graduates after it – and the two respective teacher groups participated in the evaluation.

How these strategies are translated into action is discussed in detail by Spiel et al. (2006). Therefore, in the following only some issues are discussed. All these issues are related to the sample selection and therefore have consequences for the

evaluation results and their interpretation. That means the definition and recruitment of the participants of curriculum evaluation remain a challenge.

Concerning the data collection (we defined four independent subgroups) only the student group was easy to recruit. In particular, difficulties arose for the teacher and the graduate subgroups. For the latter it was impossible to calculate the response rate. During their obligatory internship graduates tend not only to change the department but also the hospital and we often had invalid addresses. In the teacher subgroup, participants did not indicate their age and gender to avoid identification. Furthermore, their response rate was systematically related to their attitudes towards the reform of the curriculum: university teachers of basic theoretical subjects who were exhibiting a high degree of fear concern that the reform would reduce theoretical parts of medical education showed a higher response rate than the other teachers but also assessed the necessity of the reform lower than the other groups. That means evaluators have to bear in mind possible effects of self-selection in the final evaluation sample caused by attitude biases.

The analysis of the concrete goals of the curriculum and its realization showed (except for factual knowledge in natural and social sciences) a huge discrepancy between the assessments of teachers and supervisors (ideal) and the assessments of students and graduates (real; see Figure 1; for details see Spiel et al., 2006). These findings support the importance of evaluating both views. What should be imparted obviously diverges from what is actually imparted. Both teacher groups were in complete agreement, but students and graduates gave different answers although the general trend is the same for both.

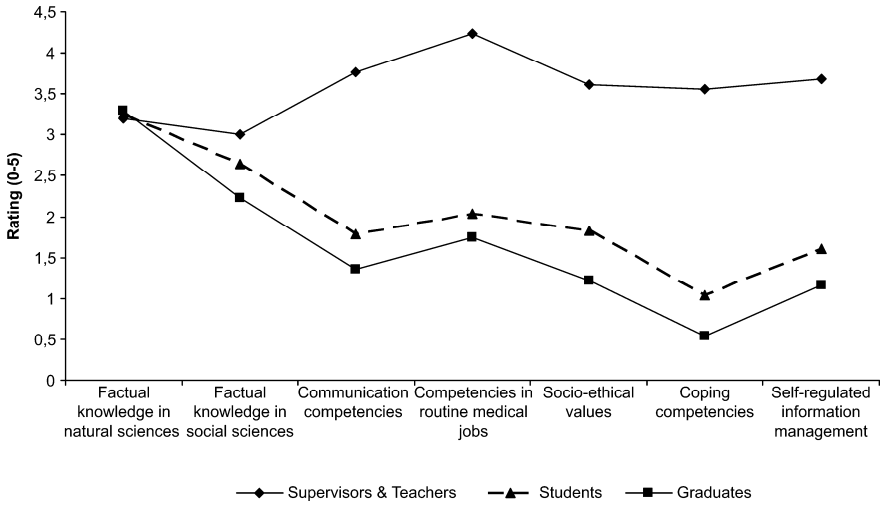


Figure 1. Competencies medical students should acquire at university (ideal; teachers' and supervisors' perspective) and how these competencies are actually imparted (real; students' and graduates' perspective).

To find out whether the discrepancy between students' and graduates' ratings is an effect of time and possibly a consequence of a practice shock, we divided the sample of graduates into two subgroups, a beginners' group, who had started their three-year internships within the past year, and an advanced group who had started their internships at least one year ago, and compared both groups with the students. Results indicated that the hypothesis concerning an immediate shock of practice could not be confirmed (see Spiel et al., 2006). The longer the practical experience the poorer the assessment of the imparting of relevant medical competencies at university. This finding, however, raises a significant question: is there an ideal time lag after the final exams when graduates should be interviewed about the university education they have received? In our opinion and experience this question cannot be answered satisfactorily and no general recommendations can be made. The only conclusion we can draw from our results is that evaluators should be aware of possible temporal effects of assessments when interpreting data.

As mentioned before, effectiveness of the medical curriculum was investigated by asking students and graduates to assess their own competencies and university teachers and supervisors to evaluate the competencies of their particular learning group.

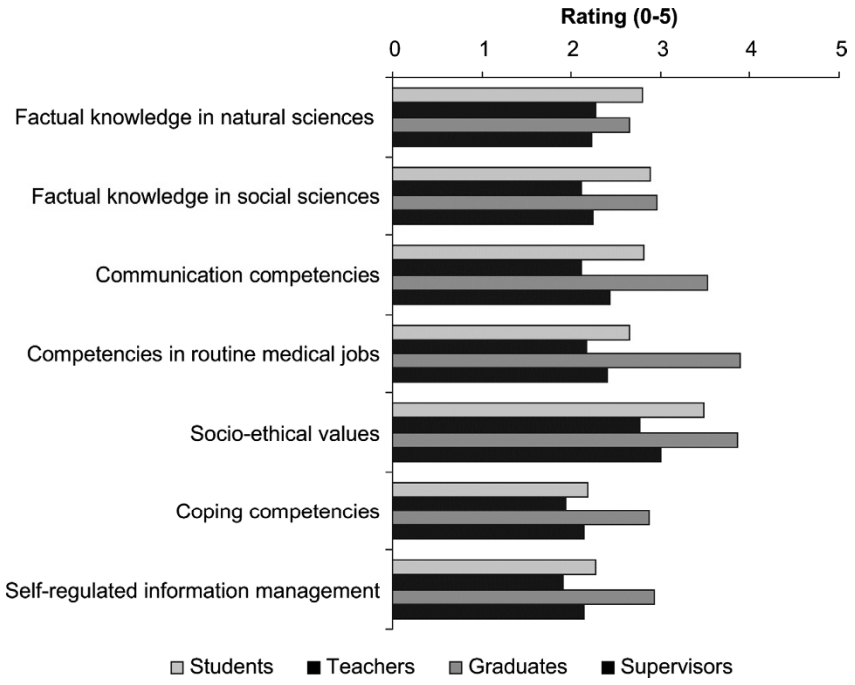


Figure 2. Competencies of students and graduates (as indicators of the effectiveness of medical education) assessed by themselves and by their instructors (teachers and supervisors) on a six-point rating scale (0 = not at all to 5 = very good).

The results obviously demonstrate the necessity to involve various stakeholder groups in data collection to get a reasonably valid impression of the situation. Both teacher groups (university teachers and supervisors) assessed the level of competencies possessed by their learner groups (students and graduates) to be considerably lower than was self-assessed by them (Figure 2; for details see Spiel et al., 2006). There were no significant differences between the two teacher groups.

In sum, based on our experiences from the baseline evaluation of medical education we recommend evaluators of curricula in higher education to focus on competencies. If data about actual performance of graduates in natural settings cannot be gathered, information both from learners (self-assessment) and teachers (external/expert assessment) and from an ideal perspective (what should be imparted) and a real perspective (what is in fact imparted) should be collected. In addition, we propose to gather data about the study behavior of students to get information about their experience with teaching methods provided in the curriculum.

Of course, it cannot be assumed (as demonstrated by the evaluation data) that the different assessments described strongly correspond. It must be assumed that there is not just one truth here but the influence of different criteria and different valuable information. In the ideal scenario these differences lead to a further step in the evaluation process: all views have to be considered as relevant and important data; members of the different groups have to start a process of negotiation as recommended by Guba and Lincoln (1989).

TAKE-HOME MESSAGES FOR MODELING AND MEASUREMENT OF COMPETENCIES IN HIGHER EDUCATION

From our experience and knowledge in the field of evaluation and, in particular, in the field of evaluation of curricula including the evaluation of entrance examinations (Spiel, 2010; Spiel, Litzenberger & Haiden, 2007; Spiel, Schober & Litzenberger, 2008) and the accreditation of non-formal and informal learning experiences at universities (Spiel, Finsterwald & Schober, 2009) we strongly recommend considering the relation between the competence profile of freshers and the competence profile of graduates as a framework when one is conceptualizing projects for modeling and measuring competencies in higher education. Whereas the profile of freshers represents the bottom line of performance in higher education, the contribution of higher education, concretely the specific curriculum and its realization at a specific higher education institution, represents the value-added to reach the competence profile of graduates. Consequently, it should be taken into account how the curriculum contributes to the development and promotion of competencies in students (ideal = goals of the curriculum, experts' perspective, real = quality of implementation, students' and graduates' perspective). Furthermore, the context for education has to be considered and contextual data collected. Here the teaching quality and context conditions such as student-teacher relations are of relevance.

Based on this framework we recommend evaluators to …

- …consider the different concepts of evaluation (from baseline to impact evaluation) when developing and evaluating a curriculum; e.g. both prospective and process evaluation of a curriculum help to avoid such big discrepancies as we have observed between the assessment of the curriculum's goals (teachers' and supervisors' view) and its realization (students' and graduates' view); furthermore, if differences occur the process evaluation might identify them very early and provide information about how to overcome them;
- …bear in mind the four levels of evaluation; we strongly recommend them not to stop at either the first level as it is done by students' ratings of courses or at the second level (learning); modeling and measuring competencies are obviously evaluation at the second level; however, given how the acquired competencies could be transferred in later professional life is of high relevance for the single student, for the higher education institute, and also for society as a whole; the latter one also corresponds to the fourth level of evaluation (results);
- …take into account the whole competence profile even if the focus of modeling and measurement is on single competencies; competencies are not independent of each other; moreover, they are developed and promoted in relation to (or dependence on) other competencies; generic competencies such as those for self-regulated learning have been shown to support the promotion of discipline-related competencies (e.g. Schunk & Ertmer, 2000); consequently, curricula should explicitly support such competencies in students (see e.g. Schober, Wagner, Reimann & Spiel, 2008);
- …to intensify research which systematically combines theoretical models of competencies with adequate measurement models and respective assessment instruments (see e.g. Klieme & Leutner, 2006);
- ...to contrast findings obtained by applying such an approach with self-ratings of graduates and the external ratings of their supervisors;
- ...to consider that different education/career stages might be related to different frames of reference which could influence modeling of competencies.

