PROFESSIONAL AND VET LEARNING

Modeling and Measuring Competencies in Higher Education

Tasks and Challenges

Sigrid Blömeke, Olga Zlatkin-Troitschanskaia, Christiane Kuhn and Judith Fege (Eds.)

SensePublishers

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

PART 3: LONG-TERM OUTCOMES

PART 4: COMMENTARY

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.

INTERSUBJECTIVITY AND COMPETENT JUDGMENTS

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_s^2	817.47	20.90
Task (*t*)	σ_t^2	0[a]	0
Judge (*j*)	σ_j^2	62.50	
1.60			
s × *t*	σ_{st}^2	671.42	17.10
s × *j*	σ_{sj}^2	62.18	
1.60			
t × *j*	σ_{tj}^2	0[a]	0
s × *t* × *j*, *e*	$\sigma_{stj,e}^2$	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