Modeling and Measuring
Competencies in Higher Education
Http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php

KoKoHs Working Papers

No. 6

Christiane Kuhn, Miriam Toepper, & Olga Zlatkin-Troitschanskaia

Current International State and Future Perspectives on Competence Assessment in Higher Education

Report from the KoKoHs Affiliated Group Meeting at the AERA Conference on April 4, 2014 in Philadelphia (USA)

SPONSORED BY THE

Federal Ministry
of Education
and Research

Johannes Gutenberg University Mainz

Humboldt University of Berlin

KoKoHs Working Papers from the BMBF-funded research initiative "Modeling and Measuring Competencies in Higher Education"

The *KoKoHs Working Papers* series publishes articles from the funding initiative "Modeling and Measuring Competencies in Higher Education (KoKoHs)". These may be conceptual papers or preliminary results intended for rapid dissemination or public discussion. Publication in *KoKoHs Working Papers* does not preclude publication of the texts elsewhere. The responsibility for the content lies with the authors. The content does not necessarily reflect the views of the publishers of *KoKoHs Working Papers*.

Editors:

Prof. Dr. Sigrid Blömeke
Humboldt University of Berlin
Faculty of Arts and Humanities IV
Department of Education Studies
Chair of Instructional Research
Unter den Linden 6
D-10099 Berlin

Prof. Dr. Anand Pant
Humboldt University of Berlin
Faculty of Humanities and Social Science
Department of Education Studies
Unter den Linden 6
D-10099 Berlin

Prof. Dr. Olga Zlatkin-Troitschanskaia
Johannes Gutenberg University Mainz
Department 03: Law, Management and Economics
Chair of Business Education I
Jakob Welder-Weg 9
D-55099 Mainz

Contact:

miriam.schaffer@uni-mainz.de
corinna.lautenbach@hu-berlin.de

The research initiative is funded by the German Federal Ministry of Education and Research under grant no. 01PK11100A and 01PK11100B.

The *KoKoHs Working Papers* are also available for download:

http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php

Current International State and Future Perspectives on Competence Assessment in Higher Education

Report from the KoKoHs Affiliated Group Meeting at the AERA Conference on April 4, 2014 in Philadelphia (USA)

Christiane Kuhn, Miriam Toepper, & Olga Zlatkin-Toitschanskaia

Contact:
miriam.schaffer@uni-mainz.de

Bibliographical references:

Kuhn, C., Toepper, M. & Zlatkin-Troitschanskaia, O. (2014). Current International State and Future Perspectives on Competence Assessment in Higher Education – Report from the KoKoHs Affiliated Group Meeting at the AERA Conference on April 4, 2014 in Philadelphia (USA) (KoKoHs Working Papers, 6). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Current International State and Future Perspectives on Competence Assessment in Higher Education - Report from the KoKoHs Affiliated Group Meeting at the AERA Conference on April 4, 2014 in Philadelphia (USA)

Abstract:
The research program "Modeling and Measuring Competencies in Higher Education (KoKoHs)", which is funded by the Federal Ministry of Education and Research (BMBF) aims at a systematic and internationally compatible research on competence development and assessment in higher education in Germany. To meet this challenge a KoKoHs Affiliated Group Meeting was held at the AERA Conference on April 4th, 2014 in Philadelphia. Theoretical and methodological tasks and challenges of modeling and measuring competencies in higher education were discussed by KoKoHs project members and international cooperation partners. The present working paper documents insights into the meeting, which included talks and discussions on measurement and research methodology, generic competencies and teacher training in STEM fields.

Keywords:
AERA Annual Meeting Conference, Modeling and Measuring Competencies

Table of Contents

Section I:
Measurement and Research Methodology

Section II:

Generic Competencies in Higher Education

Section III:

Teacher Training in STEM Fields

Concluding Commentary

Summary

Olga Zlatkin-Troitschanskaia, Johannes Gutenberg University Mainz, Germany

Welcoming Speech

Welcome to the meeting on "Theoretical and Methodological Tasks and Challenges of Modeling and Measuring Competencies in Higher Education – Current State and Future Perspectives on Competence Assessment"

Overview of Research Context

- Competence-oriented learning and teaching in higher education highly relevant topic due to Bologna reform
- Competencies formally included in all study and exam regulations and tested accordingly
- Need for valid information on learning success in tertiary education as a basis for sustainable development measures
- Is such an assessment possible at all?
- Little empirical groundwork on learning success in higher education
- Scientific approaches to competence-orientation in higher education had to be developed
- International research experiences were considered (e.g., AHELO, ETS, CAE, ACER)
- Need for theoretically founded competence models and valid testing methods

KoKoHs Program – Background

- "Modeling and Measuring Competencies in Higher Education" (KoKoHs)
- Funded by the Federal Ministry of Education and Research (BMBF)
- First phase 2011 – 2015
- Total budget approx. 15 million euros

KoKoHs Program – Purpose and Aims

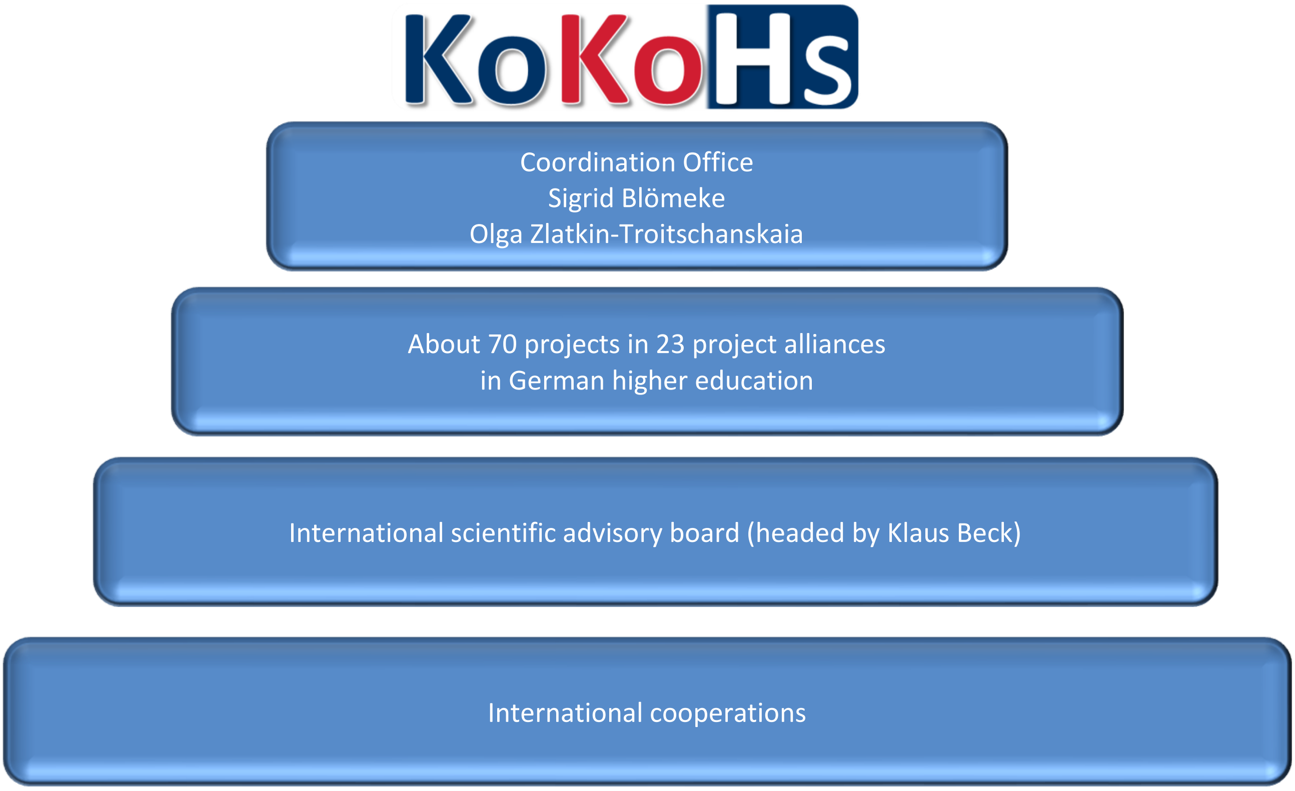
Purpose

- **Fundamental, systematic, and internationally compatible research on competence development and assessment in higher education in Germany**

Aims

- **Model** domain-specific/generic **competencies** in selected subjects (while taking into account the specific curricular and job-related features)
- Transform the theoretical models into suitable **measuring instruments**
- **Validate** test score interpretations

KoKoHs Program – Structure



KoKoHs

Coordination Office
Sigrid Blömeke
Olga Zlatkin-Troitschanskaia

About 70 projects in 23 project alliances
in German higher education

International scientific advisory board (headed by Klaus Beck)

International cooperations

International Cooperation Partners

- Centro Nacional de Evaluación para la Educación Superior (CENEVAL), Mexico – Rafael Vidal Uribe
- Council for Aid to Education (CAE), (New York) USA – Roger Benjamin, Doris Zahner, Raffaela Wolf
- Educational Testing Service (ETS), (Princeton) USA – Tom Van Essen, Ross E. Markle
- Griffith University, Australia – Royce Sadler
- Michigan State University, (East Lansing) USA – Alicia Alonzo
- OECD (AHELO-Feasibility Study), France – Karine Tremblay

- Research Center for Education and the Labour Market, Netherlands – Rolf van der Velden
- Stanford University, USA – Lee Shulman
- Stanford University & SK Partners, (Stanford) USA – Richard Shavelson
- University Luxembourg, Luxembourg – Sabine Krolak-Schwerdt
- University of Colorado, (Boulder) USA – Edward W. Wiley
- University of Illinois at Chicago, USA – James W. Pellegrino
- University of Massachusetts Amherst, (Massachusetts) USA – Ronald K. Hambleton
- University of St. Gallen, Switzerland – Christoph Metzger
- University of Twente, Netherlands – Jean-Paul Fox, Marieke van Geel
- University of West Georgia, (Carrollton) USA – Li Cao
- Vanderbilt University, USA – David Lubinski, Camilla Benbow
- University of California, USA – Mark Wilson

International Cooperation Partners and KoKoHs Project Members present at today´s meeting

Presenters

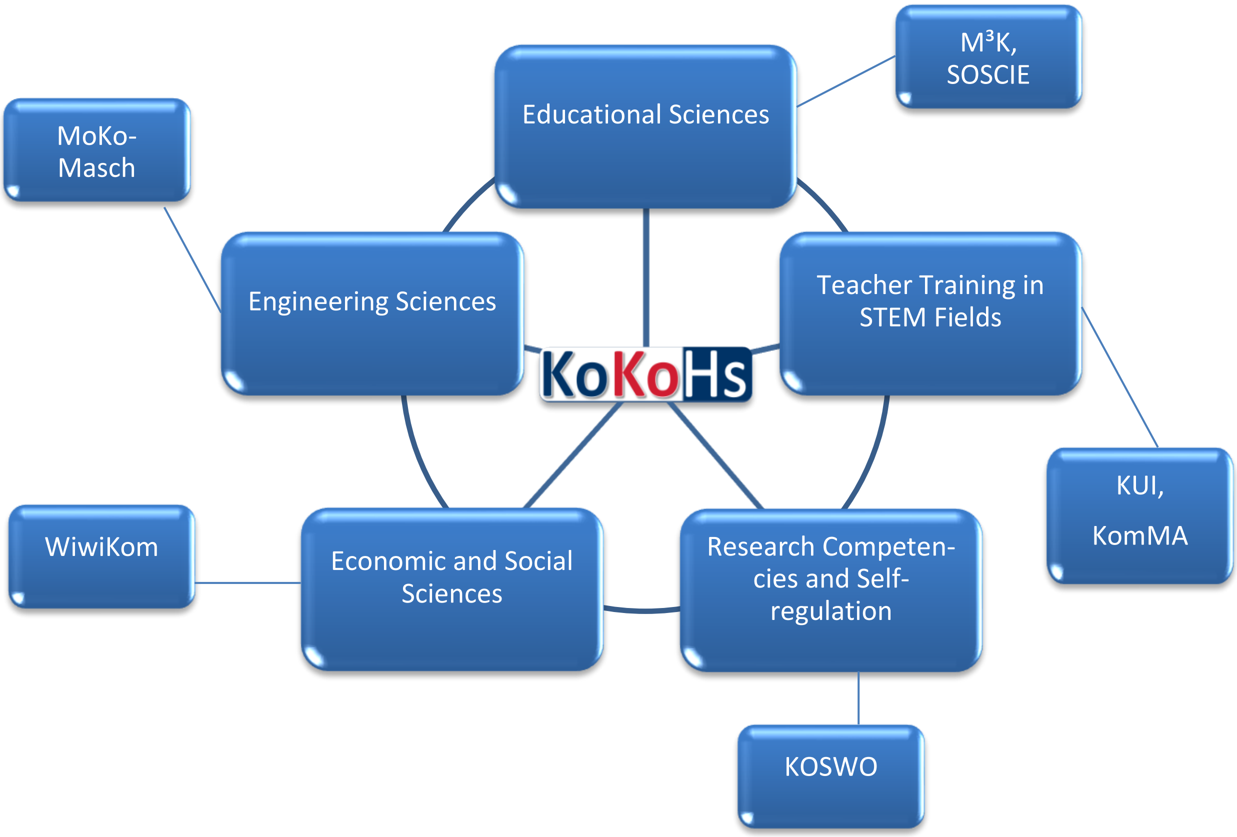
- University of West Georgia, USA – **Li Cao**
  ("Addressing Ecological Validity in Modeling and Measuring Competencies in Higher Education")
- Friedrich Schiller University Jena, Germany – **Linda Gräfe & Andreas Frey**
  ("Item Response Theory Based University Exams (MoKoMasch)")
- Johannes Gutenberg University Mainz, Germany - **Susanne Schmidt, Manuel Förster & Olga Zlatkin-Troitschanskaia**

  ("A Multilevel Analysis of Differences in the Economic Content Knowledge of University Stu dents in Germany with Individual and Contextual Covariates (WiwiKom)")
- University of Twente, Netherlands – **Marieke van Geel**
  ("The Effects of a School Wide Data-Based Decision Making Intervention on Student Achievement Growth in Dutch Primary Schools")
- Council for Aid to Education (CAE), (New York) USA – **Doris Zahner & Raffaela Wolf**
  ("A Case Study of an International Performance-Based Assessment of Critical Thinking Skills")
- Bielefeld University, Germany - E**lisabeth Marie Schmidt**
  ("Useful Strategies in Dealing With Primary Scientific Literature: An Expert-Novice Comparison (KOSWO)")

- Humboldt University Berlin, Germany – **Sigrid Blömeke**
  ("Effects of Opportunities to Learn on the Mathematics Pedagogical Content Knowledge of Prospective Kindergarten Teachers (KomMa)")
- Silke Grafe, University of Würzburg, Germany – **Silke Grafe**
  University of Bremen, Germany – **Andreas Breiter**
  ("Modeling and Measuring Pedagogical Media Competencies of Pre-Service Teachers (M^3K)")
- University of Paderborn, Germany – **Elena Bender & Niclas Schaper**
  ("Modeling Competences of Teaching Computer Science in German Schools at High School Level - Theoretical Framework, Curriculum Analysis and Critical Incident Based Expert Interviews (KUI)")

Discussants

- Educational Testing Service (ETS), (Princeton) USA – **Ross E. Markle**
- Michigan State University, (East Lansing) USA **– Alicia Alonzo**
- Stanford University, USA – **Lee Shulman**
- Stanford University & SK Partners, (Stanford) USA – **Richard Shavelson**
- University of Illinois at Chicago, USA – **James W. Pellegrino**
- University of Massachusetts Amherst, (Massachusetts) USA – **Ronald K. Hambleton**
- University of Twente, Netherlands – **Jean-Paul Fox**
- University of Fribourg, Switzerland **– Fritz Oser**
- University of Mainz, Germany **– Klaus Beck**
- University of West Georgia, USA **– Li Cao**


M³K,
SOSCIE
Educational Sciences
MoKo-
Masch
Engineering Sciences
KoKoHs
Teacher Training in STEM Fields
KUI,
KomMA
WiwiKom
Economic and Social Sciences
Research Competencies and Self-regulation
KOSWO

KoKoHs Program – Concept of Competence

Weinert (2001) defines competencies as

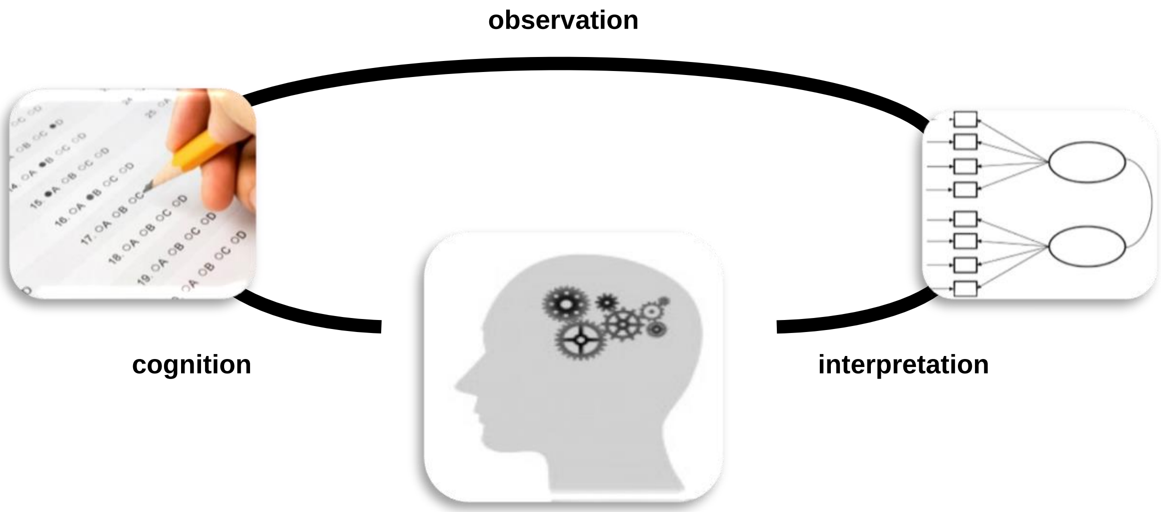
> “cognitive abilities and skills that individuals possess or acquire in order to solve certain problems as well as the aligned motivational, volitional and social dispositions and skills to apply the solutions in different situations successfully and responsibly“ (pp. 27-28).

- Holistic view
- However, limitations were necessary for practical reasons. Focus on cognitive abilities and skills.