We are very hopeful that the many lessons which evaluators have learned can be useful for modeling and measurement of competencies in higher education. It will, however, be necessary to adapt procedures to the particular competencies and institutions of higher education under study.

REFERENCES

Atria, M., Reimann, R., & Spiel, C. (2006). Qualitätssicherung durch Evaluation. Die Bedeutung von Zielexplikation und evaluativer Haltung [Quality assurance by evaluation. The importance of goal explication and an evaluative mindset]. In C. Steinebach (Ed.), *Handbuch Psychologische Beratung* [Handbook of psychological counselling] (pp. 574–586). Stuttgart: Klett-Cotta.

Berk, R. A., & Rossi, P.H. (1998). *Thinking about program evaluation.* Thousand Oaks, CA: Sage.

Bortz, J., & Döring, N. (2006) *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* [Research methods and evaluation for human and social research] (4th ed.). Berlin: Springer.

Calhoun, J. G., ten Haken, J. D., & Woolliscroft, J. O. (1990). Medical students' development of self- and peer-assessment skills: A longitudinal study. *Teaching and Learning in Medicine*, *2,* 25–29.

Clack, G. B. (1994). Medical graduates evaluate the effectiveness of their education. *Medical Education, 28*, 418–431.

Deutsche Gesellschaft für Evaluation (DeGEval)., (2002). *Standards für Evaluation* [Standards for evaluation]. Cologne: DeGEval.

Fink, A. (1995). *Evaluation for education & psychology*. Thousand Oaks, CA: Sage.

Gerrity, M. S., & Mahaffy, J. (1998).Evaluating change in medical school curricula: How did we know where we were going? *Academic Medicine, 73*, 55–59.

Guba, E.G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hewson, M. G., Copeland, H. L., & Fishleder, A. J. (2001). What's the use of faculty development? Program evaluation using retrospective self-assessments and independent performance ratings. *Teaching and Learning in Medicine, 13*, 153–160.

Hill, J., Rolfe, I. E., Pearson, S. A., & Heathcote, A. (1998). Do junior doctors feel they are prepared for hospital practice? A study of graduates from traditional and non-traditional medical schools. *Medical Education, 32*, 19–24.

Joint Committee on Standards for Educational Evaluation (JCSEE). (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.

Kirkpatrick, D. L. (1998). *Evaluating training programs*. *The four levels*. San Francisco, CA: Berrett-Koehler.

Kirkpatrick, J. D., & Kirkpatrick, D. L. (2005). *Transferring learning to behavior. Using the four levels to improve performance*. San Francisco: Berrett-Koehler.

Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? [What are competencies and how can these be measured?] *Pädagogik, 56*(6), 10–13.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen – Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG [Competence models for the recording of individual learning results and for the assessment of educational processes – The German Research Association (DFG)'s plans for new fields of emphasis]. *Zeitschrift für Pädagogik, 52(6),* 876–903.

Marsh, H. W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74,* 264–279.

Merl, P. A., Csanyi, G. S., Petta, P., Lischka, M., & Marz, R. (2000). The process of defining a profile of student competencies at the University of Vienna Medical School. *Medical Education, 34*, 216–221.

Morrison, J. (2003). ABC of learning and teaching in medicine: Evaluation. *British Medical Journal, 326*, 385–387.

Patton, M. Qu. (1996). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Rolfe, I. E., Andren, J. M., Pearson, S., Hensley, M. J., & Gordon, J. J. (1995). Clinical competence of interns. *Medical Education, 29*, 225–230.

Rossi, P.H., & Freeman, H.E. (1993). *Evaluation. A systematic approach* (5th ed.). Newbury Park, CA: Sage.

Schober, B., Spiel, C.,& Reimann, R.,(2004). Young physicians' competences from different points of view. *Medical Teacher, 26*, 451–457.

Schober, B., Wagner, P., Reimann, R., & Spiel, C. (2008). Vienna E-Lecturing (VEL): Learning how to learn self-regulated in an internet based blended learning setting. *International Journal on E-learning, 7*, 703–723.

Schunk, D. H., & Ertmer, P. H. (2000). Self-regulation and academic learning: Self efficacy enhancing interventions. In M. Boekaerts, P.R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 631–650). San Diego, CA: Academic Press.

Schweer, M. K. W. (2001). Evaluation der Lehre [Evaluation of teaching]. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* [Handbook of educational psychology] (2nd ed., pp. 466–471). Weinheim: PVU.

Shadish, W. R., Newman, D. L., Scheirer, M. A., & Wye, Ch. (1995). *Guiding Principles for Evaluators (New Directions for Program Evaluation, No. 66)*. San Francisco, CA: Jossey-Bass.

Sonneck, G.. (1994). *Bildungsziele und Lehrveranstaltungen im Medizinstudium* [Goals and courses of medical education]. Vienna: Facultas.

Spiel, C. (2001), Program evaluation. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (Vol. 18, pp. 12169–12173). Oxford: Elsevier Science.

Spiel, C. (2010). Hochschulzugang: Bestandsaufnahme und Empfehlungen [University admission: Review and regards]. In AQA – Österreichische Qualitätssicherungsagentur [Austrian quality assurance agency], *Mobilität, Durchlässigkeit und Qualität* [Mobility, permeability and quality] (pp. 39–43). Vienna: Facultas.

Spiel, C., Finsterwald, M., & Schober, B. (2009). Anerkennung non-formalen und informellen Lernens an Universitäten [Admission of non-formal and informal learning at universities]. In E. Westphal & M. Friedrich (Eds.), *Anerkennung von non-formalen und informellen Lernen an Universitäten* [Admission of non-formal and informal learning at universities] (pp. 29–83). Graz: Leykam.

Spiel, C., Gradinger, P., & Lüftenegger, M. (2010). Grundlagen der Evaluationsforschung [Principles of scientific evaluation]. In H. Holling H. & B. Schmitz (Eds.), *Handbuch Statistik, Methoden und Evaluation* [Handbook of statistics, methods and evaluation] (pp. 223–232). Göttingen: Hogrefe.

Spiel, C., Litzenberger, M., & Haiden, D. (2007). Bildungswissenschaftliche und psychologische Aspekte von Auswahlverfahren [Educational and psychological aspects of selection procedures]. In C. Badelt, W. Wegscheider & H. Wulz (Eds.), *Hochschulzugang in Österreich* [university admission in Austria] (pp. 479–552). Graz: Grazer Universitätsverlag.

Spiel, C., Lüftenegger, M., Gradinger, P., & Reimann, R. (2010). Zielexplikation und Standards in der Evaluationsforschung [Goal explication and standards in scientific evaluation]. In H. Holling H. & B. Schmitz (Eds.), *Handbuch Statistik, Methoden und Evaluation* [Handbook of statistics, methods and evaluation] (pp. 252–260). Göttingen: Hogrefe.

Spiel, C., Schober, B., & Litzenberger, M. (2008). *Evaluation der Eignungstests für das Medizinstudium in Österreich* [Evaluation of the selection procedures for medical studies in Austria]. Projektbericht für das Bundesministerium für Wissenschaft und Forschung, Wien [Project report for the ministry of science and research, Vienna]. Vienna: University of Vienna, Faculty of Psychology.

Spiel, C., Schober, B., & Reimann, R. (2006). Evaluation of curricula in higher education: Challenges for evaluators. *Evaluation Review, 30,* 430–450.

Widmer, T., Landert, C., & Bachmann, N. (2000). *Evaluations-Standards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards)* [Evaluation standards of the Swiss evaluation association]. Retrieved from http://www.seval.ch/de/documents/seval_Standards_2001_dt.pdf.

Wottawa, H., & Thierau, H. (1998). *Lehrbuch Evaluation* [Textbook evaluation]. Bern: Hans Huber.

Christiane Spiel & Barbara Schober
Department for Economic Psychology, Educational Psychology and Evaluation
Faculty of Psychology
University of Vienna

Ralph Reimann
Quality Management
University of Natural Resources and Life Sciences, Vienna

ROLF VAN DER VELDEN

MEASURING COMPETENCES IN HIGHER EDUCATION: WHAT NEXT?

INTRODUCTION

The theme of this conference, "Modeling and Measurement of Competences in Higher Education", indicates that there is a growing awareness that the measurement of competences should not be restricted to primary and secondary education, but should also enter into the domain of higher education. In this contribution, I will provide a sketch of the current state of large-scale skills assessments and the challenges that lie ahead. I will argue that it is important to link the measurement of competences in higher education to economic and social outcomes, and I will provide a short impression of the kind of information that graduate surveys can provide. I will discuss some of the main implications that can be drawn from these graduate surveys for higher education.

THE STATE OF THE ART[1]

In the last few decades we have seen an increased awareness of human capital as one of the driving forces behind economic development. Research has provided sound evidence that investments in education provide large economic and social returns both for the individual and for society at large. As a result, different actors in society (policy-makers, employers, employees, students) have increasingly invested in education and training as a way of improving the existing stock of human capital.

A development which accompanied this increased interest in education and learning was the need to monitor and assess the stock of human capital. The Organisation for Economic Co-operation and Development (OECD) played a prominent role in this by initiating the so-called Indicators of Education Systems (INES) project, aimed at developing indicators of the input, process and output of education and training. The results of this project are published annually in the publication "Education at a Glance".

What soon became clear is that education as such is only a poor indicator of the stock of human capital. Individuals of the same level of education show a strong heterogeneity in skills. Likewise, countries that have more or less comparable levels of educational attainment still show large differences in the proficiency levels of different skills. Moreover, skills acquisition does not only take place in education. People also learn through work experience and in daily life, which leads to a further loosening of the link between educational qualifications and the later stock of skills.

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 207–216.

This has caused a paradigm shift, from measuring educational attainment to measuring competences or skills. The basic idea is that educational attainment as such may be important, but the driving mechanism behind the effect of education on economic and social outcomes operates through the skills and competences that these educational qualifications represent. A suitable illustration is given in a recent overview by Hanushek and Woessmann (2011). They argue that a net improvement in the literacy scores of 15-year-olds by a quarter of a standard deviation would increase economic growth in OECD countries by almost 300%, or approximately US$125 trillion, by 2090. They show that skills and competences completely account for the projected effect of increased educational attainment on economic growth in OECD countries.

Over the past few decades, a great deal of progress has been made in assessing so-called generic basic skills, mainly in the areas of literacy, numeracy, science and civics. The most well-known examples are international cross-sectional surveys like the Trends in International Mathematics and Science Study (TIMMS), aimed at assessing the mathematics and science levels of students in primary and secondary education, the Progress in International Reading Literacy Study (PIRLS), aimed at assessing the literacy levels of primary school students, the Program for International Student Assessment (PISA), aimed at assessing the literacy, numeracy and science levels of 15-year-olds, the Civic Education Study (CIVED) and its successor the International Civic and Citizenship Education Study (ICCS), aimed at assessing the civic competences of secondary school students, and the International Adult Literacy Survey (IALS) and its successors the Adult Literacy and Life Skills Survey (ALL) and the Program of International Assessment of Adult Competences (PIAAC), all aimed at assessing the literacy and numeracy levels of 16 to 64-year-olds. However, there are also important national examples of surveys and panel studies that aim to assess the stock of basic skills, like the U.S. National Longitudinal Survey of Youth (NLSY), the Collegiate Learning Assessment (CLA), the Longitudinal Survey of Australian Youth (LSAY), the Dutch Secondary Education Cohort of Students (VOCL) or the recent German National Education Panel Study (NEPS).

As indicated above, all of these surveys assess generic skills. Up to now, no comparable surveys of this kind have been carried out in order to assess vocation-specific skills (one exception is the Mexican Higher Education Exit Assessments, see contribution of Rafael Vidal Uribe in this volume). Nevertheless, there is clear evidence that these specific skills are just as important as general skills (Bishop, 1995; Van der Velden, 2006), especially for economic outcomes. They also constitute a large part of what is being learned in vocational education and higher education. Based on the notion of the importance of specific skills, feasibility studies have been carried out to study such skills in vocational education and training (PISA-VET, Baethge et al., 2006) and higher education (see the contribution by Karine Tremblay of the OECD on the Assessment of Higher Education Learning Outcomes (AHELO) project in this volume). The main reason that assessments in specific areas have not yet become widespread is the sheer variety of specific domains that can be distinguished. While it is easier to define a

limited number of key generic skills and to develop tests in order to assess these skills, the number of specific domains is, by definition, much larger. In addition, in terms of design, it is easier to administer generic skills assessments, as these can be applied to the whole population. For obvious reasons, this is not the case for specific skills, like car mechanics, accounting or carpeting. Nevertheless, there is no reason to believe that these domains are more difficult to assess than generic skills, like literacy or problem-solving.

For higher education, it is important to measure both general academic competences as well as discipline-specific competences (as is the objective of the AHELO project). Both have been shown to strongly determine graduates' success in the labor market (Meng, 2006; Allen & Van der Velden, 2011b). However, there are a few obstacles which are yet to be tackled, which will be discussed below.

THE DIFFICULT CONCEPT OF COMPETENCES

Educational research has shown the importance of focusing on competences rather than skills, and we have seen a strong movement toward competence-based education. There are many definitions in existence, but that which was formulated during the so-called DeSeCo (Definition and Selection of Competences) project seems to be very useful. This project was initiated by the OECD in order to provide an overarching framework for international skills assessments. Emphasizing the need for competence assessment rather than a narrow focus on skills, competences are defined in this project as "the ability to successfully meet complex demands in a particular context through the mobilization of psychosocial prerequisites (including both cognitive and non-cognitive aspects)" (Rychen & Salganik, 2003, p. 43). The basic difference from previous concepts of skills is the holistic nature of the concept of competence. First, there is a direct link to performance, in the sense that competence relates to meeting demands successfully. Second, the definition clearly refers to a range of cognitive and noncognitive skills rather than just one skill. Thirdly, the concept of competence, refers to the notion of "orchestration", that is the ability to use these different skills in a meaningful and deliberate way. In this regard, the "whole" that makes up a competence is more than just the "sum of its parts". Skills are therefore best considered as one of the constituent elements of a competence. Given this definition, all of the aforementioned assessments measure skills rather than competences, which is the reason why I have refrained from using the word competence when referring to these assessments.

THE CHALLENGES

One of the main challenges for large-scale assessments is to try to broaden their scope to measure competences rather than skills. Some of the current initiatives in the area of problem-solving (PISA and PIAAC) already contain elements that are more holistic in nature and refer to the notion of orchestration. However, they are still far from what has been defined here as competence. It seems to be especially

difficult to broaden the scope to include non-cognitive aspects. Attempts to develop tests of soft skills like teamwork and communication skills have proven very difficult (Murray et al., 2005) and have not yet resulted in tests that can be compared fully across relevant populations. However, many of the so-called 21st century skills involve both cognitive and noncognitive aspects. The term "21st century skills" has been coined to refer to those skills that seem to be specifically relevant in the modern knowledge economy, such as creativity, critical thinking, learning skills, socio-communication skills and self-management skills. It is therefore crucial to include such skills in a competence assessment in higher education.

Another major challenge in this line of research is the identification of underlying causal mechanisms. It is easy to identify a correlation between educational attainment and skill level, or between skill level and earnings, but this is not the same as claiming that education actually imparts these skills, or that these skills actually increase productivity at work. The aim to improve our understanding of causal mechanisms in competence assessments in higher education has a number of implications.

First, we need to be able to assess the added value of higher education. Higher education institutes differ in terms of their selection mechanisms, and part of the difference in output results simply from differences in input. It is therefore important to have a longitudinal design with repeated measures of competence at the beginning and end of higher education.

Second, we need to be able to link the development of these competences to characteristics of the study program, for example, the modes of teaching and learning. Are certain competences better developed in student-centered environments like project and problem-based learning than in more traditional classroom settings? What is the effect of internships on the development of specific skills? What is the role of engaging in research activities? A competence assessment in higher education needs to be able to link differences in learning outcomes to different settings/environments in order to improve our understanding of the causes of these differences.

Third, we need to have a design that will help us to deal with unobserved heterogeneity. Having good control variables like input characteristics and good measures of the competences at the start of higher education is one step, but it will still not rule out the problem of unobserved heterogeneity. This is a general problem in large-scale assessments which are based on survey data rather than experimental data. Over the past two decades, new statistical techniques have been developed in order to address this problem, such as propensity score matching, difference-in-difference or instrumental variables. However, the success of applying these techniques depends greatly on the quality of the relevant control variables or instrumental variables. Although researchers have been very innovative in the creation of instruments, this process usually occurs after the data have been collected. In order to successfully identify causal relations, it is important to build some experimental variation or good instrumental variables into competence assessments in advance.

THE NEED TO MEASURE OUTCOMES

The value of a competence assessment in higher education will be greatly enhanced when it is linked to graduate surveys for two reasons. The most obvious reason is that it is not enough to measure acquired competences in higher education. We also need to examine the way in which graduates cope with the requirements of their work. The proof of the pudding is in the eating, and it is important to assess how competences determine labor market success. Graduate surveys are also important for another reason, as they provide a means to evaluate the standards that have been set for the assessment in question. Usually, these standards are formulated by subject matter experts, but graduate surveys provide an opportunity to evaluate how these standards work in practice. Are the levels that have been set sufficient to function well on the labor market? Moreover, having information on both the acquired level of competences (through the assessment) and the required level (through the graduate survey) provides an opportunity to examine the mismatch in the labor market. Skills shortages and the underutilization of skills are both undesirable and lead to a loss of productivity. Linking competence assessments in higher education to graduate surveys will provide further insight into the prevalence of these types of skills mismatch.

Many countries have already conducted some form of graduate survey, but there are only a few international comparative studies: the CHEERS survey (Careers After Higher Education: A European Research Study; see http://www.uni-kassel.de/wz1/tseregs.htm) and its successor the REFLEX project (Research into Employment and Professional Flexibility, see http://www.reflexproject.org). Both surveys focus on the transition from higher education to work and were carried out three (CHEERS) and five years (REFLEX) after graduation.

THE REFLEX PROJECT

The REFLEX project[2] was carried out in approximately 20 European countries and Japan. It is based on a representative sample of around 100,000 graduates, five years after they left higher education. The survey was carried out in 2005 (initial REFLEX sample) and 2008 (the extension to Eastern Europe, the Higher Education as a Generator of Strategic Competences (HEGESCO) project, see http://www.hegesco.org). Today, the REFLEX project constitutes the largest existing graduate database. REFLEX gathered detailed information on careers in higher education, the characteristics of the study program, the modes of teaching and learning, entry into the labor market, the characteristics of graduates' first and current job and the acquired and required level of skills. In this section, I will provide a short impression of the most relevant findings. For more detailed information, see Allen and Van der Velden (2009, 2011b).

What Does the World of Work Look Like to Higher Education Graduates?