KoKoHs Program – Study Design

“Assessment Triangle” by Pellegrino, Chudowsky & Glaser (2001)

> “a model of student *cognition* and learning in the domain, a set of beliefs about the kinds of *observations* that will provide evidence of students’ competencies, and an *interpretation* process for making sense of the evidence” (p. 44).



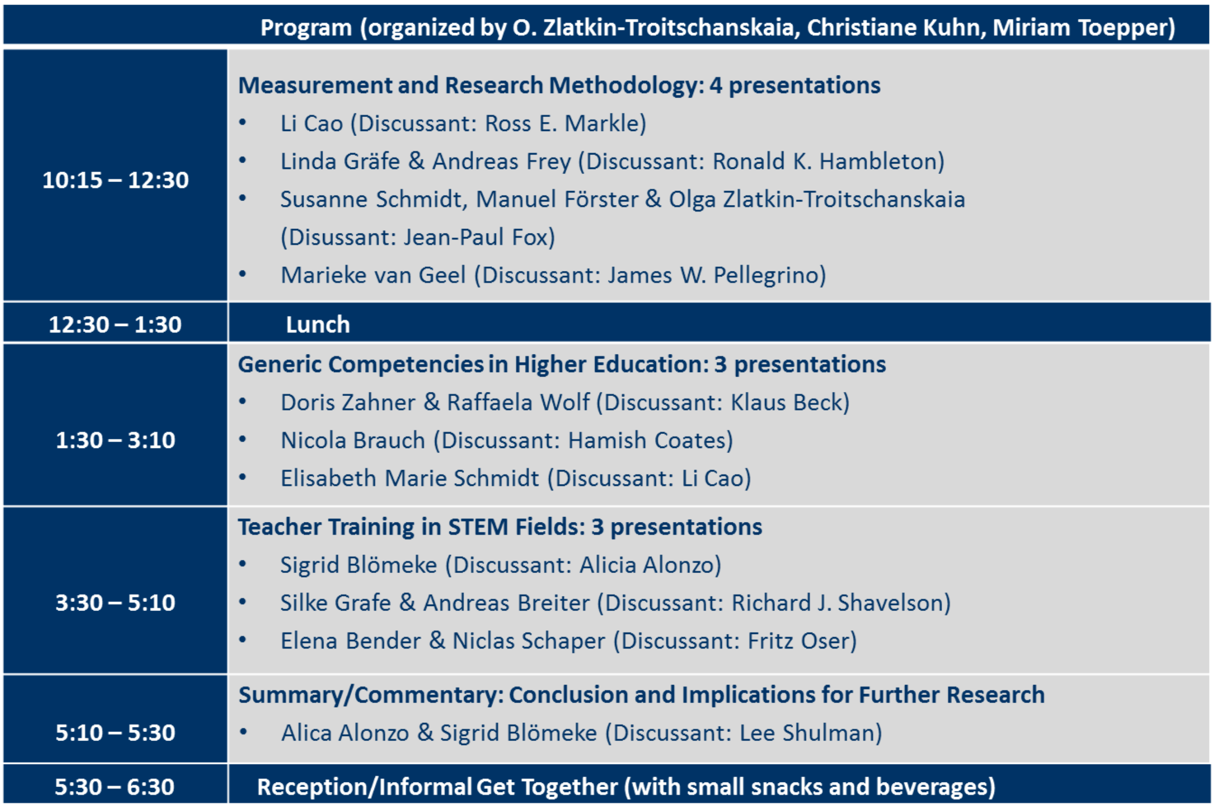
(Pellegrino et al., 2001)

Challenges in Competence Measuring

Measuring competence means

- designing or adapting items systematically
- taking into account framework conditions (time, method, format)
- analyzing data with complex IRT-based methods
- confirming psychometric quality criteria

Today's Objectives



	Program (organized by O. Zlatkin-Troitschanskaia, Christiane Kuhn, Miriam Toepper)
10:15 – 12:30	**Measurement and Research Methodology: 4 presentations** • Li Cao (Discussant: Ross E. Markle) • Linda Gräfe & Andreas Frey (Discussant: Ronald K. Hambleton) • Susanne Schmidt, Manuel Förster & Olga Zlatkin-Troitschanskaia (Disussant: Jean-Paul Fox) • Marieke van Geel (Discussant: James W. Pellegrino)
12:30 – 1:30	Lunch
1:30 – 3:10	**Generic Competencies in Higher Education: 3 presentations** • Doris Zahner & Raffaela Wolf (Discussant: Klaus Beck) • Nicola Brauch (Discussant: Hamish Coates) • Elisabeth Marie Schmidt (Discussant: Li Cao)
3:30 – 5:10	**Teacher Training in STEM Fields: 3 presentations** • Sigrid Blömeke (Discussant: Alicia Alonzo) • Silke Grafe & Andreas Breiter (Discussant: Richard J. Shavelson) • Elena Bender & Niclas Schaper (Discussant: Fritz Oser)
5:10 – 5:30	**Summary/Commentary: Conclusion and Implications for Further Research** • Alica Alonzo & Sigrid Blömeke (Discussant: Lee Shulman)
5:30 – 6:30	Reception/Informal Get Together (with small snacks and beverages)

http://www.kompetenzen-im-hochschulsektor.de

References

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C. & Fege, J. (2013). Modeling and Measuring Competencies in Higher Education: Tasks and Challenges. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education* (pp. 1-12). Rotterdam: Sense Publishers.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C. & Fege, J. (Eds.) (2013). *Modeling and Measuring Competencies in Higher Education*. Rotterdam: Sense Publishers.

Kuhn, C. & Zlatkin-Troitschanskaia, O. (2011). *Assessment of Competencies among University Students and Graduates – Analyzing the State of Research and Perspectives.* (Working paper of business education, 59). Mainz: Johannes Gutenberg University Mainz.

Schaffer, M., Zlatkin-Troitschanskaia, O., Kuhn, C., Schmidt, S. & Brückner, S. (Eds.) (2013). *International Colloquium for Young Researchers from 14th till 16th November 2013 in Mainz – Review and Impressions.* (KoKoHs Working Papers, 5). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Section I:

Measurement and Research Methodology

Li Cao, University of West Georgia, USA

Edith Braun, University of Kassel, Germany

Addressing Ecological Validity in Modeling and Measuring Competencies in Higher Education (KoKoHs)

Acknowledgement: We would like to express sincere gratitude to Dr. Ross Markle and Dr. Richard Shavelson for their valuable comments on the early version of this paper.

As a sequel, this concept paper follows up an earlier discussion (Cao, 2013) about the prospects and challenges in modeling and measuring competencies in higher education (KoKoHs). KoKoHs is a research program which is funded by the German Federal Ministry of Education and Research. The program aims at developing models and tests to measure competences in higher education. There are 23 research projects conducted by 220 researchers from over 70 universities and colleges across Germany. These projects are coordinated by Dr. Prof. Blömeke at Humboldt Universität Berlin and Dr. Prof. Zlatkin-Troitschanskaia at Universität Mainz. Both papers contribute to the discussion of status and current challenges in modeling and measuring competencies in higher education (Blömeke, Zlatkin-Troitschanskaia, Kuhn & Fege, 2013). The present paper takes a pragmatic perspective and addresses ecological validity in the context of KoKoHs. The purpose is to offer ecological validity as a means to assess the efficacy of education programs in developing competencies that meet the demands of job market, including workforce. Addressing ecological validity can also offer invaluable information for curriculum improvement of vocational training in higher education.

The ultimate goal of higher education is to educate subsequent generations and develop their competencies to meet challenges of the 21st century as productive citizens. As in many other nations, the value of higher education is greatly appreciated in Germany. Also like other nations, however, German higher education is facing unprecedented challenges with an increase of student enrollment in higher education of about 30% in 1990, 46.2% in 2010, 49.1% in 2014, and a projected 53.2% in 2020 (International Futures, No Date). This trend of increased enrollment in higher education is politically welcomed because it provides an increased access to higher education to a broader range of students (OECD, 2013). However, the German higher education system has met a hard time in adopting appropriate methods of instruction and assessing outcomes of student learning so as to maintain high quality teaching that the German higher education has traditionally developed. Furthermore, recent development makes it more clear that higher education system needs to prepare a broader range of students for professional occupations (Felstead et al., 2007; Peterson et al., 2001).

In addition, it has become much more salient for many nations recently that a high percentage of university graduates were unable to find employment upon graduation while industrial enterprises struggled to find a qualified workforce (Markle, Olivera-Aguilar, Jackson, Noeth & Robbins, 2013). Politically, this issue can be attributed to poor education policy and procedures and a flawed education system that failed to connect competencies of university graduates with expectations of employers. Methodologically, this is due to a lack of communication and misalignment of the expectations between education and workplaces. The KoKoHs program is one of the efforts to address this serious issue by focusing on developing competencies of university students so that they can meet the challenges at workplaces. It should be pointed out that higher education may never be able to completely meet the demands of employers. However, it is the duty of higher education to strive for preparing their graduates for labor market, more than what the Germany university system might have done so far. An immediate challenge for the KoKoHs program is to identify competencies that can inform curriculum development and meet the demands on the labor market simultaneously. In a sense, the KoKoHs program represents a trend towards more ecologically-sensitive service delivery practices within the assessment literature across many fields (e.g., Cleary, 2009; DiBenedetto & Zimmerman, 2013; Guthrie, Wigfield & Perencevich, 2004; Kitsantas & Zimmerman, 2002; Leiman & Stiles, 2001; Reschly, 2008; Schmitz & Wiese, 2006). In this paper, we propose that addressing ecological validity might serve as a specific means to address this pressing issue.

Defining Ecological Validity

As many other constructs in education research, the concept of ecological validity has evolved over the course of its development. Ecological validity is used in different ways and “is often confused with external validity” (Shadish, Cook & Campbell, 2002, p. 37). At its origin, Brunswik (1943, 1956) conceived the term ecological validity through his investigations of the organism-environment interactions. Instead of following Wundt’s (1874/1999), one of the founders of experimental psychology, suggestion to eliminate the messy surface features of the environment through the use of experiments, Brunswik (1943, 1956) proposed an ecological approach to psychological observations by sampling widely the environments within which particular "proximal" tasks are embedded. Brunswik's overall goal was to prevent psychology from being restricted to artificially isolated proximal or peripheral circumstances that are not representative of the "larger patterns of life." In particular, Brunswik (1943, 1956) argued whether it is possible to strip the phenomenon of all its accessory conditions, but whether it is necessary and even appropriate to do so if we can. Instead, he suggested an ecological approach that allows understanding of the organism’s adaptation to the confusing concatenation of events that disguises the regularities of its interactions with the world.

He higlighted that “proper sampling of situations and problems may in the end be more important than proper sampling of subjects, considering the fact that individuals are probably on the whole much more alike than are situations among one another” (Brunswik, 1956, p. 39).

In order to avoid this problem, Brunswik (1956) suggested that situations, or tasks, rather than people, should be considered the basic units of psychological analysis. In particular, these situations or tasks must be "carefully drawn from the universe of the requirements a person happens to face in his commerce with the physical and social environment" (p. 263). To illustrate his approach, Brunswik studied size constancy by accompanying an individual who was interrupted frequently in the course of her normal daily activities and asked to estimate the size of some object she had just been looking at. This person's size estimates correlated highly with physical size of the objects and not with their retinal image size. Brunswik claimed that this result "possesses a certain generality with regard to normal life conditions" (p. 265) (No date, MIT Website http://ai.ato.ms/MITECS/Entry/cole2.html).

Along a similar line, Bronfenbrenner (1979) approached the concept of ecological validity from developmental psychology. He defined ecological validity as “the extent to which the environment experienced by the subjects in a scientific investigation has the properties it is supposed or assumed to have by the investigator” (p. 29). In this view, Bronfenbrenner highlighted the pivotal role of researcher in establishing ecological validity, that is, to ensure accordance of the properties and outcomes of the education interventions with those at workplace. Shadish, Cook, and Campbell (2002) even viewed ecological validity more “as a method that calls for research with samples of settings and participants that reflect the ecology of application” than as a separate validity type (p. 37).

Since Brunswik’s ground breaking work, more attention has been drawn to the importance of generalizing from the particular circumstances of research investigations to wider ecological constraints under which each individual functions outside the laboratory. The term *ecological validity* has typically been used interchangeably to designate the external validity of research designs (Araújo, Davids & Passos, 2007). Ecological validity is the degree to which the behaviors observed and recorded in a study reflect the behaviors that actually occur in natural settings. Ecological validity is associated with "generalizability," that is, the extent to which the findings from a study realistically mimic (or extend to) activities and behaviors in life. The control created by the laboratory setting can potentially alter ecological validity (Gall, Gall & Borc, 2003; Walker, 2012). If the treatment effects can be obtained only under a limited set of conditions or only by the original researcher, the experimental findings are said to have low ecological validity.

Threats to Ecological Validity

We mentioned before that ecological validity is often confused with external validity. In fact, ecological validity has been defined in some cases as a subset of external validity. For instance, building on Campbell and Stanley's (1963) seminal work on internal validity, Bracht and Glass (1968) defined external validity as "the extent and manner in which the results of an experiment can be generalized to different subjects, settings, experimenters, and, possibly, tests" (p. 438). In particular, Bracht and Glass (1968) elaborated on the threats to two types of external validity: population validity and ecological validity (Table 1). The threats to population validity include those dealing with generalizations to populations of persons (What population of subjects can be expected to behave in the same way as did the sample experimental subjects?). These threats to ecological validity include those dealing with the "environment" of the experiment (Under what conditions, i.e., settings, treatments, experimenters, dependent variables, etc., can the same results be expected?).

According to Bracht and Glass (1968), external validity consisted of subcategorizes: population validity and ecological validity (Table 1). More specifically Table 1 shows, there are two treats to population validity: (1) experimentally accessible population vs. target population and (2) interaction of personological variables and treatment Effects. The former threat concerns with the generalization from the sample to the target population. The latter threat concerns with the interaction effect of the characteristics of participants with the intervention. Both threats focus on participants and deal with generalization of the results from the sample to the target population. Unlike many typical experimental studies, the two threats to population validity apply to the KoKoHs programs in two particular ways. First, in the KoKoHs program students are viewed as participants who go through an educational program, then graduate, and move to work in the target workplace. The students, graduates, and workforce are the same individuals. In this case, our sample of university students is in fact the population itself and the generalization from the sample to the population is not a concern. The threat to the generalization from the experimentally accessible population to the target population does not exist. Second, since the sample and the population in the KoKoHs program consist of the same individuals, their personological variables, such as ability, personality, motivation, anxiety, stress, and depression, etc. fall into the within individual factors which would have much a less degree of variation than those of the between individual factors. It is the interaction effects of different types of personological variables with the education intervention programs that may be of interest for research for personal and professional development.

Table 1. Factors affecting external validity: Reasons why inferences about how study results would hold over variations in persons, settings, treatments, and outcomes may be incorrect

Population Validity		
A	Experimentally Accessible Population vs. Target Population	Generalization requires a thorough knowledge of the characteristics of both the population accessible to the experimenter and the total population of the target. The results of an experiment might apply only for the population from whom the experimental subjects were selected and not for the total target population.
B	Interaction of Personological Variables and Treatment Effects	The superiority of one experimental treatment over another is reversed when subjects at a different level of some variable descriptive of persons are exposed to the treatments.
Ecological Validity		
A	Describing the Independent Variable Explicitly:	Generalization and replication of the experimental results presuppose a complete knowledge of all aspects of the treatment and experimental setting.
B	Multiple-Treatment Interference:	When two or more treatments are administered consecutively to the same persons within the same or different studies, it is difficult and sometimes impossible to ascertain the cause of the experimental results or to generalize the results to settings in which only one treatment is present.
C	Hawthorne Effect:	A subject's behavior may be influenced partly by his perception of the experiment and how he should respond to the experimental stimuli. His awareness of participating in an experiment may precipitate behavior which would not occur in a setting which is not perceived as experimental.
D	Novelty and Disruption Effects:	The experimental results may be due partly to the enthusiasm or disruption generated by the newness of the treatment. The effect of some new program in a setting where change is common may be quite different from the effect in a setting where very few changes have been experienced.
E	Experimenter Effect:	The behavior of the subjects may be un-intentionally influenced by certain characteristics or behaviors of the experimenter. The expectations of the experimenter may also bias the administration of the treatment and the observation of the subjects' behavior.

F	Pretest Sensitization:	When a pretest has been administered, the experimental results may partly be a result of the sensitization to the content of the treatment. The results of the experiment might not apply to a second group of persons who were not pre-tested.
G	Post-test Sensitization:	Treatment effects may be latent or in-complete and appear only when a post-experimental test is administered.
H	Interaction of History and Treatment Effects:	The results may be unique because of "extraneous" events occurring at the time of the experiment.
I	Measurement of the Dependent Variable:	Generalization of results depends on the identification of the dependent variables and the selection of instruments to measure these variables.
J	Interaction of Time of Measurement and Treatment Effects:	Measurement of the dependent variable at two different times may produce different results. A treatment effect which is observed immediately after the administration of the treatment may not be observed at some later time, and vice versa.

Source: Based on Bracht, G. H. & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437-474.

As discussed above, external validity is related to ecological validity in the context of KoKoHs, but each with a different focus. In education research, external validity focuses on the concept of generalizability and addresses the question: How generalizable is the locally embedded causal relationship over varied persons, treatments, observations, and settings? The focus of external validity is on establishing task equivalence in order to generalize beyond the experimental circumstances and impose a close system on a more open behavior system (Campbell, 1957; Campbell & Stanley, 1963; Gall, Gall & Borc, 2003; Valsiner & Benigni, 1986). The most relevant approach to address external validity is through multidimensional and latent variable testing analysis between groups and settings.

As Table 1 shows, there are two types of external validity – person-based and situation-based (ecological). Since our population is essentially stable (that is, we're not necessarily concerned about generalizing beyond our sample of university students, because they are, in fact, the population itself), then we should really focus on ecological validity as a threat to higher education. In the context of KoKoHs, ecological validity focuses on the concept of accordance and addresses the question: To what extent the competencies developed through education programs are in accordance with those required at workplace? This focus allows the KoKoHs programs to assess efficacy of the education intervention programs in producing the expected outcomes that conformed to workplace. The most relevant approach to address ecological validity is through multivariate analysis of repeated measures within individuals across settings. The ultimate purpose is to enhance efficacy of higher education in producing productive workforce for society. Increasing the awareness of ecological validity urges educators and researchers to develop a rigorous method for answering this question in research and curriculum design, teacher induction, program evaluation, and performance assessment etc.