The world of work for higher education graduates is best described by the following keywords:

- International: almost 40% of the graduates work in organizations that have an international scope;
- Competitive: approximately 85% of the graduates working in the private sector work in firms which face strong competition, mainly on quality. Even in the public sector, a sizeable minority of the organizations in which the graduates work face strong competition;
- Innovative: 70% of the graduates are in some way engaged in innovative activities. These innovative activities are often related to innovation in knowledge and methods;
- Insecure: 50% of the graduates have experienced some form of reorganization since starting work with their current employer. Note that these figures date from the period prior to the economic crisis of 2009 and 2010, and so this is more a structural than a temporary phenomenon;
- Professional: graduates are often required to act as an authoritative source of advice for their colleagues. They are expected to take the initiative to engage in professional contact and they experience a high level of professional autonomy.

What are the Relevant Areas of Competence?

In the REFLEX project, we have identified five relevant areas of competence for graduates: professional expertise; functional flexibility; innovation and knowledge management; mobilization of human resources; and international experience.

Items per area of competence
Professional expertise
Mastery of one's own field or discipline
Analytical thinking
Ability to assert authority
Functional flexibility
Knowledge of other fields or disciplines
Ability to acquire new knowledge rapidly
Ability to negotiate effectively
Innovation and knowledge management
Ability to use computers and the Internet
Ability to generate new ideas and solutions
Willingness to question one's own and others' ideas
Alertness to new opportunities

Mobilization of human resources
Ability to perform well under pressure
Ability to use time efficiently
Ability to work productively with others
Ability to mobilize the capacities of others
Ability to make one's meaning clear to others
Ability to coordinate activities

For the first four domains, we developed scales with items reflecting different aspects of these domains. Respondents were asked to indicate the extent to which these competences were required in their current job and the degree to which they actually possessed them. The items relating to these domains are listed in the table above. In addition, questions were asked about the respondents' foreign language proficiency and international experience in order to indicate their international orientation.

How were these domains evaluated? The results show that first and foremost, graduates need to be professional experts, as professional expertise drives labor market success. A high level of professional expertise is directly linked to a lower chance of becoming unemployed, a better chance of finding a job that matches one's level of education and higher earnings. Moreover, professional expertise is also relevant when graduates are working outside their own domain. Graduates who are experts in their own domain are also able to use their knowledge and skills to the full when working outside their own domain.

The second most important area of competence is what we have called the "mobilization of human resources". This is defined as the ability of graduates to use their own and others' capacities. The basic idea is simple: if human capital is the driving force behind the economy, then mobilizing this human capital is crucial. It is obvious that a graduate who is unemployed or who is working in a job in which his/her knowledge and skills are not fully utilized is contributing very little to the overall economic growth. Therefore, it is important that graduates possess the competences to mobilize their own capacities, such as self-management skills, organizational skills and so on. Moreover, graduates can be called upon to mobilize the capacities of others. This can relate to direct leadership skills, but it also includes the ability to create synergy in teams. Moreover, it involves the capacity to mould the work environment so as to better fit one's own competences and those of one's colleagues or subordinates. The results show that possessing the ability to mobilize human resources has a positive effect on finding a job quickly as well as on the earnings associated with that job.

The third important area relates to international orientation. As a result of globalization, graduates are increasingly expected to work in an international environment. This not only implies a good command of foreign languages, but also requires an ability to understand and empathize with other cultures and to reflect on the limitations of one's own culture. An important way of improving one's international orientation is to spend time abroad for study or work. Unsurprisingly, the results show that this kind of international experience has a positive effect on the probability of being internationally mobile after graduation or the chance of

obtaining work that requires international competences. In addition, it is also related to higher wages in general.

A fourth area which is relevant relates to functional flexibility, which is the ability to cope with changes in the work environment. As indicated above, many graduates are faced with important changes in their work environment (such as reorganization) which have an impact on their work tasks. Graduates must be able to deal with such changes and, in some cases, must even be prepared to take up tasks which are not directly related to their own field of expertise. Our findings show that competences in this area are needed, but not directly rewarded. Having a certain level of functional flexibility should not be seen as an investment that will directly pay off, but more as an insurance policy that will safeguard job opportunities when graduates are faced with change. In this sense, this concept is closely related to employability.

Finally, graduates are expected to play a role in the area of innovation and knowledge management. This relates not only to pure research and development activities but more generally to the ability to create an environment in which knowledge production and diffusion is optimized and in which innovations are implemented. This involves creativity, an ability to notice new opportunities as well as the organizational ability to implement innovations within the organization. Our findings show that innovative competences are important, but only when graduates are directly involved in actual innovative activities. Although most innovation takes place in large organizations, these organizations do not always fully utilize the innovative competences of the graduates they employ as a result of the internal division of labor.

What are The Lessons That can be Drawn from The REFLEX Study for Higher Education?

The findings of the REFLEX project suggest that there is clear evidence of a dual orientation in higher education. The most successful programs in terms of labor market outcomes are programs that have either a strong academic reputation or a strong vocational orientation. Both types of program give their graduates a competitive advantage on the labor market. Strong academic programs signal a high learning ability and strong academic skills based on selective entry requirements as well as a high-quality academic learning environment. Strong vocational programs, on the other hand, impart relevant vocational skills that can be deployed immediately on the labor market. Due to the heavy involvement of employers in the latter type of program, they also ensure a smooth transition to the labor market. There is a need for both types of program on the labor market, and it is important for higher education programs to opt for either one or the other.

Another relevant finding is that the acquisition of relevant competences is closely related to the demands of the study program. More demanding programs impart more relevant competences, which is a simple result of "time spent on task". More demanding programs require students to work harder and spend more hours studying. The overall impression is that higher education programs are not

very demanding, although there are significant differences both between and within countries.

As indicated above, professional expertise is the driver of labor market success, even when graduates are working outside their own domain. This finding is particularly relevant in the context of a discussion about whether higher education should produce generalists or specialists. It suggests that a good education in a particular field not only provides graduates with skills that are needed in jobs that match that field, but also provides a basis for the development of more general analytical skills that can be applied in other areas as well. In this sense, training in a specific field of knowledge serves as a carrier through which generic skills can be developed. Academic skills cannot be developed without some relation to content and it is this content that constitutes the heart of the specific discipline or field of study. The importance of specific knowledge should therefore not be underestimated.

In the past decade, higher education has become more internationally oriented. More and more students spend part of their study period abroad. However, it seems that today, the world in which graduates are working is changing even more rapidly. Graduates are now expected to have a strong international orientation and excellent foreign/English language proficiency. The findings show, however, that this is the area in which most graduates indicate that they have serious shortages.

There is evidence from the graduate surveys that student-centered methods like project and problem-based learning positively affect general academic skills (Meng, 2006). These generic academic skills are better developed in such environments than in a traditional classroom setting. However, there is also a strong indication that discipline-specific skills are better developed in an environment in which the teacher acts as an important source of advice. Meng (2006) therefore concludes that problem-based learning in which the teacher plays a strong role provides the best environment in which to develop both kinds of skills. This implies that the role of the teacher in student-centered methods should not be limited to merely a "coach of the learning process", but should also be more active, particularly in transferring specific knowledge (e.g., through lectures).

Finally, there is evidence that the learning process is determined not only by the curriculum, but also by assessment, as assessment drives learning. Higher education institutes should be more aware that the way in which they assess students should be in line with the kind of skills they want to develop. Multiple-choice exams do not foster academic skills, but merely reflect the short-term memory of students. The findings from the REFLEX survey confirm this theory and show that students learn more from essays and oral exams than from multiple-choice exams.

NOTES

[1] The following section is based largely on Allen and Van der Velden (2005) and Allen and Van der Velden (2011a).

[2] The initial REFLEX project was financed by the European Sixth Framework Program (Contract No: CIT2-CT-2004-506-352) and the HEGESCO project was financed by the ERASMUS lifelong learning program.

REFERENCES

Allen, J., & Van der Velden, R. (2005). *The role of self-assessment in measuring skills*. REFLEX Working Paper 2. Maastricht: ROA.

Allen, J., & Van der Velden, R.. (2009). *Competencies and labour market careers of higher education graduates*. Maastricht: ROA.

Allen, J., & Van der Velden, R. (2011a). *Skills for the 21st century: Implications for education*. Maastricht: ROA.

Allen, J., & Van der Velden, R. (2011b). *The flexible professional in the knowledge society: New challenges for higher education*. Dordrecht: Springer.

Baethge, M., Achtenhagen, F., Arends, L., Babic, E., Baethge-Kinsky, V., & Weber, S. (2006). *PISA-VET: A feasibility study.* Stuttgart: Frans Steiner Verlag.

Bishop, J. (1995). Expertise and excellence. *CAHRS Working Paper Series, 95*(13).

Hanushek, E., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Amsterdam: North Holland.

Meng, C. (2006). *Discipline – specific or academic?* Acquisition, role and value of higher education competencies [Thesis]. Maastricht: Maastricht University.

Murray, S., Clermont, Y., & Binkley, M.. (2005). *Measuring adult literacy and life skills: New frameworks for assessment*. Catalogue No. 89-552-MIE, No. 13. Ottawa: Statistics Canada.

Rychen, D. S., & Salganic, L. H. (2003). A holistic model of competence. In D. S. Rychen & L. H. Salganic (Eds.), *Key competencies for a successful life and a well-functioning society* (pp. 41–62). Göttingen: Hogrefe & Huber.

Van der Velden, R. K. W. (2006). *Generiek of specifiek opleiden?* [Generalists or specialists?] Inaugural address. Maastricht: Maastricht University.

Rolf van der Velden
Research Center for Education and the Labor Market,
Maastricht University, the Netherlands

ULRICH TEICHLER AND HARALD SCHOMBURG

ANALYZING THE RESULTS OF STUDY IN HIGHER EDUCATION AND THE REQUIREMENTS OF THE WORLD OF WORK

INTRODUCTION

Since the 1990s, interest in obtaining more precise information about the results of study and their relation to the requirements of the world of work has increased in many countries across the world, including Germany. Three reasons for this growing interest can be identified.

First, *evaluation* activities in higher education have increased dramatically in recent decades, and in this context, the focus has increasingly shifted to "output awareness": the measurement of the core activities of higher education – teaching and learning as well as research – and the factors which contribute to success in these domains as part of an information and monitoring system. Measurements should serve more regularly, more comprehensively and more systematically the purposes of both reflection and efforts for improvement on the part of the actors and control in terms of providing the basis for rewards and sanctions. Both mechanisms are expected to contribute to the enhancement of the quality, relevance and efficiency of higher education. In the arena of teaching and learning, scholars are encouraged not only to focus on the transmission and assessment of knowledge acquisition, but also to reflect on the abilities acquired up to the point of graduation. Terms such as "learning outcomes" (a completely misleading term, because "learning outputs" are referred to as a rule!), "competences" and "skills" signal a paradigm shift from the acquisition of knowledge – i.e., the essence of the academic system – to knowledge-enriched abilities which help to cope with practical problem-solving (see the overview in Cavalli, 2007).

Second, *quantitative-structural changes in higher education* have led to greater interest in more detailed information about the results of study. For example, experts agree that the dramatic expansion of student enrolment (which has increased more than tenfold over five decades worldwide) has been accompanied not only by an increase in the variety of students' motives, talents and career perspectives, but also – as a response – by increasing diversity within higher education. In addition, as higher education has become more diverse, simple indicators of competences, such as the field of study of a graduate or the grade on his or her final certificate, are less likely to be viewed as sufficient for assessing graduates' competences and the relationships between demand and supply on the graduate labor market. Similarly, reforms of study programs, such as the introduction of a convergent cycle system and degrees across Europe in the framework of the Bologna Process, raise the question of how

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 217–228.

the competences encompassed in a bachelor's degree or a master's degree differ from those acquired on completion of the preceding or persisting single-cycle programs. Qualifications Frameworks have been established at the European level since approximately 2005 in order to set standards for these competences (see Gehmlich, 2009).

Third, the pressure is increasing on higher education to ensure that the results of teaching/learning and research are useful for technology, the economy, social well-being and cultural enhancement. The spread of the terms "knowledge society" and "knowledge economy" and the increasing popularity of the term "employability" indicate rising *expectations that higher education ought to deliver more relevant results*, and so scholars feel exposed to instrumental pressures. As a consequence, interest is growing in the actual relevance of study for employment and work.

It is possible to name a fourth trend, although it is mentioned less frequently by experts addressing the issues discussed here. For various reasons, *trust is fading in the customary modes and sources* that have traditionally been used in the assessment of competences. New and possibly more elaborate measurements are often called for.

Efforts to improve information about the output and outcomes of study are moving in various directions. We may note a growth in statistics and indicators. We have noted an increasing role of graduate surveys in rating job requirements and competences acquired during the course of study. We have also observed increasing efforts to develop test-like measurements of competences. We believe that the rating of job requirements and competences, as is done in research (in graduate surveys) or in practice (in job interviews), is likely to remain the most widespread method and that efforts should concentrate on improving the quality of such rating activities.

It is not the intention of this article to review the state of the academic literature on the relationship between higher education and the world of work; this has already been done at regular intervals by one of the authors over several decades (Hartung, Nuthmann, & Teichler, 1980; Teichler, 1988, 1999a, 1999b, 2009). Nor is it the intention of this article to review the way in which job requirements and competences are addressed in major international comparative studies on graduate employment and work; sources for this purpose are already available (Schomburg & Teichler, 2006; Teichler, 2007; Allen & van der Velden, 2011; Schomburg & Teichler, 2011). Instead, we aim to sketch the character of these ratings of job requirements and competences upon graduation in comparison with other approaches and to use this framework to illustrate the potential of graduate surveys.

CUSTOMARY MEASURES AND SOURCES

In looking at the spectrum of research on the relationship between higher education and the world of work, we can note that all of the major modes of inquiry which are customary in social science or behavioral science research are used. We can also note, however, that the mode of inquiry has enormous consequences for the "findings" of a study and thus for the substantive debates on these links.

Five types of sources and measures are usually employed in assessing the results of study and the utilization/demand in employment and work:

– Indicators;
– Measurement during the actual processes of the higher education system;
– Actors' ratings (possibly self-ratings);
– Experts' ratings;
– Tests.

The term *indicator* usually encompasses quantitative-structural measures produced for purposes other than research that are viewed as suitable "proxies" (approximate measures for the issues one would ideally like to measure). For example, the most frequently employed indicator, the Gross National Product, measures financial flows over a certain period in a specific country, but owes its popularity to the aim of knowing the wealth of a country. Similarly, the high remuneration of university graduates is often taken as an indicator of the use of graduates' competences and work tasks.

Quantitative-structural measures play an enormous role in the assessment of the results of study and the links between study and the world of work. Drop-out rates can be viewed as partial indicators of inappropriate study provisions and conditions. Graduate unemployment rates are often interpreted as signs of wrong delivery of competences on the part of the higher education system. Low income advantages of graduates are often viewed as indicating "over-education"; similarly, the employment of an engineer as a social worker or a philosopher as a tourist guide is often viewed as indicating a substantive "mismatch" between higher education and graduate employment.

The frequent use of quantitative-structural data to assess the links between the results of study and the world of work can be seen as fairly rational, because the labor market is not an arena for linking individual competences to individual work tasks, but rather for trading competences between individual persons and work tasks in terms of job offers for individual persons. However, analyses of the links between the level of educational attainment and areas of expertise on the one hand and occupational groups, remuneration, etc. on the other have often resulted in researchers being too eager to diagnose "over-education" and "mismatches" and thereby underestimating the diversity of links between competences and work tasks.

Measurement during the actual processes of the higher education system most notably takes place in the *grading of study achievements*. The grades stated on the certificates awarded upon graduation are the most salient in this context. All of the information on such certificates (except for the name of the recipient) can be understood as an aggregate measure or as an indicator of competences: the field of study; the areas of specialization; the higher education institution; the length of study; etc. In many countries, a degree in a certain field of study is understood as entailing an "*effectus civilis*", i.e., an entry qualification (in the English sense of the word) for a profession. Certificates demand others to trust that those who are awarded them have acquired the necessary competences, at least to a certain

standard level. Employers and customers take certificates as indicators of competences, although they may use additional measures to assess the quality of graduates.

Measurements during the actual process tend to be more elaborate and specific than mere indicators. They rely on the actors' experiences. They are the incarnation of a pragmatic compromise between the desire to undertake sophisticated measurements and the intention to keep efforts within realistic bounds. They are based on the strengths and weaknesses of the practitioners' actions.

Actors' ratings are understood here as ratings assigned by actors outside of their normal work tasks within the system. Students may be asked within the framework of evaluation studies to rate the quality of teaching. Similarly, graduates may be asked within the framework of graduate surveys to rate their own competences upon graduation or the requirements of their job (see, for example, Schaeper & Wolter, 2008; Schaeper, 2009; Schomburg & Teichler, 2006). Employers may be asked in surveys to compare the usual quality of graduates with international experience and those without (see, for example, Janson, Schomburg, & Teichler, 2009). Scholars may be asked to rate the competences which are enhanced by a study program.

As a rule, the respondents to such studies are provided with lists of competences and work tasks (e.g., "leadership") and asked to rate the extent to which a job requires such competences and whether a graduate has acquired such competences. Through such procedures, more refined information can be gathered on the relationships between study and work than with the help of quantitative-structural indicators; however, there is widespread criticism that job requirements and competences may be misjudged by those involved (see, for example, Arum & Roksa, 2011).

Experts' ratings – i.e., ratings by highly knowledgeable persons who are not directly involved in the setting which is being rated – play an enormous role in higher education. In the area of research, proposals and texts to be published are rated by "peers" and publications or citations are indirectly determined by experts' ratings because the experts decide whether or not a text shall be published. In the area of teaching and learning, experts review study programs in formal processes of evaluation or accreditation (see Mittag, 2006) and are expected to state whether or not they believe that the curricula under scrutiny are likely to enhance the competences that they are intended to enhance.

Expert ratings' play such an important role because they cover a broad thematic range and because there is widespread belief within the higher education system that experienced actors who are asked to rate academic quality without being personally involved are knowledgeable and fair assessors. Research on peer reviews and related themes, however, shows that experts' ratings are not without bias (see, for example, Fröhlich, 2006).

Finally, *tests* differ from ratings as they evoke not only communication about competences, but also the emanation of competences in response to an artificial,

standardized stimulus (questions, tasks, etc.). They are generally viewed as the most sophisticated and thorough method of assessing competences.

Tests are not only very time-consuming in both their development and administration, but their role in higher education is also limited because there is widespread scepticism that tests, which must always take for granted a certain generally agreed core of knowledge and competences, do not fit with the diversity of concepts in higher education and could lead to undue pressure to homogenize; moreover, there are widespread doubts about whether tests would be suitable to measure the competences in the area of higher education that contribute to innovation and help to cope with indeterminate work tasks. Moreover, many efforts to measure the competences of students in higher education and graduates have focussed on the assessment of "generic skills", "general competence" and "key skills". Hence, many experts are concerned that a spread of testing in higher education would lead to an over-emphasis on general competences and a disregard for subject-specific and professional competences (see the critique voiced by Banta, 2007):

For nearly 50 years of measurement scholars have warned against pursuing the blind alley of value added assessment. Our research has demonstrated yet again that the reliability of gain scores and residual scores – the two chief methods of calculating value added – is negligible (i.e., 0.1). We conclude that standardized tests of generic intellectual skills do not provide valid evidence of institutional differences in the quality of education provided to students. Moreover, we see no virtue in attempting to compare institutions, since by design they are pursuing diverse missions and thus attracting students with different interests, abilities, levels of motivation, and career aspirations.