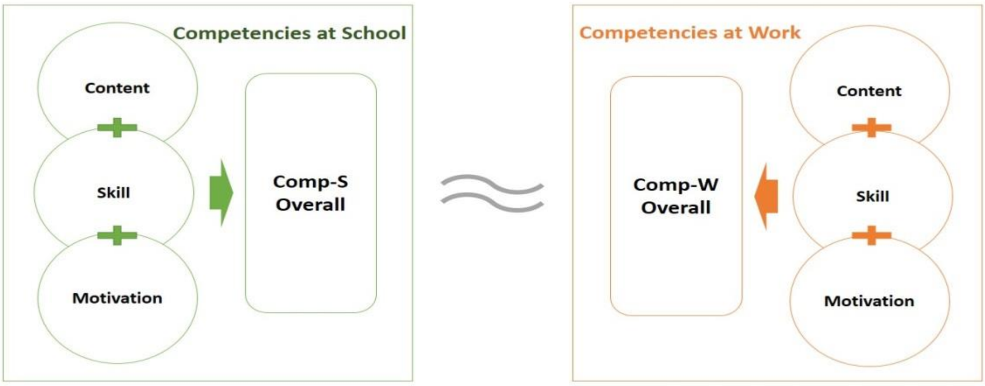
Applying Ecological Validity to KoKoHs

In the aspect of theory, all projects of KoKoHs relied on Weinert's (2001) definition of competencies as the latent cognitive and affective-motivational underpinnings of performance. In this theoretical framework, competencies include cognitive disposition, i.e., academically gained knowledge, as well as the motivational, volitional, and social dispositions to apply the gained knowledge flexibly in different situations (Blömeke, Zlatkin-Troitschanskaia, Kuhn & Fege, 2013). As Macha and Schuhen (2011, p. 38) summarized:

> One can specify the definition by Weinert (2001) as follows: "[Competencies are] the readily available or learnable cognitive [structures or processes of cognition and knowledge] abilities [memory, language, perception, attention, etc.] and skills [actions which are applied in recurring tasks] which are needed for solving problems [overcome barriers between a given state and a desired goal] as well as the associated motivational [concerning the motives which have an impact on the action or decision], volitional and social capabilities and skills which are required for successful and responsible problem solving in variable situations". Thus the existence of competence relies on three crucial dimensions: (1) cognitive abilities, (2) skills, and (3) the necessary motivational, volitional, and social capabilities and skills to solve new problems.

Building on Weinert's (2001) theory, the following model (Figure 1) is proposed to address competencies in higher education from the ecological validity perspective.

Figure 1. A Componential Model of Competencies in higher education from the ecological validity perspective



Sources: adapted from Weinert's (2001) specification of competencies in higher education in three domains: knowledge and cognitive abilities; skills to solve new problems; and motivational, volitional, and social capabilities and skills.

As Figure 1 indicates, in the KoKoHs context, addressing ecological validity is to examine the extent that competencies developed in higher education settings are in accordance with those expected in workplace. More specifically, this model can be transformed into a set of equations that correspond to a specific dimension and the overall competencies (Figure 2).

Figure 2. Operationalization of the Componential Model of Competencies in higher education from the ecological validity perspective

$$\text{Comp-S}_{\text{1-n}} \approx \text{Comp-W}_{\text{1-n}}\ldots\ldots\ldots\ldots\text{Overall}$$

$$\text{Comp-S}_{\text{knowledge}} \approx \text{Comp-W}_{\text{knolwedge}}\ldots\ldots\text{subject matter content}$$

$$\text{Comp-S}_{\text{skill}} \approx \text{Comp-W}_{\text{skill}}\ldots\ldots\ldots\text{procedural knowledge}$$

$$\text{Comp-S}_{\text{motivation}} \approx \text{Comp-W}_{\text{motivation}}\ldots\ldots\text{affective factors}$$

Note: **Comp-** stands for Competencies, **-S** stands for school, **-W** stands for workplace, and **-Overall** stands for competencies of the three dimensions combined.

As formula 1-4 indicate below indicate, ecological coefficients can be calculated for three individual dimensions and for all three dimensions of competencies (i.e., knowledge, skill, and motivation) combined. It is particularly noted here that this model is at the risk of oversimplification because each of the three dimensions entails a multilayer structure and the overall competencies includes a complex interaction of knowledge, skills, and personal variables.

Nevertheless, these equations provide a means to examine differences between measures of dimensional and overall competencies developed in university settings and those at workplace. These ecological coefficients serve as a measure to assess efficacy of a particular education program in achieving its purpose of producing capable workforce for the society.

$$\text{Efficacy}_{knowledge} = Comp\_S_{knowledge} - Comp\_W_{knowledge} \text{ ……………………..(1)}$$

$$\text{Efficacy}_{Skill} = Comp\_S_{Skill} - Comp\_W_{Skill} \text{ …………………………………..(2)}$$

$$\text{Efficacy}_{Motivation} = Comp\_S_{Motivation} - Comp\_W_{Motivation} \text{…………(3)}$$

$$\text{Efficacy}_{Overall} = Comp\_S_{Overall} - Comp\_W_{Overall} \text{………………………..(4)}$$

These equations can generate index coefficients regarding the efficacy of higher education programs in producing competencies for workplace. These coefficients indicate the degree of ecological validity for education programs covered in the KoKoHs program.

1. When the difference between Comp-S and Comp-W is **greater** than 0, i.e., in positive numbers, it suggests an over training in school.
2. When the difference between Comp-S and Comp-W is **less** than 0, i.e., in negative numbers, it suggests a under training in school.
3. When the difference between Comp-S and Comp-W larger **equals** to 0, it suggests there is a perfect match on competencies between school and workplace. Efficacy of school training reaches to the degree of 100% effective.

The index for each dimension indicates the effectiveness of an education program in producing competencies in a specific aspect, i.e., problem solving skills in computer programming. The overall index indicates the extent of the overall efficacy/effectiveness of higher education training program in producing competencies expected at workplace. These indexes provide an indicator of the program effectiveness: Whether the training/educational interventions adequately prepare students for the situations they will see in the workforce. They can also serve as assessment tools: Do the inferences we draw apply to the workforce setting? Are there are similarities and differences in addressing these questions across different subject matter areas. The answers to these questions have significant implications for how we design, implement, and assess higher education programs.

Again, we would like to point out, it is not the demand of higher education to 'produce' 100% effective graduate in regard of the labor market. But this should stimulate the reflection of what competencies should be trained by higher education. Needless to say, graduates will spread over all kinds of sectors. Therefore, it is almost impossible and even desirable to cover all of the demands of the labor market. However, addressing ecological validity offers a specific means to model and measure competencies of university graduates for workplaces and assess efficacy and efficiency of individual programs in the KoKoHs program.

Again, at the risk of oversimplification, the following formula can be used to examine the overall competencies in school and workplace assuming equal weights:

$$Comp\_S_{Overall} = \frac{Comp\_S_{knowledge} + Comp\_S_{Skill} + Comp\_S_{motivation}}{3} \text{.............................}(5)$$

$$Comp\_W_{Overall} = \frac{Comp\_W_{knowledge} + Comp\_W_{Skill} + Comp\_W_{motivation}}{3} \text{........................}(6)$$

Also, the degree of ecological validity in percentage can be calculated with this formulate:

$$\text{Efficacy}_{Overall} = \frac{Comp\_S_{Overall}}{Comp\_W_{Overall}} \times 100\% \text{ ....}(7)$$

Implications of Addressing Ecological Validity in KoKoHs

This paper proposed a componential model which refers ecological validity to the agreement of the competencies developed in universities with those expected at workplaces. This model elaborates on the classic description (Brunswik, 1943, 1956) of ecological validity and highlights the agreement of competencies between school and workplace. The focus is on improving the agreement between measures of competencies of university students observed and recorded in a KoKoHs program and those that actually occur in natural settings at workplaces. It is hoped that addressing ecological validity would help develop capabilities of our university graduates to "cope with the multiple, noisy, messy situations, which occur in the environment (Araújo et al., 2007, p. 70)". An efficient way to achieve a high degree of agreement is for universities and workplaces to work in tandem to represent the complex, and sometimes irregular, conditions in which university graduates will function at workplaces.

It is important note that addressing ecological validity is more than developing psychometric instruments to establish "the functional and predictive relationship between the test taker's performance on a particular test and the test taker's behavior in a real-world setting, such as work" (Walker, 2012). Increasing the awareness of ecological validity urges educators and researchers to develop a rigorous method for answering this question in research and curriculum design, teacher induction, program evaluation, and performance assessment etc. As Hammond and Stewart (2001) pointed out, it is crucially to ask “To what set of circumstances do we wish to generalize, or apply, our results?” before an education program starts rather than after it is finished.

Currently, higher education and the real-world workplace entail two different settings. Both function mostly as independent entities with little dialogue with each other. “Anyone with the responsibilities of hiring, training, and supervising recent college graduates for workplace success has more than likely questioned whether scholastic test performances and college grades have anything to do with workplace competencies” (Walker, 2012). Addressing ecological validity calls for higher education and workplace to converge the settings and get the two separate standards aligned and approximated to each other as closely as possible, so that each standard works in its home setting and informs its corresponding setting. In this sense, we agree with Shadish et al’s position of viewing ecological validity more “as a method that calls for research with samples of settings and participants that reflect the ecology of application” than as a separate validity type” (2002, p. 37). As Mark (1986) eloquently pointed out, “a validity typology can greatly aid … design, but it does not substitute for critical analysis of the particular case or for logic” (p. 63).

Addressing the discrepancy between school and workplace is not new. The last century has witnessed multiple curriculum reform movements (Powell, 2007). As a consequence of the sputnik shock, the budget of the National science foundation had been raised four times and the idea of educating world-leading engineers for the labor market became central in US and West Europe (Kerr, 1991). Another reform has been massive influenced by students, which called for more democratic decisions and opening access to higher education and to get prepared for the labor market (Teichler, Hartung, Nuthmann, 1980; Allen, Ramaekers & Van der Velden, 2005).

A revitalized attention has been generated to this issue since last decade which resulted in various programs and initiatives. Such efforts included the competence-based initiatives in the US (US Department of Education, 2002), partnership for 21st century skills (http://www.p21.org/), market-based approaches to teacher education (Apple, 2001; DiBenedetto & Zimmerman, 2013; Sitzmann & Ely, 2011), case-based instruction at Harvard Business School (“Case Method Teaching,” 2014) and the University at Buffalo-Michigan State University (“Assessing Case-Based Instruction,” 2014), and various internship programs in professional education, such as teacher education, law, nursing, counseling, and social work. Outside school, similar efforts could be found in skills and competency management in industrial and commercial training (Homer, 2001; National Restaurant Association, 2012), competence-based recruitment and selection (Wood & Payne, 1998), human resources management (Dubois & Rothwell, 2004; Sanchez & Levine, 2009; Wilton, 2013), and organizational management

(Cheng & Dainty, 2005; Draganidis & Mentzas, 2006; Hersey, Blanchard & Johnson, 2012). Our model of ecological validity aimed at serving as a specific means for school and workplace to inform each other in modeling and measuring competencies in order to produce productive citizens.

A most recent effort in this direction was reflected by the partnership of Purdue University with Gallup, the global polling and consulting organization, to create an index--the Gallup-Purdue Index. The Gallup-Purdue Index project facilitates the "largest representative study of college graduates in U.S. history". This index is designed to survey alumni, providing universities and employers with detailed information, including earnings data. The purpose is to create a national benchmark that evaluates the long-term success of graduates, measured by indicators including career and life satisfaction. In particular, the index takes into account workplace engagement and well-being, measured by dimensions that surround characteristics of college graduates' social, physical, financial and community lives. "What we're measuring is really to what degree these graduates have great jobs and great lives," said Brandon Busteed, executive director of Gallup Education. "We hope this is something that the higher education sector is really excited about. It sends a clear message that this is about higher ed, for higher ed, by higher ed" (Vedder & Denhart, 2014). The Gallup-Purdue Index is hailed as an ambitious and challenging undertaking that offers a thoughtful, research-based approach to evaluating the outcomes of students' higher education experiences. It offers a means to individually track student growth while they're at Purdue, and therefore provides powerful new evidence to measure whether colleges and universities deliver on the improved life and job outcomes that Americans expect of them (Colombo, 2013).

The importance of such a connection becomes more obvious for university graduates in arts and humanities who often had a harder time of finding a job in the private sector than other graduates do. It is starkly ironic that the graduates don't know what required competences they are obsessing while representatives of the labor market don't know what competencies these graduates bring with them (Briedis et al., 2008). As far as we know, there is no program that is designed with systematic connections between higher education institutes and the labor market in Germany. However, the Federal Ministry of Education and Research, scientists, and representatives of business are working together to address this pressing issue. For instance, the ***Job Requirements Approach*** **(JRA**) (Felstead et al., 2007; Peterson et al., 2001) has been recognized internationally as a methodological approach to identifying tasks and activities at work. Based on the JRA, survey instruments have been developed. They include the O*NET *Generalized Work Activities Questionnaire* (GWA) (Jeanneret et al., 2002; O*NET 2012; Peterson et al., 2001), the *UK Skills Survey* (BMRB 2006; Felstead et al., 2007), the OECD *Programme for the International Assessment of Adult Competencies* (PIAAC) (OECD, 2013a; OECD 2013b), and the Dutch Version of the GWA (Toolsema, 2003). Similarly, employees are surveyed by the *BIBB/BAuA* (Rohrbach-Schmidt, 2009) as well as in the German National Education Panel Study (Matthes & Christoph, 2013).

All these governmental and industrial efforts aimed at identifying important areas of job-related activities. However, these instruments are largely generic in nature and they do not speak to development of competencies of higher education graduates. Addressing ecological validity helps enhancing communications and synchronization between school and workplace. In this communication and synchronization, the traditional issues such as using foreign languages or scientific techniques, group management, working in a holistic way, and working under pressure of time, should be addressed. So should the new demands for both school and employers that emerged with advancement of science and technology. For instance, the function and influence of social media on learning and instruction should be considered both in school and at workplace.

One of the projects within KoKoHs, the project *KomPaed* (*Job-related competences in educational fields of work, Braun et al., 2013*), is set up to identify daily performed job-related activities and requirements in order to measure competencies indirectly. This project aimed at producing specific descriptions of the expectations from university graduates after they enter the labor market. Addressing ecological validity would help interpret results of this project and serve as benchmarks to improve the programs in higher education.

One of the primary challenges in addressing ecological validity in modeling and measuring competencies in higher education points to the insufficient characterization of the concept of competency in higher education. Meeting this challenge requires further clarification of the nature of this construct. As discussed above, all projects of KoKoHs adopted Weinert's (2001) definition which viewed competencies as the latent cognitive and affective-motivational underpinnings of performance (Blömeke, Zlatkin-Troitschanskaia, Kuhn & Fege, 2013; Macha & Schuhen, 2011). On the other hand, competencies were viewed as explicit and fully manifest: What can students do to demonstrate knowledge, skills, learning, etc. (Markle, Olivera-Aguilar, Jackson, Noeth & Robbins, 2013). Clarification of the nature of competencies helps address important questions such as: To what extent is it the responsibility of higher education to prepare students for the workforce? Can we build competency models that are theoretically sound, logical, measurable, and relevant to the workforce? One way to advance this important work is to establish a reciprocal communication between school and workplace in real life. This communication allows higher education and workplace work in tandem to develop innovative and practical models for describing and developing this concept that are founded on integrative logic with a particular focus on joint efforts of school and workplace in real life. Again, what we proposed here is not to downplay the existing programs in higher education that aim at educating and producing competent academia and researchers. Our intent is to advocate for more effort in innovative programs that aim at preparing graduates for a workplace outside universities. We believe that addressing ecological validity offer a specific means to approach this task.

Another challenge is the development of valid and reliable instrument to measure competencies at workplace in natural settings with temporally and spatially rich stimuli (DiBenedetto & Zimmerman, 2013; Edelbring, 2012; Newell & Simon, 1972; Sitzmann & Ely, 2011). Important questions to be addressed in this regard include: What types of assessments are needed in this space? Can we have a one-size-fits-all assessment that helps institutions evaluate students and improve curricula while still certifying skills for the workforce? Apparently, there is no obvious answer to these challenges because no clear mechanism for judging ecological validity has been set forth; nor are there any suggestions as to the nature of the critical factors for this judgment (Schmuckler, 2001). However, it points towards a correct direction in bringing the attention of educators and researchers to the issue of ecological validity. It is unequivocally clear that continuous effort is needed in order to model and measure competencies in higher education in a valid and reliable fashion. More importantly, progress in this area helps address the ultimate question: To what extent that university graduates are prepared so that they are ready to carry out various tasks that their profession demands and function as a valuable contributing citizen in their social, physical, financial, and community lives?

References

Allen, J., Ramaekers, G. & Van der Velden, R. (2005). Measuring competencies of higher education graduates. New Directions for institutional research, 2005(126), 49-59.

Apple, M. W. (2001). Markets, standards, teaching and teacher education. Journal of Teacher Education, 52(3), 182–196.

Araújo, D., Davids, K. & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comments on Rogers, Kadar, and Costall (2005). Ecological Psychology, 19, 69-78.

Assessing case-based instruction (2014). Available: http://edr1.educ.msu.edu/references/.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C. & Fege, J. (Eds.) (2013). *Modeling and measuring competencies in higher education: Tasks and challenges.* Rotterdam, Netherlands: Sense Publishers.

BMRB Social Research (2006). *2006 Skills Survey. Technical Report*. Available: http://www.esds.ac.uk/doc/6004/mrdoc/pdf/6004userguide.pdf (October 9th, 2013).

Bracht, G. H. & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437-474.

Braun, E., Schwippert, K., Prinz, D., Schaeper, H., Fickermann, D., Brachem, J.-C. & Pfeiffer, J. (2013). Competencies in Fields of Educational Activities. In S. Blömeke & O. Zlatkin-Troitschanskaia, (Eds.), *KoKoHs Working Paper No. 3. The German funding initiative "Modeling and Measuring Competencies in Higher Education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students* (S. 67-69). Berlin/Mainz: Humboldt Universität zu Berlin/Johannes Gutenberg Universität Mainz.

Briedis, K., Fabian, G., Kerst, C. & Schaeper, H. (2008). *Berufsverbleib von Geisteswissenschaftlerinnen und Geisteswissenschaftlern*. HIS.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Brunswik, E. (1943). Organismic Achievement and Environmental Probability. *Psychological Review 50*, 255-272.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed., rev. & enl.). Berkeley: University of California Press.