THE COMMON NEED TO IMPROVE THE UNDERSTANDING OF COMPETENCES

We can observe a multitude of valuable conceptual frameworks as regards the links between job requirements and competences (see the overviews of concepts in Bennet, Dunne, & Carrée, 2000; Knight & Yorke, 2002, 2003). In addition, the recent debates on "employability" have triggered efforts for further improvement in this direction (cf. the overview in Yorke, 2007).

Nevertheless, the public debate among practitioners, experts and researchers in this area has by no means reached a satisfactory conclusion. As previously addressed in the overview of modes of inquiry employed, the discourse on the relationships between higher education and the world of work seems to be distorted endemically by certain fallacies.

The first could be called *subordination fallacy*. Higher education is called upon to gear programs toward the presumed demands of the employment system. The frequently employed term "employability" stirred up controversial debates because it seemed to call for such a subordination (cf. the discussion in Vucasovic, 2007). We have argued that a term such as "professional relevance" could have avoided such a misunderstanding (see Teichler, 2009), if a misunderstanding exists. Any

call for such subordination underestimates the role of higher education in preparing graduates – beyond proficiency in the usual rules and tools of professional work – for indeterminate work tasks, competent handling of unexpected tasks and innovation in general.

Closely related to this is a second issue which could be called the *employers know best fallacy*. Of course, representatives of companies and other employer organizations have the best first-hand experience in this respect. However, this theory has many limitations: views could be shaped by the specific traditions of individual companies and the preoccupations of those in charge rather than by "demands"; the persons responsible for management and personnel matters could have different views from those who are active in other departments of a company; there could be an over-emphasis on short-term demands; and we have often noted an emphasis on current shortages of certain competences or on "personality", as a result of which higher education is accused of over-emphasizing the cognitive dimension of competences.

Third, we have noted a *mismatch and over-education avoidance fallacy*. Many actors and experts are calling for a close link between fields of study and occupational areas as well as for a close link between levels of educational attainment and related occupational hierarchies. Such an emphasis on a link which is as close as possible does not sufficiently take into consideration that a close link between educational awards and occupational categories is not always the result of a good link between competences and work tasks, but could be due to "credentialism", e.g., excessive rewards for formal educational attainments, the professional power of "closed" occupations and other not necessarily functional and meritocratic mechanisms. Second, there are imperfections in the planning and market-steering of demand and supply, and higher education is bound to be broader than the job requirements of individual positions; thus, higher education cannot provide the closest possible match, but must also prepare students for flexible adjustment. Third, those emphasizing a close match have underestimated the large proportion of graduates who are employed in areas that are not closely linked to their field of study and their level of educational attainment, but who considered their competences as useful for their work tasks and who become proactive agents of change in the world of work.

Fourth, we can note a *practice fallacy*. For example, internships are often hailed as the *non plus ultra* for linking learning and professional work, or the emphasis is placed predominantly on applied knowledge. Useful as these approaches may be, they are in part a response to the difficulty of understanding the links between abstract knowledge in the world of academia and professional problem-solving.

Fifth, we can observe a *general competence fallacy*. We often hear that specific knowledge rapidly becomes obsolete and that training in clear logical thinking will help to cope with various job tasks and the demand for further knowledge acquisition. Once again, useful as general competences may be, we cannot offset the need to acquire specific knowledge and related professional competences in order to cope with the challenging tasks of an "expert society".

These fallacies have been so pervasive not only in the public discourse, but also in research on the relationship between educational output and its subsequent use in the world of work or in other spheres of life, that all efforts to improve our knowledge base in these areas have to start with a demystification of the prevailing misconceptions and with a search for an appropriate balance. Operational improvement of the measurement of competences remains of limited value if it is not embedded within a convincing conceptual framework of desirable links between study and work.

POTENTIAL AND CHALLENGES OF GRADUATE SURVEYS

Currently, substantial efforts are underway to improve the knowledge basis on the relationships between higher education and work with the help of competence testing. The OECD initiative for AHELO as a functional equivalent to PISA is the best-known activity in this domain. We can also note national initiatives such as the promotion of research on competence measurement in Germany. The authors of this contribution share the view that these efforts are unlikely to achieve their ambitious aims and that the risk of generating pressure to conform in higher education is salient. However, irrespective of this sceptical view as regards competence testing, we can assume that graduate surveys will play an important role in the future in establishing the links between job requirements and the tasks of higher education. We argue that substantial research efforts are needed that aim to improve the quality of ratings of job requirements and graduates' competences within the framework of graduate surveys.

Admittedly, not all graduate surveys aim to measure job requirements and graduates' competences. In some countries, relatively simple large-scale national graduate surveys have been established. Information is collected on the distribution of graduates according to field of study, type of higher education institution, individual higher education institution, type of degree, gender, (employment) status, possibly economic sector, occupational category and possibly income. First-generation national graduate surveys were established decades ago in the U.K. and Japan which have continued to the present time.

These graduate surveys provide core descriptive information on the employment situation of graduates from higher education institutions. They raise public awareness, primarily of customary over-interpretations, e.g., studying humanities at university x leads to a highly successful career in sectors a, b and possibly c. Many economists believe that such datasets allow us to draw far-reaching conclusions on the utilization of competences (e.g., small income advantages indicating over-supply or under-utilization of competences).

Over the years, however, most graduate surveys have moved toward a more complex model in order to explain the relationship between higher education and the world of work. Some key issues inherent in restricted approaches and some directions for widening these approaches could be cited.

1. The proportion of graduates successfully making use of their competences in their job is underestimated if the typical occupational categories of graduates

and traditional income advantages are taken as the sole criteria. *Graduates' ratings of the links between their level of competence and their position and of the extent to which their competences are used* in their job can be included and graduate surveys can show a greater proportion of the productive links between study and work than the seemingly "objective" employment criteria outlined above.

2. The professional success of graduates from certain institutions, study programs or sectors of higher education may not be due to study, but may have to be attributed partly or completely to differences at the point of entry into higher education. The "value added" of less prestigious universities may be higher, as a look at the careers of graduates from prestigious universities suggests. *Data on the socio-biographic background of students and on prior schooling* can be collected in graduate surveys that help to control such effects. In principle, competence measures could be used at the beginning and end of study in order to measure the "value added".
3. The success of study may not be solely the result of the study conditions and provisions, but may also be shaped by the *students' study behavior*. In order to disentangle these effects, questions are useful in graduate surveys which ask respondents, for example, to characterize their major approaches as regards teaching and curricula, their choices and their study behavior, e.g., the time spent on learning and their learning styles.
4. Graduates may be less professionally successful than they could be on the basis of their competences because they handled the process of transition from study to employment less well than usual. Therefore, *questions regarding the search and recruitment process* may be raised in graduate surveys in order to establish the extent to which such "intervenient variables" come into play.
5. The extent of the professional success (or failure) of a graduate may be caused by a *misjudgement of her of his actual competences on the part of the employers*. They may have overrated the information stated in the graduate's credentials (e.g., the reputation of institutions, grades, etc.) – so-called "credentialism" – or recruitment interviews may not have sufficiently revealed his or her competences. Graduate surveys which ask graduates to rate their own competences at the point of graduation and their perceived job requirements help to measure such phenomena.
6. Not every graduate is driven in his or her study and work behavior by a set of motivations and values which corresponds to that of a "*homo economicus*" or "status seeker". *Questions regarding graduates' orientation* show that those seeking interesting work, opportunities to contribute to social change or another work-life balance far outnumber those matching the construct of a *homo economicus*.

In sum, readiness has grown to *design questionnaires for graduate surveys* in a way that ensures *sufficiently complex information* in order to exclude basic misunderstandings of the relationships between higher education and the world of work. This helps to measure "success" in transition, employment and work in various respects, to include a certain breadth of study conditions and provisions as

well as study behaviors that could be professionally relevant, and to cover the socio-biographic and schooling variables needed to avoid simplification, in terms of assumed causalities concerning the conditions, processes and impact of study. Thereby, the self-rating of competences and their comparison with perceived job requirements are valuable for examining both the weight of the various competences for subsequent employment and the extent to which other factors explain professional success, e.g., credentialism, misidentification of competences on the part of employers, misfortunes in the transition process and deliberate career choices that do not aim for the maximum professional use of the competences acquired in the course of study.

This thematic area has played a substantial role in the two major international comparative surveys on graduate employment and work. For example, in the CHEERS (Careers after Higher Education: A European Research Survey) survey, comprising around 36,000 persons from 12 countries who graduated in 1994/95 surveyed three to four years after graduation, five dimensions of competences and job requirements in addition to disciplinary and professional specialization were addressed in the questionnaire design with the help of 36 items: the ability to transmit systematic knowledge to work tasks; competences which are relevant for reflection, innovation and creativity; working style; socio-communication skills; motives and values (Teichler, 2007, p. 9). Two further classifications were developed in the course of the analysis: on the one hand, general-cognitive, systematic-operative, professionally knowledgeable, socio-reflexive and physiologically/manually skilled (Kellermann, 2007); on the other, knowledge, methodical skills, intelligence, socio-communication skills and organizational skills (Kiviven & Nurmi, 2007). In the REFLEX (The Flexible Professional in the Knowledge Society) survey which again comprised around 36,000 persons from 15 countries who graduated in 1999/2000 surveyed approximately five years later, many similar items were included, but a distinct classification was presented, between professional expertise, functional flexibility, innovation and knowledge management, the mobilization of human resources and international orientation (Allen & van der Velden, 2011).

The authors of the studies had developed these categories for the rating of competences and job requirements after careful study of the available research literature and obviously considered their approaches to be relatively ambitious. However, both studies state that it has remained difficult to define categories of competences in such a way that they are both theoretically satisfying and match the perceptions and understandings of the respondents.

Therefore, graduate surveys would profit enormously from accompanying research on the concepts and measurement of competences and the related demand for and utilization of competences. Cooperation is helpful with research activities striving for other modes of measurement (e.g., indicator development, test development), but certainly *research would be helpful which aims explicitly for the improvement of graduates' ratings in this domain*.

TRENDS AND FUTURE PROSPECTS

The scenery of assessing the competences of students and graduates and their possible relations to job requirements and tasks in other spheres of life is extremely varied – in terms of narrow vs. broad coverage, in the level of sophistication of the measurement and in the workload involved in the assessment.

This does not mean, however, that a move can be expected away from a gradual abandoning of relatively simple measurements toward predominantly complex measurements of competences. Instead, we can note that there is currently:

- The persistence of the vital role of higher education credentials in employers' recruitment of graduates, whereby the assessment of students' achievements by their teachers is the core of the information provided by the credentials;
- The persistence of the traditional teacher assessments of student achievement;
- The growing popularity of indicators which rely on readily available, simple aggregate measurements, as the "ranking" discourse shows;
- A spread of graduate surveys in which competences and job requirements are measured with the help of graduates' (self-)ratings;
- The persistence of the dominant final selection of job applicants by employers with the help of open or semi-structured interviews, occasionally accompanied by tests in specific areas (e.g., IT or a foreign language) or by testing behavior in semi-standardized simulations of work situations (e.g. "assessment centers");
- Increasing efforts to develop competence testing in this domain.

It should not come as a surprise to note the persistence of such a broad range of efforts in assessing competences and their links to work tasks.

First, it is by no means clear that highly complex measurements help to predict future performance. Testing at the point of entry into higher education continues to be viewed as controversial by experts as regards the greater predictive validity of achievement during the course of study in higher education, and there are sound reasons to believe that testing at the time of graduation produces even weaker results for predicting professional performance.

Second, the popularity of simple measures has increased: obviously, the desire and pressure to have easily readable data at hand is greater than the desire to enhance the quality of data.

Third, there is the issue of "value for money" as regards investment in elaborate measurements of competences; those in charge of recruiting graduates often spend less than five minutes assessing the documents provided by the candidates and less than one hour communicating with a candidate invited for final selection – this looks efficient and is also based on the assumption that more time invested would not lead to a substantial increase in predictive validity.

The scholars in charge of teaching and assessment at higher education institutions must assess the competences emphasized in their curricula and need feedback as to whether or not these competences have been enhanced by graduation and whether they are important in the graduates' subsequent work. This

can be realized through ratings, but not through tests in order to meet the specific profiles of individual study programs.

Finally, competence testing must be based on the belief that there is a broad common core of competences that are needed and must be strived for; as previously pointed out, this may be not only an illusion, but unduly homogeneous testing could create undue pressure to conform in higher education at a time when calls for diversity in higher education are increasing.

Therefore, a practice-oriented research program on the measurement of competences enhanced by higher education would be most valuable if it served the various needs for objectives to improve measurement in this domain, notably:

- The need for teachers in higher education to change their mode of measuring achievement in their classes if they want to move from knowledge-oriented teaching and learning toward competence-oriented teaching and learning;
- The needs for those in charge of recruiting graduates to enhance the quality of job interviews;
- The need for graduate surveys to increase the quality of self-rating of competences as well as their links to job requirements;
- Efforts to enhance the testing of competences in higher education.

Such a broad approach to improving the measurement of competences and enhancing our knowledge of the character, causes and consequences of competences would have three significant advantages over an approach to improve testing of competence. Our knowledge about the strengths and weaknesses of existing measurements of competences could grow substantially. Moreover, the potential of experts could be pooled in order to enhance the conceptual basis for measuring competences. Finally, a broad range of practical and relevant means of measuring competences could benefit and eventually improve.

REFERENCES

Allen, J., & van der Velden, R. (2011). *The flexible professional in the knowledge society: New challenges for higher education*. Dordrecht: Springer.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: The University of Chicago Press.

Banta, T. (2007). A warning on measuring learning outcomes. *Inside Higher Ed*. Retrieved from http://www.insidehighered.com/views/2007/01/26/banta.

Bennet, N., Dunne, E., & Carré, C. (2000). *Skills development in higher education and employment.* Buckingham: SRHE & Open University Press.

Cavalli, A.. (2007). *Quality assessment for higher education in Europe*. London: Portland Press.

Fröhlich, G. (2006). Informed peer review – Ausgleich der Fehler und Verzerrungen? In HRK (Hochschulrektorenkonferenz), (Ed.), *Von der Qualitätssicherung der Lehre zur Qualitätsentwicklung als Prinzip der Hochschulsteuerung* (pp. 193–204). Bonn: HRK.

Gehmlich, V. (2009). *Die Einführung eines Nationalen Qualifikationsrahmens in Deutschland (DQR). Untersuchung der Möglichkeiten für den Bereich des formalen Lernens.* Bonn/Berlin: BMBF.

Janson, K., Schomburg, H., & Teichler, U. (2009). *The professional value of ERASMUS mobility*. Bonn: Lemmens.

Kellermann, P. (2007). Acquired competences and job requirements. In Teichler, U. (Ed.), *Careers of university graduates* (pp. 115–129). Dordrecht: Springer.

Knight, P. T., & Yorke, M. (2002). Employability through the curriculum. *Tertiary Education and Management*, *8*, 261–276.

Knight, P. T., & Yorke, M. (2003). *Learning, curriculum, and employability in higher education*. London: RoutledgeFalmer.

Kivinen, O., & Nurmi, J. (2007). Job requirements and competences: Do qualifications matter? In Teichler, U. (Ed.), *Careers of university graduates* (pp. 131–142). Dordrecht: Springer.

Mittag, S. (2006). *Qualitätssicherung an Hochschulen. Eine Untersuchung zu den Folgen der Evaluation von Studium und Lehre*. Münster: Waxmann.

Schaeper, H., & Wolter, A. (2008). Hochschule und Arbeitsmarkt im Bologna-Prozess. Der Stellenwert von "Employability" und Schlüsselkompetenzen. *Zeitschrift für Erziehungswissenschaft, 11*, 607–625.

Schaeper, H. (2009). Development of competencies and teaching-learning arrangements in higher education: Findings from Germany. *Studies in Higher Education*, *34*, 677–697.

Schomburg, H., & Teichler, U. (2006). *Higher education and graduate employment in Europe: Results of graduates surveys from 12 countries*. Dordrecht: Springer.

Schomburg, H., & Teichler, U. (2011). *Employability and mobility of bachelor graduates in Europe: Key results of the Bologna Process*. Rotterdam: Sense Publishers.

Teichler, U. (2007). *Careers of university graduates*. Dordrecht: Springer.

Teichler, U. (2009). *Higher education and the world of work: Conceptual frameworks, comparative perspectives, empirical findings*. Rotterdam: Sense Publishers.

Vukasovic, M. (2007). Deconstructing and reconstructing employability. In E. Fromment, J. Kohler, L. Purser & L. Wilson (Eds.), *EUA Bologna handbook* (part 1.4–2). Berlin: Raabe.

Yorke, M. (2007). Employability in higher education. In E. Fromment, J. Kohler, L. Purser & L. Wilson (Eds.), *EUA Bologna handbook* (part 1.4-1). Berlin: Raabe.

Ulrich Teichler & Harald Schomburg
International Center for Higher Education Research,
University of Kassel, Germany

PART 4

COMMENTARY

JUDITH GULIKERS AND MARTIN MULDER

"MODELING AND MEASURING COMPETENCIES" CONFERENCE, BERLIN, 24–25 FEBRUARY, 2011

In February 2011, I visited Berlin in order to attend a special conference on "Modeling and Measuring Competencies in Higher Education", organized by the Humboldt University of Berlin and the Johannes Gutenberg University of Mainz. It was a special conference as it addressed this topic only over a period of two days and featured mainly invited (keynote) speakers. While the audience consisted mainly of German participants, leading researchers from universities and testing institutes in Germany, the U.S.A. and Australia presented their views and work. After three elaborate keynotes in the morning, the afternoon on both days consisted of panel discussions or a so-called "town hall meeting" in which three to five researchers introduced their work in 15 minutes, followed by a lively and interactive discussion. During lunchtime on Friday, there was a poster round in which 14 posters were presented which were mainly from Germany, with some from Finland. It was a very interesting and inspiring but also confronting experience. I would like to discuss two crucial controversies that were illuminated during this two-day meeting (the first day focused mainly on large-scale, high-stakes assessments while on the second day, we paid much more attention to the individual student in the assessment process):

1. Are we measuring the cognitive aspects of competencies or something more?
2. Curriculum validity versus professional/labor market validity.

These discussions, and the positions and frame of reference which researchers adopt in this respect (either large-scale (high-stakes) or focusing on the individual student), have, in my opinion, a large impact on both the modeling and measurement of competencies (or competencies), as well as on the impact such an assessment can or should have on the teaching and learning process.

THE FIRST DAY: LARGE-SCALE (HIGH-STAKES) MEASUREMENTS AT THE INSTITUTIONAL LEVEL

The openings words of Prof. Sigrid Blömeke informed the audience about a large German Ministry of Education and Research funding initiative called "Modeling and Measurement of Competencies in Higher Education". This initiative stimulates new, creative but also fundamental research, emphasizing *more evidence-based innovation* in teaching and learning in (higher) education. Ninety-four proposals were submitted from various disciplines, showing the interest in and relevance of this topic and the necessity of carrying out research in this field.

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 231–236.