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychological Bulletin, 54,* 297–312.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cao, L. (2013). Prospects and challenges in modeling and measuring competencies in higher education: A reflection. A working paper presented at the International Colloquium for Young Researchers at Mainz, Germany.

Cleary, T. J. (2009). School-based motivation and self-regulation assessments: an examination of school psychologist beliefs and practices. Journal of Applied School Psychology, 25(1). 71–94.

Cleary, T. J., Callan, G. L. & Zimmerman, B. J. (2012). *Assessing Self-Regulation as a Cyclical, Context-Specific Phenomenon: Overview and Analysis of SRL Microanalytic Protocols.* Educational Research International, 2012.

Colombo, H. (2013). *Purdue, Gallup join to create measures higher-ed learning outcomes.* Available: http://www.newssentinel.com/apps/pbcs.dll/article?AID=/20131218/NEWS/312189957/0/FRONTPAGE.

Case method teaching (2014). Available: http://hbsp.harvard.edu/product/casemethodteaching (March 25th, 2014).

Cheng, M. I. &. Dainty, R. I. J. (2005). Toward a multidimensional competency-based managerial performance framework: A hybrid approach. *Journal of Managerial Psychology, 20*, 380–396.

DiBenedetto, M. & Zimmerman, B. (2013). Construct and predictive validity of microanalytic measures of students' self-regulation of science learning. *Learning and Individual Differences, 26*, 30-41.

Dubois, D. & Rothwell, W. (2004). *Competency-Based Human Resource Management*. Davies–Black Publishing.

Draganidis, F. & Mentzas, G. (2006). Competency-based management: A review of systems and approaches. *Information Management &Computer Security, 14,* 51–64.

Edelbring, S. (2012). Measuring strategies for learning regulation in medical education: Scale reliability and dimensionality in a Swedish sample. *BMC Medical Education, 12*, 76.

Felstead, A., Gallie, D., Green, F. & Zhou, Y. (2007). *Skills at Work, 1986-2006*. Oxford: Skope.

Guthrie, J. T., Wigfield, A. & Perencevich, K. C. (2004). "Scaffolding for motivation and engagement in reading," in J. T. Guthrie, A. Wigfield. & K. C. Perencevich, (Eds.), *Motivating Reading Comprehension: Concept-Oriented Reading Instruction* (pp. 55–86). Mahwah, NJ: Lawrence Erlbaum.

Hammond, K. & Stewart, T. (2001). *The essential Brunswik: Beginnings, explications, applications.* New York: Oxford University Press.

Hersey, P., Blanchard, K. H. & Johnson, D. E. (2012). *Management of Organizational Behavior* (10th Edition). Upper Saddle River, NJ: Prentice Hall.

Homer, M. (2001). Skills and competency management. *Industrial and Commercial training, 33*(2), 59–62.

International Futures (IFs) modeling system, Version 7.00. Frederick S. Pardee Center for International Futures, Josef Korbel School of International Studies, University of Denver, Denver, CO. Available: http://www.ifs.du.edu/ifs/frm_CountryProfile.aspx?Country=DE.

Jeanneret, P. R., Borman, W. C., Kubisiak, U. C. & Hanson, M. A. (2002). Generalized work activities. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret & E. A. Fleishman (Eds.), An occupational information system for the 21st century: The development of O*NET (S. 105-125). Washington, DC: American Psychological Association.

Kitsantas, A. & Zimmerman, B. J. (2002). Comparing self-regulatory processes among novice, non-expert, and expert volleyball players: amicroanalytic study," *Journal of Applied Sport Psychology*, *14*(2), 91–105.

Leiman, M. & W. Stiles, W. B. (2001). Dialogical sequence analysis and the zone of proximal development as conceptual enhancements to the assimilation model: the case of Jan revisited, *Psychotherapy Research, 11*(3), 311–330.

Kerr, Clark (1991). *The great transformation in higher education, 1960-1980*. SUNY Press.

Macha, K. & Schuhen, M. (2011). Framework of Measuring Economic Competencies. *Journal of Social Science Education 10*(3), 36–45.

Mark, M. M. (1986). Validity typologies and the logic and practice of quasi-experimentation. In W. M. K. Trochim (Ed.), *Advances in quasi-experimental design and analysis* (pp. 47-66). San Francisco: Jossey-Bass.

Markle, R., Olivera-Aguilar, M., Jackson, T., Noeth, R. & Robbins, S. (2013). *Examining Evidence of Reliability, Validity, and Fairness for the SuccessNavigator™ Assessment.* A research report**.** Princeton, NJ: Educational Testing Service.

Matthes, B. & Christoph, B. (2011). *Nationales Bildungspanel. Großpilot E8 Feldversion. Version 1.02.*

National Restaurant Association (2012). *Hospitality and restaurant management.* Upper Saddle River, NJ: Pearson.

Newell, A. & Simon, H. (1972). *Human Problem Solving*. New Jersey, Englewood Cliffs: Prentice-Hall Inc.

Reschly, D. J. (2008). School psychology paradigm shift and beyond. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology,* (5th ed., pp. 3–15). Bethesda, MD: National Association of School Psychology.

OECD (2013a). *Education at a Glance 2013: OECD Indicators,* OECD Publishing. Available: http://dx.doi.org/10.1787/eag-2013-e.

OECD (2013b). *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris: OECD.

O*NET (2012). *O*NET Generalized Work Activities Questionnaire*. Available: http://www.onetcenter.org/dl_files/MS_Word/Generalized_Work_Activities.pdf (December 4th, 2012).

Peterson, N. G., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K. Y., Campion, M. A., Mayfield, M. S., Morgeson, F. P., Pearlman, K., Gowing, M. K., Silver, M. B. & Dye, D. M. (2001). Understanding work using the Occupational Information Network (O*NET): Implications for practice and research. *Personnel Psychology*, 54, 451-492.

Powell, A. (2007). How Sputnik changed US education: Fifty years later, panelists consider a new science education 'surge'. Harvard Gazette. Available: http://news.harvard.edu/gazette/story/2007/10/how-sputnik-changed-u-s-education/.

Sanchez, J. I. &. Levine, E. L. (2009). What is (or should be) the difference between competency modeling and traditional job analysis? Human Resource Management Review, 19, 53–63.

Schmitz, B. & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: time-series analyses of diary data, *Contemporary Educational Psychology, 31*(1), 64–96.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*, 419-436.

Sitzmann, R. & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin, 137*(3), 421-442.

Teichler, U., Hartung, D. & Nuthmann, R. (1980). *Higher education and the needs of society*. Windsor: NFER Publishing Company.

The Partnership for 21st Century Skills (2014). Framework for 21st Century Learning. Available: http://www.p21.org/.

Toolsema, B. (2003). *Werken met cempetenties. Naar een instrument voor de identificatie van competenties*. Enschede: University of Twente.

U.S. Department of Education, National Center for Education Statistics (2002). *Defining and Assessing Learning: Exploring Competency-Based Initiatives*, NCES 2002-159, prepared by Elizabeth A. Jones and Richard A. Voorhees, with Karen Paulson, for the Council of the National Postsecondary Education Cooperative Working Group on Competency-Based Initiatives. Washington, DC: U.S. Department of Education.

Valsiner, J. & Benigni, L. (1986). Naturalistic research and ecological thinking in the study of child development. *Developmental Review, 6,* 203-223.

Vedder, R. & Denhart, C. (2014). *How the College Bubble Will Pop.* Available: http://online.wsj.com/news/articles/SB10001424052702303933104579302951214561682.

Walker, J. (2012). *Ecological Validity in Vocational Assessments.* Available: http://www.cecassoc.com/NewWorkerSpring2012.html.

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.

Wilton, N. (2013). *An Introduction to Human Resource Management.* Thousand Oaks, CA: Sage.

Wood. R. & Payne, T. (1998). *Competency-Based Recruitment and Selection*. Wiley.

Wundt, W. (1999). *Grundzüge der physiologischen Psychologie.* Bristol, England: Thoemmes Press. (Original work published 1874).

Linda Gräfe, Andreas Frey, Sebastian Born, Raphael Bernhardt, Gernot Herzer, Anna Mikolajetz, and S. Franziska C. Wenzel, Friedrich Schiller University Jena, Germany

Written University Exams based on Item Response Theory (MoKoMasch)

In order to determine whether students have acquired the competencies and/or knowledge regarded as necessary to assign credit points and to justify pass/fail decisions, written exams are a common instrument which is broadly used at universities. The results of such exams are directly linked to decisions with a high individual relevance for each student. Unfortunately, written university exams often lack in common measurement standards, which is problematic for three major reasons. First, the learning objectives focused by the course are not systematically represented by the exams. Hence, the extent to which an exam measures what it is supposed to measure remains unclear. Second, the relation between the assigned grades and the fulfillment of the learning objectives is regularly kept indistinct and thus, the results of students are interpreted in a norm-referenced way only. Third, the scales of written exams are typically not connected across cohorts. This missing connection, however, makes the exams unfair as the same performance of a student could lead to different grades in different cohorts.

With this paper we therefore want to (a) draw the attention of measurement experts and university teachers to this issue, (b) outline a procedure to overcome the mentioned shortcomings, and (c) illustrate this procedure with empirical results.

Proposed Procedure

In order to avoid or at least substantially reduce the problems associated with typical written university exams we suggest applying a combination of well-established and modern measurement procedures. Specifically, the learning objectives need to be described by a detailed assessment framework and operationalized thoroughly by high-quality test items (e.g., Osterlind, 2002). These items should be given to the students in a standardized setting. The gathered responses are then scaled using item response theory (IRT) models (e.g., van der Linden & Hambleton, 1997). In order to make criterion-referenced test score interpretations possible, a standard-setting procedure may then be used to define cut-off points between grade levels and/or between pass and fail. Finally, tests presented to different cohorts should be connected by appropriate linking or equating methods (e.g., Kolen & Brennan, 2014) to establish a consistent evaluation standard across several cohorts.

Empirical Application

The proposed procedure was applied for the written exam at the end of a course on “Introduction to Research Methods in Education”. The assessment framework consisted of ten content areas combined with the cognitive processes of Bloom’s taxonomy (Bloom, Englehart, Furst, Hill & Krathwohl, 1956). The assessment framework was operationalized by an item pool of 80 test items. From this item pool, in the year 2012 as well as in the year 2013 an exam in paper-and-pencil format was assembled. The item set used in each exam covered the assessment framework and contained 37 (2012) and 35 (2013) items, respectively. The second exam in the year 2013 comprised 17 link items which had also been used in the first exam. Thus, a common item nonequivalent group design (Kolen & Brennan, 2014) was used to link the two assessments. This made it possible to report the results obtained in the second assessment on the same scale as in the first assessment. The exams were given to two cohorts of educational science students with $N_{2012} = 114$ and $N_{2013} = 97$ (84% female in both years). The gathered responses were scaled with the one-parameter logistic IRT model. A model with low complexity was chosen to maximize the probability that item parameter estimates remain stable over time. In order to evaluate the item fit, the mean squared error (MNSQ) and the weighted mean squared error (WMNSQ) as well as their corresponding *t*-values were analyzed. The ability of the students was estimated with maximum likelihood. Finally, a simplified bookmarking procedure (e.g., Mitzel, Lewis, Patz & Green, 2001) was used to set the cut-offs between grade levels.

Results

There was no item in the first assessment showing a significant misfit. One item had to be excluded because all responses to that item were incorrect. All in all, 36 items remained in the first test for the following analyses. The mean of the difficulty distribution was -1.03 (*SD* = 1.50). The mean of the point biserial correlation between the single items and the sum of solved items (corresponding to the item discrimination from classical test theory) was .37 with a range of .09 - .61. For model identification purposes the mean of the latent ability distribution was fixed to 0.00. The variance was freely estimated as 0.75. The reliability of the ability estimates was .80.

Concerning the linking between the two cohorts, 12 of the 17 link items showed item parameter invariance and could be included in the analysis of the year 2013 with fixed difficulty parameters. The difficulties of the remaining five link items showed significant differences between the two assessments. Consequently, for the second assessment their difficulty parameters were estimated freely.

One item of the second assessment showed a significant misfit (WMNSQ = 1.22; t=2.6.). (However, this item was kept in the test because it had an acceptable point biserial correlation with the total score (.21). In addition, providing feedback to the students is much easier without item exclusions.

The mean of the difficulty distribution in the second year was -0.69 (*SD* = 1.26). The point biserial correlation between the single items and the total score (mean: .43, range: .08 - .66) was a bit higher compared to the findings of the first assessment. While the mean of the latent ability distribution was slightly lower (-.11) after linking compared to the first assessment, the variance (1.11) of the latent ability distribution was higher. The reliability of the ability estimates was .85 and thus even a bit higher compared to the year before.

Discussion

With the proposed procedure we are advocating a combination of methods that makes it possible to directly connect test scores and/or given grades to the fulfilment of learning objectives. Furthermore, the procedure offers the possibility to establish stable evaluation criteria over different assessments. Thus, the requirements of what students should know and can do to reach a certain grade level can be kept constant over time. Both, criterion-referenced test score interpretations and time-invariant cut-scores are achieved by using an IRT model. Exams based on sum scores or classical test theory would not be able to achieve both goals which are very important to resolve major problems of typical university exams.

The empirical results obtained from two applications of a newly developed written university exam show that the procedure can be well applied in typical university settings. The reliabilities of the ability measures achieved for the two exams were good to very good. Nevertheless, it has to be noted that the standard errors of the ability estimates are rather large (average standard error: 0.44). As a consequence, the 95%-confidence interval around the ability estimate of a student typically covers several grade levels. If not restricted by the local university, the usage of a smaller number of grade levels is hence recommended.

Regarding the aim of maintaining the same reporting scale over time, for the case of written university exams it has to be considered that the invariance of item parameters over assessments depends to some degree upon the instruction in the preceding course. Nevertheless, the present study underlines that establishing a solid linking between assessments is possible if the written exam is based on a common assessment framework which is underlying both exams.

Summarizing, the proposed procedure proved to be a promising method capable to increase the validity and fairness of written university exams. These are important steps towards a quality of written university exams which reflects the high individual relevance of the test results. We are recommending the use and further development of IRT-based written university exams.

References

Bloom, B., Englehart, M. Furst, E., Hill, W. & Krathwohl, D. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, Toronto: Longmans, Green.

Kolen, M. J. & Brennan, R. L. (2014). Test equating, scaling, and linking. Methods and practices (3rd Ed.). New York: Springer.

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 249-281). Mahwah, NJ: Erlbaum.

Osterlind, S. J. (2002). Constructing Test Items: Multiple-Choice, constructed-response, performance, and other formats (2nd Ed.). Boston, Dordrecht, London: Kluwer.

van der Linden, W. J. & Hambleton, R. K. (Eds.). (1997). Handbook of modern item response theory. New York, NY: Springer.

The preparation of this article was supported in part by grant 01PK11005C (project “Modeling Competencies of Mechanical Engineering Students in the Areas of Construction, Design and Production Engineering” – MoKoMasch) from the German Federal Ministry of Education and Research within the research initiative “Modeling and Measuring Competencies in Higher Education” (KoKoHs).

Correspondence concerning this article should be addressed to Linda Gräfe, Institute of Educational Science, Department of Research Methods in Education, Friedrich Schiller University Jena, Am Planetarium 4, D-07737 Jena, Germany, Fon: +49 (0)3641-945393, e-mail: linda.graefe@uni-jena.de.

Ronald K. Hambleton, University of Massachusetts Amherst, USA

Comment on - Item Response Theory Based University Exams

Thank-you to many individuals and agencies

- German Federal Ministry of Education and Research for supporting this huge and very important study to improve curriculum building and assessment in higher education.
- Professor Olga Zlatkin-Troitschanskaia at Mainz University and her faculty colleagues in Germany who are participating.
- Large number of graduate students (over 70) who are working hard on many research projects in higher education, several of which will be presented here today.

Background to this IRT-Based University Exams Research

Begin with three problems in higher education in Germany:

(1) Learning outcomes are <u>not</u> properly reflected on the exams.

(2) Norm-referenced tests (NRTs) are common.

(3) These NRTs are not linked from one semester to the next.

(1) This one is common in my country too - professors are rarely trained in assessment practices, and so not only are tests not content validity, but the targets of instruction (learning outcomes) are rarely well defined either. (Jim Popham talked about “cloud referenced tests” in 1974.)

(2) Distinctions between NRTs and CRTs did not become common until the 1970s (in areas of purposes, test development, and evaluation). CRTs are mainly needed.

(3) A common reporting scale across years would permit common standards to be used across time and even instructors.

--An IRT approach would be helpful for linking the final exams from year to year. (Though classical equating methods would be fine too.)

One or Two Paradigm Shifts?

- Problems 1 and 2 can be addressed with a paradigm shift from NRT to CRT methods and practices.

--This is critical and would address the first two problems. Defining learning outcomes, new types of items, developing CRTs, setting performance standards, etc.

--In Secolsky's "Assessment in Higher Education," we wrote a chapter on item analysis, but many other chapters too that are very practical.

--IRT is not needed, and still two huge problems for assessment in higher education could be solved.

Problems 1 and 2

- The authors make a strong case for features of TIMSS and PISA. I agree, content specs., item writing, etc. are well handled.
- But, in these projects, international committees agree on the content frameworks. If items are linked to the frameworks, the tests have content validity.

--The reality is though that the content specifications do not necessarily link to the content specs. of users, such as countries, and this complicates score interpretations.

--With good CRTs that a professor might use items need to assess the content specifications which need to line up with what is actually taught in the course.

For this reason, I would recommend that the researchers also look at the way valid CRTs are constructed - for example look at the websites of the states in the US. Nearly all of them specify learning outcomes, build tests to measure them, teach them, and then assess. Documentation is tremendous - 100s of pages.

- The focus too with PISA and TIMSS are complicated group based IRT models using plausible values methodology. Individual scores are not even estimated. Again, the US state reports may be more relevant (though I agree that the IRT work would still be complicated) but at least the focus is on the individual student.

One or Two Paradigm Shifts?

- Problem 3 can be addressed with a paradigm shift from CTT to IRT methods and practices, but classical methods of equating would be fine too.

--What does not appear in the short paper are the reasons for shifting to IRT.

--My own preference would be to focus on problems 1 and 2, and work on problem 3 later or simultaneously, with or without IRT.

The Case for Item Response Model (IRT)?