After the opening speech, the first day was filled with contributions from German and North American researchers, which the exception of the keynote speaker, Karine Tremblay, the Senior Survey Manager of the "Assessment for Higher Education Learning Outcomes (AHELO)" international project of the Organisation for Economic Co-operation and Development (OECD). Overall, this day was characterized by the view of assessment as a large-scale, high-stakes undertaking, for the purpose of comparing (or even ranking) institutions at a national or international level. Probably because of this purpose, all assessments discussed this day were written tests, often containing multiple-choice formats, for gauging the cognitive aspects of competence. This issue was heavily discussed in the panel meeting in the afternoon during a discussion of four German projects. The majority of these research projects defined competence by its "narrow definition", described by Klieme and Leutner (2006) as "context specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in a specific domain". Blömeke adds to these cognitive aspects the importance of taking emotional/motivational aspects into account and therefore assesses competence by addressing not only students' cognition, but also their beliefs. The decision to deal with cognitive aspects only was defended by arguing that these elements can be measured objectively through written items, which is almost inevitable in large-scale assessments used to compare institutions. However, an additional argument was made for "curriculum validity": measuring those elements that are also present in the curriculum. The assessments were drawn up after consultations with faculty members from various higher education (HE) institutions about the content of their curricula. The involvement of the labor market was not an issue, as the labor market for HE graduates was argued to be too vague and too broad. The question which came immediately to mind was obvious:

> If the purpose is to innovate higher education (see the funding initiative), then assessing only those elements that are present in the current curriculum will not stimulate innovation in the curriculum, will it?

Prof. Richard Shavelson, a well-known American assessment specialist, offers us a way out of assessing cognitive aspects only in large-scale tests: he provides technical insight into assessing competencies through written test items that are also authentic performance assessment tasks and representative of competencies used in the real world. He argues that the starting point for developing such an assessment should be a careful description of the criterion behavioral domain in real life (see also Shavelson, 2010), and thus not the current curricula of HE institutions. His technical presentation was enforced by Roger Benjamin on the second day of the conference, who discussed the Collegiate Learning Assessment (CLA), which is used all over the U.S. in HE institutions (see also Benjamin, Chun, & Shavelson, 2009). This is a large-scale, high-stakes assessment of critical analysis and evaluation, problem-solving, persuasive writing and the mechanics of writing through an array of written (but relevant to real life) performance tasks scored by trained computer systems. The degree to which written tasks can be called performance tasks can be debated, but Benjamin convinced me that these

assessments are definitely more performance, real-life and competence-oriented in the HE context than I had understood from the examples I had heard on the first day of this conference. In addition to the focus on real-life performance tasks designed to assess generic competencies instead of disciplinary content, the CLA also stresses the necessity of holistic (instead of analytic or atomized) scoring. For each performance task, three integrative (or holistic) model answers are developed, characterizing a high, moderate or low performance. Perhaps even more important, from my point of view, was Benjamin's emphasis on using the CLA as an instrument for improving the teaching and learning process in HE institutions. The CLA feeds back at the institutional level, indicating the value added of a certain institution, and thus its curriculum or educational program. In this way, the CLA does not aim at curriculum validity (that is, assessing only what is taught in the curriculum), but to offer institutions a handle with which to improve their educational programs in order to address the real-life challenges that their graduates will face in the future. During the discussion, Blömeke addressed the issue of taking students' entry level of competence into account as an indicator of the value added of HE institutions. This issue was discussed further on day two by Prof. Spiel.

THE SECOND DAY: INDIVIDUAL STUDENTS AND COMPETENCE ASSESSMENT

The discussion outlined above focused mainly on large-scale, written assessments, with the aim of comparing, ranking and improving at the institutional level. Various speakers on the second day of the conference, however, strongly emphasized the need to examine the individual (student) level when talking about modeling and measuring competencies. Prof. Michaela Pfadenhauer (Germany) discussed the word "competence" from a sociological perspective, and built up an argument for what it means for the modeling and measurement of competencies. She advocated a broad definition of competence that includes its social aspects, which are inherently connected to the individual, his or her interpretation of the world and his or her feeling of responsibility for his or her actions. Competence, in Pfadenhauer's words, means that the actor combines "können, wollen und dürfen" (i.e., being able, willing and allowed to) in repeated and responsible actions in various situations. This implies that modeling and measuring competence cannot be done without a careful examination of the actor's actions and must include the actor's own perspective and reflections on his or her performance. If we follow this line of reasoning, Pfadenhauer argues for the impossibility of objectifying competence and using standardized measurements and assessments that do not involve a dialogue with the actor or an examination (through introspection or retrospection) of his/her actions. She takes this even further, arguing that the actor should have the decisive vote in determining whether or not he/she possesses a certain competence. An obvious point of discussion is raised, namely that various scientific studies have shown the unreliability of self-assessments. Pfadenhauer responded to this issue by saying:

> I am not arguing that you are on the safe side of competence assessment when you use self-assessments. I argue that we should reframe how we view and use self-assessments. The self-assessments that are often used in practice and research are not addressing competence as I see it, namely as a personal and responsible implementation of action. If you see competence like this, then self-assessment through introspection based on various performances is inevitable.

The problem, she states, is that this perspective on competence development and measurement requires the alteration of both the (German) education and accountability systems, which is not happening. This problem has been recognized by other researchers in various countries (e.g., Knight (UK) & Kvale (Denmark), 2007) and it is safe to say that these problems are being faced in the Dutch movement towards competence-based assessment and education as well.

Prof. Sadler (Australia) agrees with Pfadenhauer on the importance of involving students in the assessment and development of their competencies, and of engaging in a dialogue with students on the meaning of competence and being competent within a certain area or task. He also strongly stresses the need for a holistic approach to competence assessment (see also Sadler, 2009):

> Decomposing competence into manageable (or even atomized) components in order to facilitate judgements may have some interim value in certain contexts, but the act of decomposition can obscure how a practitioner would work the various bits together to form a coherent whole.

In this holistic approach and regarding the need for competence assessment to relate to the professional world, Sadler agrees with the arguments of Shavelson and Benjamin concerning the CLA. However, contrary to Pfadenhauer, who places the main responsibility for competence assessment on the actor, and the CLA, which uses written performance tasks scored by a computer, he elaborates on the crucial role of complex performance tasks that can only by assessed by knowledgeable judges who have calibrated their views on what competence means. This opens up a new discussion on the pros and cons as well as the (im)possibilities of human judges in complex, open-ended assessment tasks. A link is made to assessment quality, the heightened focus on the assessment process and procedure and the professionalism of human judges (i.e., teachers).

Prof. Christiane Spiel (Austria; second day of the conference) adopts a somewhat different viewpoint, namely that of program evaluation and evaluation research. Her talk combined the institutional perspective with the student-centered perspective of competence assessment discussed above, by exploring the issue of evaluating the value added of a HE curriculum for the development of student competencies. She emphasized using outcome evaluations that assess the extent to which programs achieve their goals, which seems obvious, but she stresses that goals should address both generic and domain-specific competencies instead of disciplinary knowledge alone. In order to examine the value added, evaluation systems should compare the competencies of graduates and freshmen (baseline data), as well as comparing graduates' competencies with defined graduates profiles.

Prof. van der Velden (the Netherlands) reinforced this value added discussion by exploring large-scale (but not high-stakes) graduate surveys. He argues that these surveys should measure both cognitive and non-cognitive aspects of students' competencies, as developed through a certain curriculum containing specific characteristics, and that they should relate these aspects to the outcomes which students achieve in the labor market. Then, these data can provide useful information on: (1) the types of output (i.e., competencies or something else) on which a HE institution should focus, based on the requirements of the dynamic world of work; and (2) what HE institutions can do in order to better foster these outputs. The results show that the HE graduate labor market strongly emphasizes professional expertise, including domain-specific as well as generic academic competencies. The latter category is especially important for equipping graduates to deal with uncertainty in society and to become functionally flexible in the labor market. While van der Velden bases his arguments on large-scale written survey data which combine and compare HE institutions, he stresses that at the institutional level, the development and measurement of competence requires assessment methods other than written (specifically multiple-choice) tests, as these will not stimulate students to develop generic academic competencies, nor will they stimulate teachers to educate their students in generic competence development. This reveals what van der Velden calls "the assessment gap": most HE institutions still rely mainly on multiple-choice tests of disciplinary skills instead of (also) using other types of assessment and assessing professional expertise in terms of both domain-specific and generic academic competencies which are representative of the labor market.

Thus, there is a great deal to be done in order to improve the modeling and measurement of competencies in HE that drive institutions to innovate towards professional validity. This conference showed that various countries have different perspectives on competences (or competencies), their modelling and the quality and form of their measurement. This conference showed that there are many sides to competence assessment: a scale-related side; a cognitive, a performance and a social/beliefs-related side; an individual student and an institutional side; a curricular and a labor market side; a dialectical/reflective versus a standardized/one-size-fits-all side; and moreover, a summative, comparative and ranking purpose and a formative purpose which feeds back to the institution, program or student level in order to stimulate improvement and development. However, independent of their viewpoint or purpose, all participants agreed that competence assessment is increasing in relevance and importance, both at the national and international level, as shown by the large funding initiative being run by the host country's Ministry of Education. However, it is fraught with controversies and difficulties, as it challenges the way in which we define the quality of (higher) education, how (higher) education institutions change, or should change, to address these challenges, and how institutional evaluation systems (as Spiel was talking about), as well as external accountability bodies, examine, grant and stimulate the improvement of the quality of educational curricula, teaching and assessments.

REFERENCES

Benjamin, R., Chun, M., & Shavelson, R. (2009). *Holistic tests in a sub-score world: The diagnostic logic of the collegiate learning assessment (CLA)*. Retrieved from http://www.collegiatelearningassessment.org/.

Knight, P. (2007). Grading, classifying and future learning. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment in higher education: Learning for the long term* (pp. 72–86). New York: Routledge.

Kvale, S. (2007). Contradictions of assessment for learning in institutions of higher learning. In D. Boud & N. Falchikov (Eds.), *Rethinking assesment in higher education: Learning for the long term* (pp. 57–71). New York: Routledge.

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 45–63). -Springer- Science + Business Media.

Shavelson, R. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training, 2*(1), 41–64.

Judith Gulikers & Martin Mulder
Chair Group Education and Competence Studies
Wageningen University, the Netherlands