- In principle, it is an easy case to make:

--Achievement estimates of persons could be independent of the specific items on the test;

--*Item statistics* could be independent of the particular sample of candidates;

--An estimate of precision of measurement for each candidate would be available

--If computer-adaptive testing were to become viable, IRT would be needed, and more.

- In practice, there are lots of problems to overcome:

--Almost no professors would know anything about it.

--Item calibrations often require much bigger samples than are available in many university courses (In this study about 100). Errors are large.

- Small samples create big problems for equating, and model fit studies are problematic (because statistical power is low). A good example: classical discrimination indices varied from .09 to .61, but model fit was excellent! This highlights lack of power to detect model misfit.

- Applications of IRT would not be easy for professors. [Perhaps universities would be equipped with staff like those on this paper in resource centers around universities to handle the complexities.]

--For example, it was mentioned in passing that 5 of 17 items were deleted from the link. This is a very high number. Knowing more about these five would be very important. Is it content, item quality, shift in dimensionality?

Finally,….

- These are clever researchers, and with excellent ideas, and their work so far appears top-quality.

- Perhaps I should have gone and read their full reports, and maybe I would feel differently about IRT—I have only read a six page summary.
- I would focus on problems 1 and 2, and with IRT I think there is a lot of research that could be done, especially with sample sizes and instructor training.

Next Steps

- Continue the excellent work so far and consider:

1. Sample sizes and their implications and consequences for using IRT models successfully. This would include studies of item calibration, assessing test dimensionality, and equating.
2. Field test approaches for training professors in the use of item banks, test development, and other uses of IRT in their work.

- Continue the excellent work so far and consider:

3. Methods for setting passing scores—bookmark may be fine, but other methods available and much can be learned about the process itself—and how to implement any judgmental methods with a very small sample of teachers, perhaps 1!

Marieke van Geel, Trynke Keuning, Jean-Paul Fox, Adrie Visscher, University of Twente, The Netherlands

Assessing the effects of a (school wide) data-based decision making intervention on student achievement growth in primary schools in the Netherlands

Abstract

Despite growing international interest in the use of data to enhance educational quality, relatively few studies examining the effects on student achievement are available. In the present study, the effects of a two-year data-based decision making intervention on student achievement growth were investigated. A total of 53 primary schools in the Netherlands participated in a project aimed at implementing data-based decision making throughout the entire school organization. Student achievement data was collected over the two school years prior to the intervention and during the two intervention years. Linear mixed models were used to analyze the differential effect of data-use on student achievement, controlling for background variables at the school and student level and accounting for individual growth in student achievement from grade three to eight.

A positive mean intervention effect over students, schools and grades, and heterogeneity in school intervention effects was estimated, with a value of approximately one extra month of schooling. Heterogeneity in performance of students in the study prior to intervention and during intervention were not attributable to differences in observed student background variables. High intervention effects were identified for low-SES schools and students, leading to the conclusion that the data-based decision making intervention especially significantly improved the achievement of students of low-SES schools.

Introduction

Today, data plays an important role in informing decisions in all sectors of society; from commercial organizations adjusting their sales strategy based on the analysis of customer behavior, to hospitals evaluating their treatment effectiveness, and teachers adapting their instruction to well-defined student needs (Lai & Schildkamp, 2013). In education, there is growing emphasis on the use of data to base decisions on, assuming that this will lead to increased student achievement. Data-based decision making (DBDM) can be defined as: “teachers, principals, and administrators systematically collecting and analyzing data to guide a range of decisions to help improve the success of students and schools” (Ikemoto & Marsh, 2007, p108). At the class, school and board level, student and school performance data is supposed to be analysed, and decisions are supposed to be based on these data. Since the aim of DBDM is to systematically maximize student achievement of all students, the focus is explicitly on evaluating and analysing student performance data, but in order to make decisions additional information is also gathered (Hamilton et al., 2009).

The intervention

Although only few studies provide empirical evidence for the effect of data-based decision making (DBDM) on the achievement of students, there is considerable empirical evidence for the elements DBDM can be decomposed into, such as the impact of feedback, setting goals, and improving instructional quality. In line with an increasing interest all over the world, the government in the Netherlands promotes the use of data to improve education.

At the University of Twente, an intervention aimed at data-based decision making was developed. The two-year training course for entire primary school teams was based on literature on professional development and aimed at acquiring the knowledge and skills related to DBDM and implementing and sustaining DBDM in the school organization.

Model and hypotheses

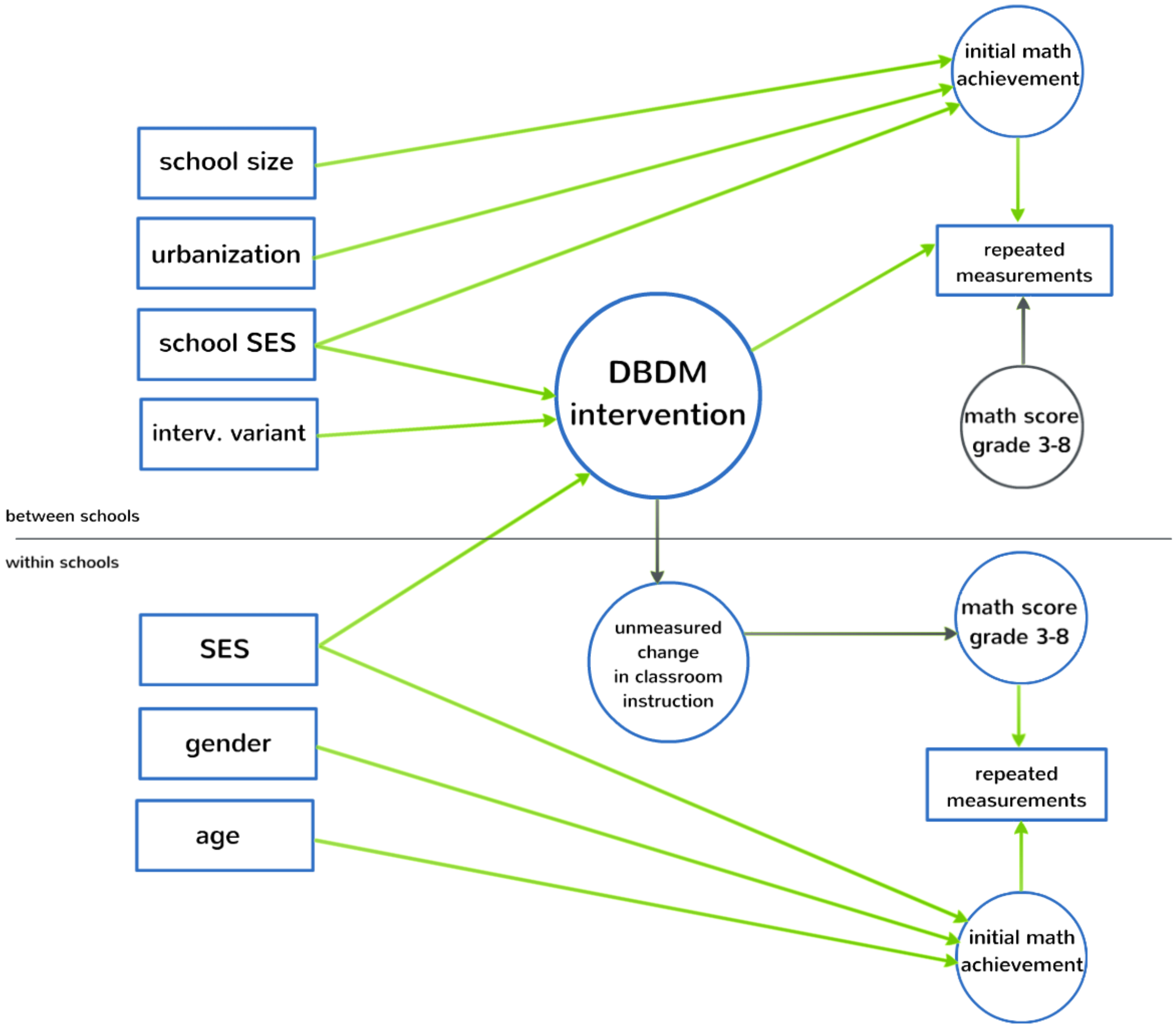
In Figure 1, the general model for this study is presented. It builds on previous studies on data-based decision making which state that the use of data can enhance student achievement (Campbell & Levin, 2008; Carlson, Borman & Robinson, 2011; Lai & McNaughton, 2013). In this multilevel model it is assumed that implementing DBDM will lead to (unmeasured) changes in teacher's classroom practices which, in turn, are responsible for raising student achievement growth in mathematics (*hypothesis 1*), and that intervention effects differ between schools (*hypothesis 2*).

At the school level, the effect of the implementation of DBDM might vary as a result of school characteristics such as school size, average student SES, and the level of urbanization. Schools with a higher percentage of students with a lower socio-economic background on average score less than schools with a high-SES student population (Carlson et al., 2011; Inspectie van het Onderwijs, 2012). Since teachers are more likely to underestimate the potential of students from a low-SES background, an interaction between intervention and average school student-SES is expected (*hypothesis 3*) because the intervention is aimed at ambitious goal setting by teachers, and improving student achievement of all students.

At the student level, achievement might differ based on students' gender, SES, initial achievement, and the grade they are in at the moment of testing, therefore achievement will be controlled for these background characteristics. At the student level, comparable with hypothesis 3 at the school level an interaction effect is expected for SES and the intervention: the intervention effect is expected to be higher for low-SES students (*hypothesis 4*).

Furthermore, schools chose one out of three intervention trajectories at the end of the first intervention year. It is expected that schools in which DBDM for mathematics was implemented successfully during this first intervention year chose to continue with DBDM for spelling immediately in or halfway the second intervention year. The intervention effect therefore will probably be largest for schools following the mathematics-spelling-spelling variant, smaller for the mathematics-mathematics-spelling trajectory, and smallest for schools that decided they needed the full two intervention years to implement DBDM for mathematics (*hypothesis 5a and 5b*).

Figure 1. *Conceptual model of the relationship between DBDM and student achievement growth.*



Participants, measures and data collection

School level

At the school level, data was collected on school size, degree of urbanization, average SES, and intervention trajectory variant. In total, 53 schools (1190 team members) fully participated in the study. School teams included on average 22 team members, with a range from 5 to 67. Sample characteristics are depicted in Table 1.

Table 1. *Sample characteristics of schools (N=53)*

School size	Small (<150)	14	(26%)
	Medium (150-350)	31	(58%)
	Large (>350)	8	(15%)
School SES	High	17	(32%)
	Medium	24	(45%)
	Low	12	(23%)
Urbanization	Rural	19	(36%)
	Suburban	23	(43%)
	Urban	11	(21%)
Trajectory	M-M-M	15	(28%)
	M-M-S	13	(25%)
	M-S-S	25	(47%)

Student level

The student achievement on standardized tests were scored on an ongoing ability scale per subject, from grade three to eight. Students take these tests twice a school year (mid and end of school year) with an exception for grade eight, where the test at the end of the school year is scaled differently. This means that there are eleven standardized assessments per student per subject over the course of their primary school career. Over the two years prior to the intervention and the two intervention years, most students took eight tests, leading to eight ability scores per subject, which makes it possible to follow student cohorts and to compare achievement of grades across years. An overview of test occasions is depicted in Figure 2. With approximately 1,500 observations per grade per test moment per school year, the total of observed achievement scores was 66,530.

Figure 2. Overview of measurement occasions. Shadings indicate cohorts.

	Prior to intervention				During intervention			
	School Year 2009-2010		School Year 2010-2011		School Year 2011-2012		School Year 2012-2013	
	Mid (Febr)	End (June)	Mid (Febr)	End (June)	Mid (Febr)	End (June)	Mid (Febr)	End (June)
Grade 3	X	X	X	X	X	X	X	X
Grade 4	X	X	X	X	X	X	X	X
Grade 5	X	X	X	X	X	X	X	X
Grade 6	X	X	X	X	X	X	X	X
Grade 7	X	X	X	X	X	X	X	X
Grade 8	X	-	X	-	X	-	X	-

Next to students' ability scores, the following data was collected at the student level: gender, student weight category indicating SES, and date of birth. Age was centered based on the expected age in months at the time of the test, based on the average age for students who do not accelerate or repeat grades, and thus indicating how many months younger or older a student was than expected.

Data analysis

Given the multilevel structure of the data, with measurements nested within students, and students nested within schools, the *lme4* package (Bates, Maechler, Bolker & Walker, 2013) in R (RCoreTeam, 2013) was used to perform linear mixed effects analyses to investigate and assess effects of the intervention on student achievement.

A full latent growth analysis, where student- and school specific achievement growth are explicitly modeled was numerically not feasible. Therefore, growth was modeled by modeling heterogeneity in (average) student achievement in grade three, grade years three to five, and grade years six to eight, while accounting for differences between measurement occasions in the different grade years in average test performance over students and schools. The differences in average achievements over grades were modeled as fixed effects such that the general mean represents the average performance of students over schools at measurement occasion mid-year grade three. Student and school achievements were allowed to vary across the general mean, which was accomplished by introducing student and school-specific random intercepts.

Furthermore, random effects were introduced for the average achievements over grades three to five and grades six to eight at the level of students. At the level of schools, a random effect was introduced representing the variability in the effect of the intervention across schools. By modeling the differential effect of the intervention, school-specific intervention effects were estimated and schools benefiting from the intervention were identified.

Interpretation of effects

Student achievement was measured using standardized tests with a national benchmark. Based on the benchmark data, the estimated average difference between student scores at two subsequent test moments is approximately 7.7 (Cito, 2009). Since there are approximately five school months between two test occasions, an effect of 1.54 (average of 7.7 ability points, divided by five months of schooling) on average can be interpreted as the expected increase in performance due to one additional month of schooling. This expected effect of an additional month of schooling will differ slightly between lower and higher grades, since the estimated differences in ability scores between two test occasions are larger in the lower grades (Cito, 2009).

Results

Results are depicted in Figure 3 (Model 3) and Figure 5 (Model 5). In Figure 4, random intercepts are plotted against random intervention effects, indicating a larger intervention effect for schools with a lower level of initial achievement.

Figure 3. *Effects in Model 3.*

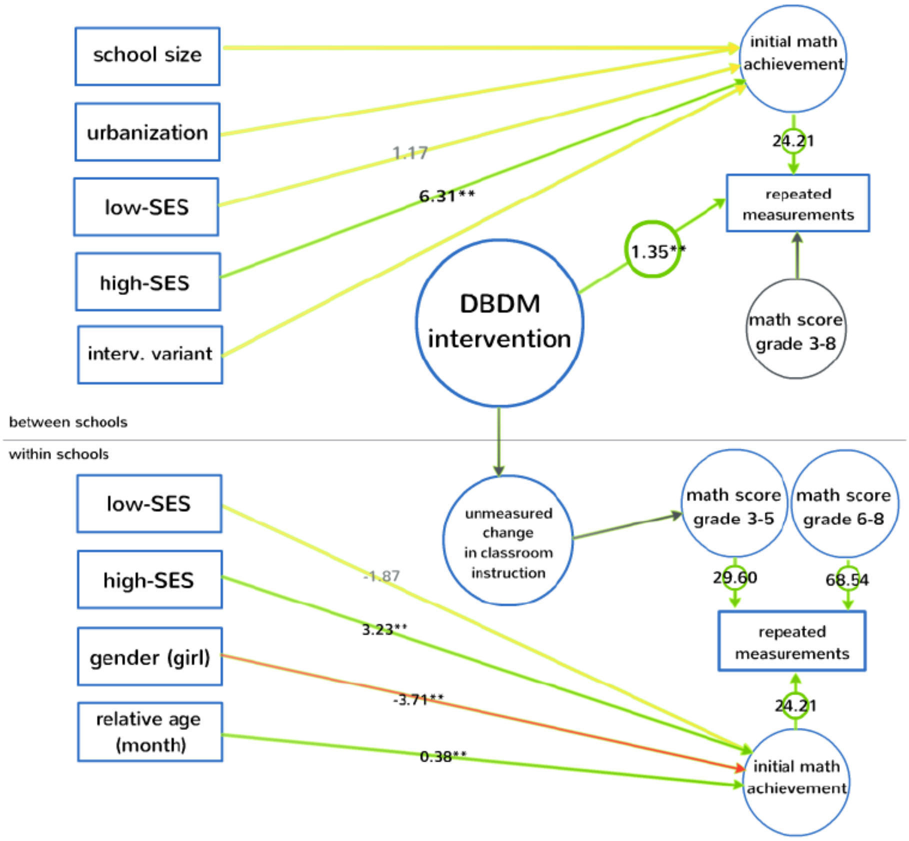


Figure 4. *Random intervention effects plotted against random intercepts (Model 3). Shapes indicate school-SES characteristics.*

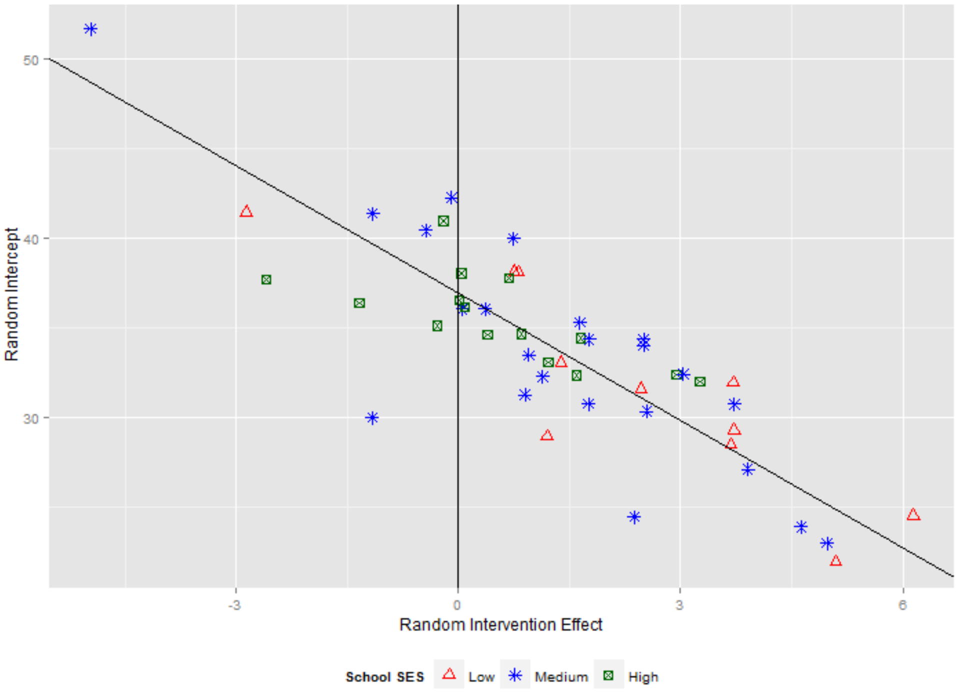
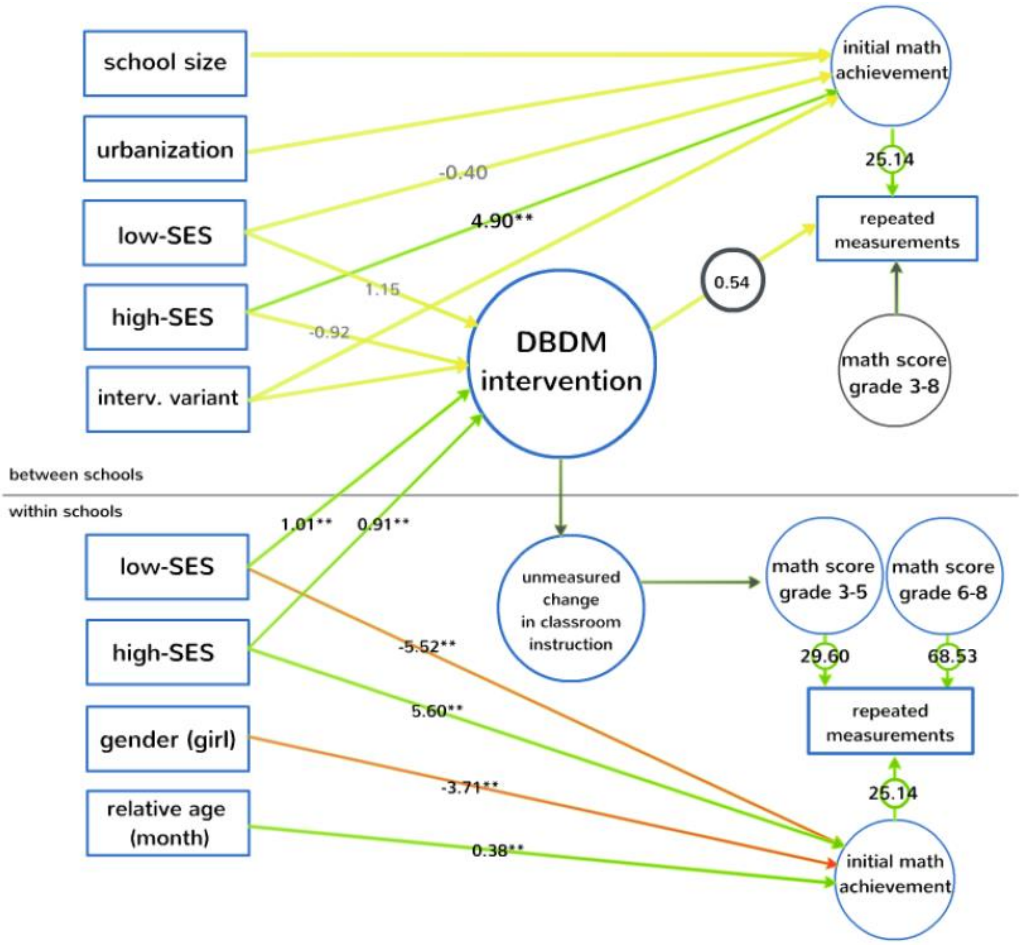


Figure 5. *Effects in Model 5.*



Conclusion & discussion

There is a worldwide interest in the use of data in order to improve education. Many studies focus on the preconditions for successful data based decision making, or describe the process of DBDM in schools, but only very few empirical studies are available on the effects of DBDM on student achievement. The present study is meant to contribute to the international knowledge base on DBDM effects. This was done by investigating heterogeneity in the effects of a DBDM intervention on student achievement for mathematics in 53 primary schools in the Netherlands.

Findings of this study indicate that DBDM can enhance student achievement (hypothesis 1, confirmed), although effects differ across schools (hypothesis 2, confirmed). The fixed effect of intervention without introducing interaction effects is 1.33, indicating an effect of almost an extra month of schooling. Interaction effects suggest that DBDM is especially effective for schools with a large population of low-SES students (hypothesis 3, confirmed). Interestingly, the effects for interaction between student-SES and intervention were not completely in line with expectations (hypothesis 4, confirmed with remark): a positive interaction effect for intervention was found for low-SES students, but the interaction effect was *also positive* for high-SES students. Combining the interaction effects of intervention and student-SES and school-SES leads to the conclusion that the effect of intervention will only lead to negative, but not significant, effect on student achievement for medium-SES students in high-SES schools. An explanation might be that medium-SES students in high-SES schools possibly often belong to the lower scoring students. Since the intervention was aimed at raising achievement for all students, it is possible that teachers decreased the amount of time dedicated to the lowest scoring students in order to distribute attention across all students more equally. However, this does not hold for low-SES students. A further analysis of the data may provide more insight into this effect.

Schools were not allocated to intervention trajectories at random, but were allowed to choose the trajectory of their preference after the first intervention year. The choice for continuing DBDM for mathematics, or broadening the scope of DBDM to spelling during the second intervention year was allowed to be made by schools in order to enhance motivation and commitment. It was expected that this choice would be related to achievement gain during the first intervention year. Analyses however showed that there were no significant differences in achievement or intervention effect across trajectories (hypothesis 5a and 5b, rejected). It may therefore be assumed that schools did not base their choice of an intervention trajectory on the student achievement results during the first intervention year.

The support from the project team finished after the two intervention years, the further implementation and sustainability from then on were schools' own responsibility. Since full implementation of school wide reform can take up to five years (Desimone, 2002), it will be interesting to monitor student achievement and DBDM implementation in the schools that participated in the intervention. Student achievement data in the first school year after completing the intervention will be collected in the summer of 2014 in order to estimate retention effects, and school leaders will be interviewed about the sustainability of DBDM in their school organization.

Further research within this project will focus on the relationship between DBDM effectiveness and the preconditions for successful DBDM, such as school leadership, an achievement-oriented culture, and collaboration within the school team. A follow-up project includes the coaching of teachers regarding to DBDM in the classroom. However, this study already indicates a positive effect of a DBDM intervention on student achievement.

References

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. Available: http://cran.r-project.org/package=lme4.

Campbell, C. & Levin, B. (2008). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability*, *21*(1), 47–65.

Carlson, D., Borman, G. D. & Robinson, M. (2011). A Multistate District-Level Cluster Randomized Trial of the Impact of Data-Driven Reform on Reading and Mathematics Achievement. *Educational Evaluation and Policy Analysis*, *33*(3), 378–398.

Cito (2009). *Rekenen-Wiskunde Handleiding*. Arnhem.

Desimone, L. M. (2002). How Can Comprehensive School Reform Models Be Successfully Implemented? *Review of Educational Research*, *72*(3), 433–479.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E. B., Supovitz, J. A. & Wayman, J. C. (2009). *Using Student Achievement Data to Support Instructional Decision Making*. Washington, DC. Available: http://ies.ed.gov/ncee/wwc/pdf/practice_guides/dddm_pg_092909.pdf.

Ikemoto, G. S. & Marsh, J. A. (2007). Cutting Through the "Data-Driven" Mantra: Different Conceptions of Data-Driven Decision Making. In *Evidence and Decision Making: Yearbook of the National Socieity of Education* (pp. 105–131).

Inspectie van het Onderwijs (2012). *Beoordeling van opbrengsten in het basisonderwijs*.

Lai, M. K. & McNaughton, S. (2013). Analysis and Discussion of Classroom and Achievement Data to Raise Student Achievement. In K. Schildkamp, M. K. Lai & L. Earl (Eds.), *Data-based Decision Making in Education: challenges and opportunities* (pp. 23–48). Dordrecht: Springer.

Lai, M. K. & Schildkamp, K. (2013). Data-based Decision Making: An Overview. In K. Schildkamp, M. K. Lai & L. Earl (Eds.), *Data-based Decision Making in Education: challenges and opportunities* (pp. 9–22). Dordrecht: Springer.

RCoreTeam (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available: http://www.r-project.org/.

Section II:

Generic Competencies in Higher Education

Raffaela Wolf, Doris Zahner, Fiorella Kostoris, Roger Benjamin, Council for Aid Education, New York, USA

A Case Study of an International Performance-Based Assessment of Critical Thinking Skills

Introduction

The measurement of higher-order competencies within a tertiary education system across countries presents methodological challenges due to differences in educational systems, socio-economic factors, and perceptions as to which constructs should be assessed (Blömeke, Zlatkin Troitschanskaia, Kuhn & Fege, 2013). According to Hart Research Associates (2009), there is substantial merit in assessing twenty-first century skills such as critical thinking and writing since about 78% of academic institutions in the United States have established cross-discipline learning outcomes, so called meta domains (Porter, McMaken, Hwang & Yang, 2011), that all undergraduate students should possess upon graduation. Furthermore, changing skill demands of graduating students have been observed around the world since the 1990s (Levy & Murname, 2004). Meeting the demands of today's world requires a shift in assessment strategies to measure the skills now prized in a complex global environment. More specifically, assessments that only foster the recall of factual knowledge have been on the decline, whereas assessments that evoke higher-order cognitive skills have seen an accelerating demand in the twenty-first century. As an example, CAE (the Council for Aid to Education) has been developing assessments that target higher-order skills. The Collegiate Learning Assessment-plus (CLA+) is a measure that emulates critical-thinking and writing skills.

In late 2012, the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) approached CAE proposing a research study to test the feasibility of adapting, translating, and administering CLA+ to higher education students in Italy. The purpose of this feasibility study was twofold. The first purpose was to see if it was possible to assess Italian students' higher-order skills as outlined in Table 1. The second purpose was to see if the Italian students' performance was comparable to their American counterparts.

It is evident that these types of competencies are desirable in many cultures around the globe, regardless of discipline or curriculum. However, measuring competencies within an international framework poses psychometric challenges that pertain to test development, scoring, and the validity of score interpretations (Hambleton & Murphy, 1992). Bias and measurement equivalence (ME) are two different, yet intertwined, pivotal notions that pertain to instrument characteristics in cross-cultural comparisons. Bias is often referred to as nuisance, or confounding factors, whereas equivalence is related to issues concerning the measurement of the instrument (Van de Vijver, 1998). Different forms of bias are considered the main sources of in-equivalence in cross-cultural research (Van de Vijver, 1998; Van de Vijver & Leung, 1997). Bias occurs when observed results systematically distort the relationships between true scores and observed variables. Thus, bias is considered a threat to the validity of the score inferences drawn within a cross-cultural context. There are two main forms of bias: construct and method, where the former refers to unintended differences in the latent constructs, while the latter represents differences in the process of measurement that are due to characteristics of the instrument or administration. Item bias was not considered in the current study.

Construct comparability rests upon the assumption that test scores are contingent upon the same definition of higher-order skills across the countries. If the constructs are comparable, then test score differences across countries may reflect a true representation of the discrepancies in student performance. However, within the context of such comparisons, differences in scores may be influenced by confounding variables, such as test adaptation (e.g. translation), familiarity with item response formats, and many other socio-cultural factors, which introduce method bias. For example, selected-response items (SRQs) are widely used in the United States, whereas many European countries make use of performance or constructed-response tasks (Wolf, 1998). The lack of familiarity with a particular item type could create a source of construct irrelevant variance and, thus, limit the validity of score interpretations. A mixed-format type assessment, consisting of both performance tasks (PTs) and SRQs, can be deemed a viable option in an attempt to ensure test fairness and to reduce the potential impact of bias across cultures.

CLA+ is a mixed-format type assessment; thus this paper presents the results from the feasibility study as a case study of the successful adaption, translation, and administration of CLA+ in 12 Italian institutions. A discussion is provided regarding how different biases may be addressed within an international context. A second analysis examined whether students from Italy and the US ascribe the same meanings to different item formats (PT and SRQs) thus addressing the issue of measurement equivalence and the feasibility of cross-cultural score comparisons. Results are interpreted within a validity framework.

Methodology

Task Selection, Translation, and Adaption of CLA+

CLA+ consists of two sections, a PT and a set of SRQs. ANVUR was presented with an assortment of PT and SRQ sets and a committee of bilingual educators and administrators decided upon the “Parks” PT and a set of SRQs that they felt were culturally appropriate and adaptable for use in the Italian context. The PT and SRQs were then translated and adapted by a third party translation group and eventually verified by ANVUR and CAE staff. ANVUR was provided with a translation and adaptation guide to help facilitate the process. Following the translation and adaptation of the PT and SRQs, ANVUR conducted cognitive labs and a small pilot study, with Italian university students, to verify that the translated and adapted version of CLA+ was clear and elicited the appropriate types of student responses.

CAE adapted its current CLA+ Testing Platform ("CLA+ Platform") to accommodate the adaptation and translation changes made to the "Parks" PT and the 25 SRQs. CAE implemented an additional platform, encompassing text translations as necessary, to facilitate the administration of the tests in Italy. The CLA+ Platform was modified to accommodate student responses in Italian.

Participants

ANVUR recruited 12 universities to participate in this feasibility study, four from three geographical regions (i.e., north, central, and south). The student participants from the 12 universities (n = 5853) comprised of graduating students in their third and fourth year at their respective institutions. These students took the Italian CLA+ during the spring semester of 2013. A sample of American students (n = 4666) were selected for comparative purposes. The American student participants were university freshmen from the fall semester of 2013. The sampled institutions (public and private) consisted of small liberal arts colleges, as well as large research institutions, from the various regions of the United States. Because CLA+ is a newly modified and upgraded version of CLA, the only comparison group available for this study was entering freshmen.

Test Administration

The Italian CLA+ was administered on ANVUR's testing platform. Students had a total of 90 minutes to complete the CLA+, 60 minutes for the PT, and 30 minutes for 20 SRQs. The American students had a similar administration of CLA+ except through a different test delivery platform. The test administration of the Italian CLA+ was vetted and approved by CAE, prior to administration, to assess comparability of the testing platforms. A customized testing platform was created for the Italian students so that testing conditions were uniform between the two countries.

CLA+

CLA+ is a performance-based authentic measure that targets higher-order competencies, such as critical-thinking and written-communication skills, by using a combination of both PTs and SRQs. The adapted version of the CLA+ consisted of one PT and 20 SRQs. Higher-order skills are emulated by presenting authentic tasks, within real-world contexts, in which students must demonstrate those skills. The PTs are designed so that students must get to the bottom of a problem and recommend a course of action after analyzing a document library that contains various sources of information, such as letters, maps, and graphs, just to name a few. As shown in Table 1, the PT is composed of three subscales: analysis and problem solving (identifying, interpreting, evaluating, and synthesizing pertinent information and proposing a solution in terms of how to proceed in case of uncertainty), writing effectiveness (producing an organized and cohesive essay with supporting arguments), and writing mechanics (demonstrating command of standard written English). Similarly to the PT, the SRQs are also developed with the intent to elicit higher-order cognitive skills rather than the recall of factual knowledge. Students are presented with a set of questions that pertain to documents from a range of information sources. The SRQ subscales were identified as critical reading and evaluation (eight items), scientific and quantitative reasoning (seven items), and critique an argument (five items). Students were given 60 minutes to construct a response to the PT and 30 minutes to respond to the 20 SRQs.

Table 1

CLA+ Tasks and Subscales

Task	**Subscale**
PT	Analysis and Problem Solving
	Writing Effectiveness
	Writing Mechanics
SRQ	Critical Reading and Evaluation
	Scientific and Quantitative Reasoning
	Critique an Argument

Scoring

The PT of the adapted version of CLA+ was scored in Italy by a team of trained scorers. CAE representatives led a series of trainings both virtually and on-site in Rome. All responses were assigned raw subscale scores and raw total scores that reflected critical-thinking and writing skills. Total CLA+ scores were computed as a weighted sum of the PT (weighted at .50) and SRQs (weighted at .50).

For the PTs, CAE measurement scientists initially trained three scorers from ANVUR via Skype, followed by an additional in-person training of the Italian lead scorers (one representative from each participating institution plus the three scorers from ANVUR) in Rome. The ANVUR scorers prepared a translated version of the CAE scoring rubric. This team of Italian lead scorers then trained a set of Italian scorers to complete the scoring of the student PT responses.

The CLA+ scoring rubric for the PTs consists of three subscores: Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM). Each of these subscales is scored from a range of 1–6, where 1 is the lowest level of performance and 6 is the highest, with each score pertaining to specific response attributes. For all task types, blank or entirely off-topic responses are flagged for removal from results. Because each prompt may have differing possible arguments or relevant information, scorers receive prompt-specific guidance in addition to the scoring rubrics. Additionally, the reported subscores are not adjusted for difficulty like the overall CLA+ scale scores, and, therefore, are not directly comparable to each other. These PT subscores are intended to facilitate criterion-referenced interpretations, as defined by the rubric.

Analysis and Problem Solving (APS) measures a student's ability to make a logical decision or conclusion (or take a position) and support it with accurate and relevant information (facts, ideas, computed values, or salient features) from the document library.

Writing Effectiveness (WE) assesses a student's ability to construct and organize logically cohesive arguments. This is accomplished by strengthening the writer's position by elaborating on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence).

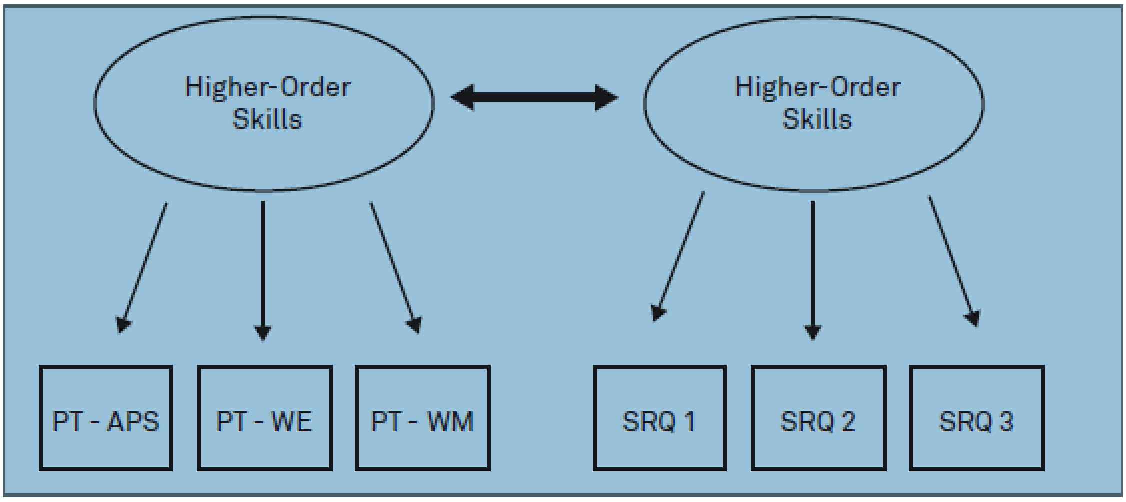
Writing Mechanics (WM) evaluates a student's facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage).

The selected-response section of CLA+ consists of 20 items distributed across three subscales: scientific and quantitative reasoning (seven items), critical reading and evaluation (eight items), and critique an argument (five items). Subscores in these sections are determined according to the number of questions correctly answered, with scores adjusted for the difficulty of the particular question set received.

Data Analysis

Independent sample t-tests were conducted to assess whether there were significant mean differences on the PT and SRQs across countries. In an attempt to examine whether students accredit the same meaning to the different item formats, a multi-group confirmatory factor analysis (MG-CFA) was conducted (Byrne, Shavelson & Muthén, 1989). In the first step, a confirmatory factor analysis (CFA) model was specified that reflected how higher-order skills were theoretically operationalized. A one-factor CFA model, a two-factor CFA model and a higher-order CFA model were tested. The two-factor model had the best model fit in both countries:

Figure 1. Example of Correlated Traits Model with 3 PT subscales and 3 SRQs 4



This model was fitted for the American and Italian students separately to ensure that the same model is valid in each group. Secondly, a baseline model was established by running a common model for both groups with unconstrained parameters. In the third step, several models were estimated to test for ME:

Table 1

Testing for Measurement Invariance with Categorical Data

Model	**Factoring loadings**	**Thresholds**	**Residual variances**	**Factor means**	**Factor Variancies**
Configural invariance	*	*	Fixed at 1	Fixed at 0	Fixed at 1
Strong invariance (1)	Fixed	Fixed	Fixed at 1	Fixed at 0/*	Fixed at 1
Strong invariance (2)	Fixed	Fixed	Fixed at 1	Fixed at 0/*	Fixed at 1/*

Note. The * indicates that the parameter is freely estimated. Fixed at 0/*= the factor means are fixed at 0 in one group and freely estimated in the other group. Fixed at 1/* = the factor variance is fixed at 1 in one group and freely estimated in the other group.

The various models were fit using an adjusted weighted least squares (WLSM) algorithm using the Mplus software (Muthén & Muthén, 2010). All model in this analysis were evaluated in terms of goodness of fit criteria. Exact fit was evaluated using the model χ2, whereas close fit was evaluated using the comparative fit index (CFI), Tucker-Lewis non-normed fit index (TLI), and root mean squared error of approximation (RMSEA). In this study, values of less than .05 were used for the RMSEA and values greater than .95 were used for the TLI (Hu & Bentler, 1999). All fit indices were used conjunctively to determine model fit.

Results

Descriptive Statistics

Table 1 provides descriptive statistics for the adapted CLA+. Both countries showed similar results for the PT (Italy: M = 9.17, SD = 2.95 ; US: M = 9.06, SD = 2.54), whereas the sample from Italy had a higher mean on the SRQs (M = 12.31, SD = 2.85) compared to the American sample (M = 10.64, SD = 3.62). Independent sample t-tests showed statistically significant differences on the SRQs (t (10564) = 25.82, p<.001) but not on the PT task. However, it is uncertain whether these differences are due to true differences in performance or whether the familiarity with item types across cultures introduced nuisance variability.

Table 1

Descriptive statistics for CLA+ for Italian vs. American students

	Italy		US	
	SRQ	**PT**	SRQ	PT
Items (N)	20	1	20	1
Students (N)	5853	5853	4638	4638
Min Score	0	3	0	3
Max Score	19	18	19	18
Mean	12.31	9.17	10.64	9.06
SD	2.85	2.95	3.62	2.54

Factor Analyses Results

The first step was to test whether the proposed two-factor model fits the empirical data for each group. Results indicate that the hypothesized model is supported in both groups (Italian: χ2 = 1280.05; df = 229; RMSEA = .028; CFI = .989; TLI = .988; American: χ2 = 2203.51; df = 229; RMSEA = .043; CFI = .992; TLI = .992). The second step was to move from a single-group CFA to MG-CFA in order to cross-validate the two-factor model across the two groups (configural invariance). Table 1 indicates that Model 1 provided a good fit (χ2 = 3455.13 ; df = 458 ; RMSEA = .035 ; CFI = .99; TLI = .99) to the data, indicating that the factorial structure of the construct is equal across the two groups. In other words, examinees ascribe the same meaning to the definition of higher-order skills across countries. Given that configural invariance was confirmed, the factor loadings and thresholds were then constrained to be equal to test for strong invariance. Model 2 fit significantly worse than Model 1, DIFFTEST(56) = 13239.55, p<.001, and Model 3 fit significantly worse than Model 2, DIFFT-

EST(2) = 1402.13, p<.001. These results suggest that students may have ascribed different meanings to the item formats across countries.

Table 2

Fit indices for invariance tests

	X^2	df	RMSEA	CFI	TLI
Model 1: Baseline (Configural invariance)	3455.13	458	.035	.99	.99
Model 2: Strong Invariance (1)	25166.68	514	.095	.92	.92
Model 3: Strong Invariance (2)	23764.55	512	.093	.92	.92

Table 3

	X^2	p
Model 1 vs Model 2	13239.55	<.001
Model 2 vs Model 3	1402.13	<.001

Discussion

The feasibility of assessing higher-order skills in two different cultures was confirmed in this study. Cross-cultural studies aim to address the question to whether valid test score inferences can be drawn across different cultural populations. This case study was an attempt to address bias as a function of the interpretation of test scores rather than an inherent property of the instrument. It is well known that test adaptations or translations are prone to introducing different types of biases (Hambleton, 1996), such as construct, method, and item bias. In this feasibility study, translation effects were mitigated through the implementation of a multi-stage translation process. Through the combined effort of colleagues and content-area experts from each culture it was possible to specify and examine the similarities in the underlying construct definition of higher-order skills and the alignment of the items with the test blueprint. As part of the adaptation phase, a small pilot study was conducted in Italy to ensure that items on the instrument were functioning as intended. Consequently, it was determined that Italian and American students appear to associate the same meaning to the definition of higher-order skills and that the items on the instrument were adequately sampled from the domain of higher-order skills. The appropriateness of construct representativeness across countries was confirmed by the results of the CFA analyses.

Method bias may be introduced through administration procedures and/or differences that pertain to the instrument itself. The test administration platform of the Italian CLA+ was examined by CAE prior to administration to ensure comparability of the testing platforms. In order to circumvent problems due to rater effects, specific scoring rubrics and guidelines were developed, and graders underwent rigorous training sessions that were facilitated through the joint effort of both countries. However, there was reason to believe that the use of different item formats could be a source of method bias since familiarity with item types varies by culture (Wolf, 1998). Post-hoc statistical analyses were conducted in an attempt to examine whether examinees from Italy and the US ascribe the same meaning to the PT and SRQs. According to these results, it is evident that higher-order skills were assessed in both countries. However, students appeared to associate different meanings with different item types across countries, which imposes the question as to whether valid score inference can be drawn from direct score comparisons of students in different countries. Psychometric evidence exists for providing valid score inferences within each country due to the successful adaptation of CLA+. However, direct score comparisons across countries should be made with caution because a total score that is comprised of PT and SRQ scores may have an altered meaning in both countries due to the dissimilar meanings that are associated with different item types. This could be due to the differences in the two populations, which is a limitation of the current study. CLA+ is a newly modified and upgraded version of the CLA; thus, the only comparison group available for this study was entering freshmen who were compared to graduating students in Italy. This implies that the groups may have varied in ability, which was not accounted for in the analyses. Plans for a future analysis include the use of U.S. CLA+ senior data in order to examine whether the effect of growth in higher-order skills from freshmen to graduating seniors may have had an impact on the results of the current study. Furthermore, when interpreting the test scores across countries, other factors that could impact test score results, such as student motivation and/or socio-economic status, need to be addressed.

During the last few decades, bias has predominantly been associated with item bias or differential item functioning; methods to address construct and method bias often appear to be neglected. While the importance of addressing item bias is evident in cross-cultural research, it is also apparent that cross-cultural comparisons can further be challenged by construct irrelevant sources of variance that go beyond individual items. Perhaps an ongoing effort, including both a priori and post-hoc considerations, could provide fruitful information in terms of construct, method, and item bias. Rather than viewing and/or treating each component in isolation, a holistic approach that combines these sources could ensure high standards in all stages of the test development and adaptation process, consequently aiding in the collection of evidence for valid cross-cultural score interpretations.

Some suggestions for future a priori activities include a focus on collaborative efforts between measurement scientists, cognitive scientists, and experts within the tertiary education system from both cultures in an attempt to develop instruments that are within appropriate cultural contexts. Different translation procedures also may be combined to ensure adequate translations. The translated instrument could be pilot tested with bilingual students to assess the appropriateness of the adapted version. However, findings may need to be interpreted with caution since the bilingual students may not be representative of the target population. In an attempt to minimize method bias, it may be worthwhile to provide practice items so that students from different cultures can become accustomed to different item formats. Individual items also should be reviewed in terms of poor translation, complex wording of items, and whether items invoke unintended additional abilities. Statistical analyses at the item level, such as differential item functioning, should be integrated into the item development process to ensure appropriateness of translated items. Comparisons of item statistics in the two versions of the instrument should consider controlling for any ability differences in the two groups.

Bias is often perceived as a nuisance factor (Van de Vijver, 1998) and thus many statistical procedures exist in an attempt to mitigate or reduce the unwanted effects of bias on cross-cultural score comparisons. However, if bias would be neglected, then perhaps one could gain information in terms of systematic cross-cultural differences, which may indeed be beneficial to the instrument development process. This would also aid in the collection of validity evidence to ensure appropriate cross-cultural comparisons. In sum, it is feasible to assess higher-order skills globally. However, in a collaborative effort across nations, numerous factors need to be taken into consideration prior, during, and after the test adaptation phase to ensure that valid cross-cultural score inferences can be drawn from the data.

References

Arum, R. & Roksa, J. (2011). Academically Adrift: Limited Learning on College Campuses. Chicago, Ill.: University of Chicago Press.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C. & Fege, J. (2013). Modeling and Measuring Competencies in Higher Education: Springer.

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105(3), 456.

Hambleton, R. K. (1996). Guidelines for Adapting Educational and Psychological Tests.

Hambleton, R. K. & Murphy, E. (1992). A psychometric perspective on authentic measurement. Applied Measurement in Education, 5(1), 1-16.

Hart Research Associates. (2009). Learning and Assessment: Trends in Undergraduate Education - A Survey Among Members of The Association of American Colleges and Universities. Washington, DC: Hart Research Associates.

Hu, L. t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1-55.

Levy, F. & Murname, R. J. (2004). Education and the Changing Job Market: An Education Centered on Complex Thinking and Communicating is a Graduate’s Passport to Prosperity. Educational Leadership, 62(2), 80-83.

Muthén, B. & Muthén, L. (2010). Mplus Version 6.1 [Software]. Los Angeles, CA: Author.

Porter, A., McMaken, J., Hwang, J. & Yang, R. (2011). Common Core Standards: The New US Intended Curriculum. Educational Researcher, 40(3), 103-116.

Van de Vijver, F. J. (1998). Towards a theory of bias and equivalence. Zuma Nachrichten, 3, 41-65.

Van de Vijver, F. J. & Leung, K. (1997). Methods and data analysis for cross-cultural research (Vol. 1): Sage.

Wolf, R. M. (1998). Validity issues in international assessments. International journal of educational research, 29(6), 491-501.

Klaus Beck, University of Mainz, Germany

Comment on- A Case Study of an International Performance-Based Assessment of Critical Thinking Skills

Let me start my comments on this paper with a personal remark which underpins the importance of those parts of your study in which the problem of careful translation is dealt with and which you might feel to be amazing: It was in 1948 and the following time, some years after World War II, when the United States decided to support Germany in becoming again a democratic republic. One facet of this support was the so-called Marshall-plan, officially the “European Recovery Plan”. This plan had been set up to provide European countries with economic resources including food. In realizing this plan Americans asked Germans what they were in need most urgently. No doubt, among the pressing needs for the ahungered population was flour to bake bread. The German umbrella term for wheat, rhy and sorghum is “*Korn*”. So, Germans messaged that they are in need of “*corn*”, not remarking that the translation of the English “corn” into German is “*Mais*”. This was the reason that Germany got corn in bulk but no wheat, no rhy, no sorghum. Germans, though very thankful for everything to eat were astonished that American people eat bread made of “*Mais*” which for our tongue - to say the least - tasted rather strange.

Though I was a very young boy at that time I remember very well that I did not like this *Mais*-bread made of corn and I grew up with the imagination that Americans are pretty different - a prejudice which wonderfully broke down when I met first with a big fat T-bone steak.

Now, thank you for your clear presentation and also for your full paper which I got in advance! I liked very much the clarity and the well-structured argumentation of the report on your interesting and also very necessary study. Though there are already some trials to adapt measuring instruments - as far as I know exclusively from English into other languages and I wonder when it will come for a first time the other way round - this field becomes more and more important, especially in respect to higher education. One reason is, of course, the megatrend of globalization and within this development a growing movement which we are used to call brain circulation. As to this we are in need of instruments allowing to assess not only general and domain specific knowledge but also cognitive competencies and capacities of academics who are ready and willing to go abroad and to apply for jobs in foreign countries.

Therefore it is a work of merit that your study on critical thinking skills goes beyond the measurement of mere knowledge using a modified version of the CLA (Collegiate Learning Assessment). As all our cognitive activities are embedded in any content it could be of interest to learn more about the conception of the generic domain the CLA+ is focused on and to which extent the results got from its administration might be generalizable across diverse academic disciplines and fields of practice.

Contrary to the many pragmatic and often less careful adaptions of tests from one language to another it is important to take into account that transgression of language boundaries does usually mean: transition from culture to culture implying that we meet here not only with a problem of *translation* but also with the more demanding problem of *adaption*.

By the way this is not always the case. Think of bi-lingual or even multilingual countries like Switzerland or Belgium or – topically in these days – Ukraine where adoption of a test does not imply the transition of a cultural boarder and thus making adaption superfluous. The other way round and, maybe, more important: Looking at large countries like the US or Russia or China where people speak and understand – in principle – the same (standard) language it seems to be rather sure that different regions or even big cities provide differences of culture (not to speak of social subcultures) which might be quite relevant in our context and thus necessitate adaption within one and the same language – a problem which has not been reconsidered with all its facets in the back dating big discussion on “cultural fairness of tests” (cf. e.g. Mensh & Mensh 1991). Cultural differences of this type might be more substantial than cultural differences, say, between Denmark and Sweden. To me a careful look at these circumstances seems to be as necessary as it is difficult – all the more with respect to multicultural classes at universities where we nowadays find rarely cultural homogeneous groups.

Following Van de Vijver and Ronald Hambleton (1996) and according to the International Test Commission (2010) you observed and analyzed two of the three types of bias relevant in test adaption, i.e. construct bias and method bias, leaving out item bias because of time restrictions and perhaps also because of confidentiality of your measuring instrument. Nevertheless, one might be eager to learn more about the content of CLA+ to get better insight in the problem of domain specificity mentioned before.

In any case, let me raise some questions and address some interesting points dealing with the two other bias problems covered in your study, namely construct and method bias:

1. In your report it is said that an Italian committee of ANVUR “was presented with an assortment of PT [performance tasks] and SRQ sets [selected-response items]” from which they chose tasks “they felt culturally appropriate”. Are the “feelings” apt and reliable enough to ensure cultural appropriateness and also to allow for an item selection which represents the construct adequately?

2. I was a little bit surprised that you took Wolf’s statement (1998, 495) for granted that in Europe constructed-response tasks are more widely-used than selected-response items. Though I have

no empirical evidence on this I dare to claim that all over Europe in performance tests multiple choice-items are prevailing, even in Italy, at least since the European-wide Bologna-reform of tertiary education has been established. By this reform the number of exams students have to pass has been multiplied. As a consequence assessments can no longer be handled if they require that teachers read written statements of more or less length which is a rather time-consuming effort.

The results of your study seem to confirm my guess because Italian students perform better on SRQs than the US students and on PT items both perform nearly equally.

3. Again, if your concerns would apply that Italian students are not so familiar with the SRQ-format and the US students are not so familiar with the PT-format: Doesn't this mean that these alternate disadvantages could compensate each other reciprocally to a more or less high extent? Are we in this case in need of weighting scores resulting from the two different item types? How could we quantify the grade of familiarity with the two item formats?

4. Constructed responses as reactions to stimuli given by PTs typically originate from more authentic stimulus configurations than MC items do. Therefore they provide more valuable information for an assessment than answers to MC-items. On the other hand it is much more complicated to distill and theoretically categorize this information in an objective and reliable way. Given this, I missed in your paper a little bit some data on inter-rater reliability occurring in the analysis of the performance tasks. And I didn't get whether the Italian students answered in their mother tongue or in English. In both cases we meet with two different inter-rater reliability problems, namely the within and the between language degree of accordance of raters. The within language reliability problem is well known and has been widely analyzed. But if Italian and American raters have to assess one and the same text and to assign its content to categories expressed in Italian and English language we have not only the problem whether these categories are sufficiently semantically equivalent but also the other problem whether this is true for the texts to be assessed if they have been translated from Italian to English. I cannot estimate to which extent these problems can be overcome by drawing on bi-lingual persons though at least at a first glance this seems to be the silver bullet for their solution. Rather, one has still to control for the question whether these persons are also "bi-*cultural*", i.e. whether they are sufficiently enough familiar with the living environments of both cultures. Again, I have no idea how this could be done satisfyingly. Additionally – and I add this far beyond any critique of your paper which reports state-of-the-art measures for dealing adequately with the problem of reliability – we know that even intensive training of graders can result in an erroneous consent on the meaning of certain expressions or phrases!

5. Next, your study is comprising graduate students in Italy and freshmen in the US. Therefore your findings showing an overall advantage of the Italian over the US students seem to be pretty plausible. But, as Van de Vijver and Hambleton state and as you yourself are discussing in your paper, it is necessary to test equivalent groups in respect to duration of their study and, additionally, by controlling for curriculum differences and for opportunities to learn in general. So, which conclusion would you draw if the US freshmen after three years show up with the expected gain in higher-order skills? Would, then, your concerns related to item format bias (PT vs. SQR) disappear though you haven't got additional independent information on the problem of item bias? And which conclusion would one have to draw if after three years US graduate students still show to be behind their Italian counterparts? Of course, and you discuss this also in your paper, there might exist different reasons for such a finding (e.g. student motivation, socio-economic status, differences in curricula, other learning opportunities). But these reasons are in doubt only the causes for differences in measuring outcomes. The crucial question for an international comparison still remains to be answered whether test scores include some hidden biases of the types sketched above. Your suggestion to follow a more holistic approach in trying to avoid bias seems to me very interesting. And I am really excited to learn more about your ideas how this could be done in detail.

6. As far as I can see in all adaption efforts on measuring instruments experts are involved to judge the adequacy of constructs and their operationalization as well as to rate on cultural appropriateness of test tasks and of (ideal) solutions. Difficult questions arise from this measure which, as far as I can see, still wait for a systematic and careful analysis:

- Which are the criteria allowing for the assessment of expertise of experts? Are we in need of tests measuring expert competence? Putting this question: Do we fall in the trap of infinite regress because we then are in need of experts who help us developing an expert test?

- How many experts are to be included? Does this depend on the type of problem which we are faced with (all the variants of validity and of reliability)? Is one expert enough – if she or he is really an "expert"? Why, or rather in which cases, do we need more than one expert and if so how many exactly? Or should we say: The more the better? Do we meet here with a problem of representativeness? Or – totally different – have we to deal with a question of truth which could be answered correctly or incorrectly, in terms of "right" or "wrong"? To give an example: The interpretation and assessment of a given statement produced by a student within a performance task (PT) could be a matter of finding out what people of this target group usually mean by uttering a statement like this (a problem of representativeness which can only be solved by an adequately sampled majority of "experts") or it could be a matter of judging whether the interpretation or the assessment is correct or not (a problem of truth which is not at all an issue of acclamation and therefore can be solved by one competent “expert”).

- In which way have we to deal with different or even contradictory statements of experts – an experience which does occur more often than we like it? Even teachers and university professors do not always agree on (best) answers to questions we would like to put in SRQ-formatted tests not to speak of PTs (cf. Beck & Krumm 1994, 193-195). Do we have any criteria other than tests (see above) at our disposal to arrange a hierarchy of credibility or dependability of experts or to weight their expertise? Why should we include experts in our studies if they do not deserve to be on the very top of scales of these types? And if they were all on top why to have more than one (see above)? Could years of experience serve as a valid measure? Are they at least a sufficient indicator for expertise? Experts are often called “experts” because they command different fields of practice. This is especially true in cases of generic competences like e.g. “critical thinking” which consists of a general and of a domain specific component as well. From which field of practice should we choose them? Do we here again stumble across a problem of representativeness (see above)? Does it make any sense to compute a mean on the basis of diverging expert ratings? In which case could this be an adequate approach?

I am afraid that the higher the pedagogical costs at risk in decisions based on tests developed by inclusion of expert opinions touching substantial features of the instrument the less one can take the responsibility for such decisions. And I refrain from speculating whether in this sense comparative studies in general fall into a section of lower risk.

7. Lastly and at least: I am curious about your results if you compute achievement scores on critical thinking separately for the industrial North of Italy, the touristic middle and the clever Cosa Nostra South. Imagine that you would find a South-North-decline of scores. In Europe – as a joke – we would be inclined to say that this could be named best as the “Berlusconi-effect”!

References

Beck, K. & Krumm, V. (1994.) Economic Literacy in German-speaking Countries and the United States: Methods and first Results of a Comparative Study. In: Walstad, W. B. (Ed.). An International Perspektive on Economic Education. Boston: Kluwer, 183-201.

International Test Commission (2010). International Test Commission Guidelines for translating and Adapting Tests. [http://www.intestcom.org]

Mensh, E. & Mensch, H. (1991).The IQ Mythology: Class, Race, Gender, and Inequality.Carbondale, IL.:Southern Illinois Univ. Pr.

Van de Vijver, F. & Hambleton, R. K. (1996). Translating Tests: Some Practical Guidelines. European Psychologist, 1(2), 89-99.

Wolf, R. M. (1998). Validity Iissues in International Assessments. International Journal of Educational Research, 29(6), 491-501.

Li Cao, University of West Georgia, USA

Comment on - Useful Strategies in Dealing with Primary Scientific Literature: An Expert-Novice Comparison (KOSWO)

Dear Elizabeth Schmidt:

This is to elaborate on my notes (of April 2, 2014) regarding your paper. My comments are based on the 5-page paper that I received before the AERA meeting. My purpose is to share my response to the paper and offer some suggestions to improve the paper and the project.

Overview of the study

Your paper reported a study that examined the quality of written descriptions of research purpose and research design between doctoral students (N = 21) and first-year undergraduate students (N = 16) of psychology. Participants in each group were asked to read two three-page psychology conference papers each in 15 minutes and then “(1) to describe the aim of the study and (2) to describe the experimental design realized in the study” (p. 3). The written responses were scored by two independent raters with a 6-point rating scale. Results of the two separate t-tests supported the hypotheses that the doctoral students group scored significantly higher on both questions as compared to the undergraduate student group. The paper suggested further analysis of the think-aloud protocols and the written notes taken during the process of reading the two articles. The intent was to design a training to help university students read primary scientific literature more effectively.

Significance of the study

The reported study joined the recent research on improving university student ability in reading scientific literature. The ability to comprehend, analyze, and evaluate primary scientific literature is fundamental for university students to complete their training in school and to function as professionals in the future. Research in this area is of great significance in understanding the reading process and designing effective training programs, as well as in extending the research of the expert-novice difference in reading. The current study focused on differences in performance and strategy use in reading between expert and novice students. The topic is partially relevant and interesting to the STEM community at large and to the reading research in particular. While the paper reported some interesting findings, addressing the following issues could improve the paper and project. Below are my concerns and suggestions:

Title

The current title of your paper is **Useful Strategies in Dealing With Primary Scientific Literature: An Expert-Novice Comparison.** However, the paper only reported data about performance. I understand the paper is part of a large project and you may have collected data on the strategy and process of reading. However, you may want to reword the title so that it reflects the data and the theme of the actual paper more accurately.

Conceptual Framework

This section did a good job in setting up the context for the study by identifying the research gap and providing a clear purpose statement. Specifically, this section described that "Previous research has mainly focused on the skilled processing of textbook materials, but only occasionally on the literacy reception of PSL. Hence, there is a need to examine students' ways of dealing with PSL and to identify the strategies that help to process it more effectively" (p. 2). This research gap has been translated into a clear statement of the study purpose: "The aim of the present study was to examine which strategies are useful when dealing with PSL" (p. 2).

Along the line of my comment on the title above, I have two concerns on this section. One is concerned with the consistence of the theme and research question and design of the paper. Your paper says that "The aim of the present study was to examine which **strategies** are useful when dealing with PSL. For this purpose, we compared scientists' and students' **performance** with regard to a deep understanding of the content of two texts" (p. 2). If the focus is on development of the university students' **strategy** in reading primary scientific literature, this should become the focus in the review of the literature and in your research question.

Related was my other concern regarding review of the literature to set up the stage for your study. Your paper included a total of six references. Of these six references, five were from the research team members and the only external reference is of three decades old, i.e., (Paris & Jacob, 1984). As my previous notes (Appendix) to you suggest, there is a large body of literature on promoting university student ability in reading scientific literature. Incorporating this literature will help you (1) describing the current status of the research of reading in science, in relation to expert-novice research (https://www.google.com/#q=expert+novice+differences), (2) identifying the problem for research, and (3) establishing the conceptual framework of the study, particularly in regards to the importance and appropriateness of the research questions.

Method

Design---The study adopted a passive response design to collect participants' responses to the two conferences papers. Participants were required to read one paper silently and to think aloud while reading the other paper. The reading modes were counterbalanced across the two participant groups to avoid the possible order effect on the participants' responses to the readings.

Participants---The study included two groups of participants. The expert group was consisted of 21 doctoral and post-doctoral students in psychology while the novice group included 16 first-year undergraduate students in psychology. While the sample size was relative small, both group shared the same academic background. The paper reported a 10-year difference of age between the two groups. You may also want to explore possible influence of age, as an independent variable or a co-variate, on the strategy use and performance in reading. Normally, people's reading ability improves with more experience.

Materials---The study used two psychology conference papers as the reading materials. The advantage of using the materials in one discipline is the elimination of subject matter difference since both groups of participants shared psychology as their background. The disadvantage of this design is its limitation of generalizability of the results to other disciplines. This limitation needs to be justified in the Method section and discussed in the Discussion section of the paper. This issue concerns with scope of the study as indicated in the title of the paper. Scientific Literature includes psychology and many other areas far beyond. You may want to consider multiple studies to manipulate this variable so as to increase generalizability of your results to a broader extent. For instance, study 1 uses materials in social science, e.g., psychology; study 2 uses materials in hard science, e.g., computer science; and study 3 uses materials from a mix of multiple disciplines etc. As suggested above, you may want to consider a similar design to include students from different disciplines in order to enhance the external validity of your study.

Procedure---The paper said the written responses were collected after a brief time-period. A description of the length of this time period would better inform the readers of your paper. Recent research in metacognition and self-regulated learning found that compared to the immediate responses, the delayed responses are a more robust correlate to monitory accuracy and academic performance (Dunlosky & Metcalfe, 2009; Dunlosky & Nelson, 1992; Thiede, Anderson & Therriault, 2003; Thiede, Dunlosky, Griffin & Wiley, 2005).

Result

The paper reported the results of two t-tests for data analysis. The two graphs clearly demonstrate the differences of understanding between the two groups. However, relying solely on scores of the written responses placed a serious limit to data analysis and made the data hardly convincing in addressing the proposed research question. As your Discussion section suggested, the study could use a mixed-method design and use results of both qualitative and quantitative data to address the research question.

Discussion

Due to the above issues, the Discussion section presented the several unwarranted statements. This section stated with the statement "The aim of the present study was to examine differences between experts and novices in terms of the reading behavior and understanding of PSL." In the present paper, no data were reported in terms of the reading behavior. Similarly, the paper offered no data to support the statement that "We assume that the experts' strategies lead to deeper processing" and the ending discussion that "This would correspond with Berthold and Renkl's (2010) findings that indicate that novice learners process texts in a more mentally passive manner rather than deeply or focused processing them, a process that would result in a better learning outcome" (p. 5). The purpose of the Discussion is to interpret the data and discuss their theoretical and practical implications in relation to the existing literature. The current Discussion tried to do that. However, it went a bit too far beyond what your data warranted.

Summary

The paper reported a study that followed the expert-novice research paradigm to examine differences of the written responses between doctoral students and undergraduate students in reading two conference papers in psychology. The study found doctoral students scored significantly higher than the undergraduate group in describing the purpose and research design after reading the papers. The study is of great significance in extending the expert-novice research, particularly in promoting university students' ability in reading primary scientific literature. The paper could be strengthened by developing a theoretical framework to guide the study, relating the study to the existing research of the past three decades, adopting a mixed-methods research design, and using more sophisticated data analysis to address the proposed research questions.

References

Dunlosky, J. & Metcalfe, J. (2009). Metacognition. Beverly Hills, CA: Sage.

Dunlosky, J. & Nelson, T.O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. Memory Cognition, 20(4), 374-380.

Thiede, K., W., Anderson, M. C. M. & Therriault, D. (2003). Accuracy of Metacognitive Monitoring Affects Learning of Texts, Journal of Educational Psychology, 95(1), 66–73.

Thiede, K., Dunlosky, J., Griffin, T. & Wiley, J. (2005). Understanding the Delayed-Keyword Effect on Metacomprehension Accuracy. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Section III:

Teacher Training in STEM Fields

Sigrid Blömeke, Simone Dunekacke, Lars Jenßen, Thomas Koinzer, Wibke Baack, Marianne Grassmann, Humboldt University Berlin, Germany

Martina Tengler, Hartmut Wedekind, Alice Salomon University of Applied Science, Berlin, Germany

Effects of opportunities to learn mathematics on pre-school teachers' mathematics pedagogical content knowledge (KomMa)

State of Research and objectives

- Development of early mathematics ability depends on support

  (van Oers, 2009)

- Research gap with respect to pre-school teachers‘ competence to provide this support in the informal setting of pre-schools

  (National Advisory Panel, 2008)

- OTL in mathematics not stressed in pre-school teachers‘ training

  (Aubrey, 1994; Copley, 2004)

Objectives

1) Conceptual model of pre-school teachers‘ competence
2) Development of a reliable and valid standardized assessment
3) Data-based model of the competence structure, levels and development during training and transition into pre-school
4) Identification of important OTL predicting teacher competence

Conceptual Model of Competence

Def.: Competence is understood as a domain-specific multidimensional latent trait including cognitive abilities and affective-motivational facets underlying performance in real-world situations

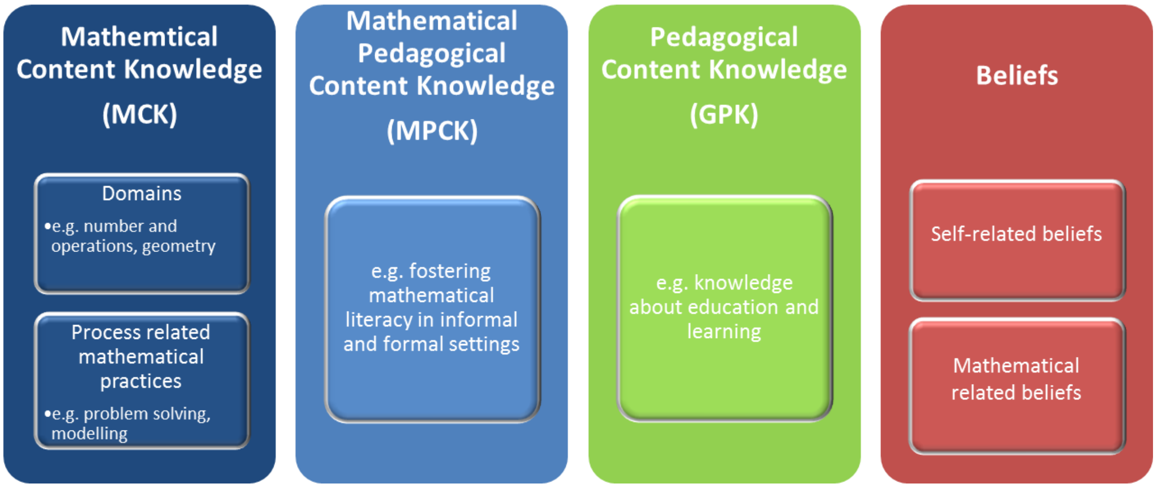
Method:

1) Half-standardized qualitative analysis of pre-school standards from all 16 states
2) Analysis of p-s teacher education curricula
3) Distinction of facets (Shulman, 1986; TEDS-M)

4) Identification of cognitive and affective- motivational characteristics by mapping pre-school standards and teacher education curricula (including identification of gaps)

Results: 4 major real-world situations

- „Utilizing informal settings to foster mathematics ability“
- „Development of children‘s math ability“
- „Diagnosing math ability“
- „Teaching math to kindergarten children“


Mathemtical Content Knowledge (MCK)
Domains
•e.g. number and operations, geometry
Process related mathematical practices
•e.g. problem solving, modelling
Mathematical Pedagogical Content Knowledge (MPCK)
e.g. fostering mathematical literacy in informal and formal settings
Pedagogical Content Knowledge (GPK)
e.g. knowledge about education and learning
Beliefs
Self-related beliefs
Mathematical related beliefs

Content validity confirmed by expert panel (Jenßen et al., in press).

MPCK assessment

Method:

1) Using the model as a heuristic
2) Development of a large item pool
3) Cognitive lab with p-s teacher education students
4) Content validity (items/test): expert panel
5) Field test with students from different states

Results: Test with 21 items covering crucial job requirements

Ex.: “You are playing “shopping tour” with two children at the daycare center. The game provides an opportunity for the children to become familiar with money as a quantity. You are actively involved in this game by taking over the role of a customer or a salesperson. Please give two brief examples how you would utilize this situation to foster the children’s mathematics ability.”

Exemplary answers accepted as correct:

Buying several goods with a specific amount of money

Appropriately price tagging different goods

Returning change in different ways

Method:

1) Using the model as a heuristic
2) Development of a large item pool
3) Cognitive lab with K-teacher education students
4) Content validity (items/test): expert panel
5) Field test with students from different states

Results: Variance =.7, split-half reliability = .8, EAP = .6

- Inter-rater reliability of CR items: Yules Y = .5-.9
- Construct validity: CFA as hypothesized (r = .4-.8)
- Criterion validity: MPCK predicts p-s teachers‘ action planning mediated by perception (Dunekacke, in press; Dunekacke et al., in press)

More validity studies are currently carried out (e.g., convergent and discriminant validity to other professions, observations).

OTL in pre-school teacher training

- Teachers‘ OTL only barely surveyed in a standardized way

  (Blömeke, Fellbrich, Müller, Kaiser & Lehmann, 2008)
- Significant impact on MPCK (besides background)

  (Sarama & Clements, 2009; Kwong et al., 2007; Blömeke, Suhl, Kaiser & Döhrmann, 2014)
- Pre-school teachers‘ training curricula: (post-)secondary institutions, generalists, all age groups, broad requirements

  (Metzinger,2006)
- OTL in math, math pedagogy vary across states/ institutions (Dunekacke et al., 2013)
- 11 items („To what extent …“) „Utilizing informal settings to foster math ability“, „Development of children‘s math ability“, „Diagnosing math ability“, „Teaching math to kindergarten children

OTL	Range	M	no.	Rel.
OLT1	1-4	1.9	4	0.9
OLT2	1-4	1.3	3	0.9
OLT3	1-4	0.8	2	0.7
OLT4	1-4	0.7	2	0.9

Method and results

Sample (Pilot study):

- N=354 future pre-school teachers (6 schools/15 classes)
- Gender: 83% = female; Age: M = 23 years (SD = 4 years)
- Status: 1st year = 41%, 2nd = 33%, 3rd = 10%, 4th = 16%
- School grades: math = 3.2 (SD = 1.1), German = 2.6 (SD = 0.8)

Data analysis:

- Latent regression analyses; ConQuest 2.0 (Wu et al., 2007)

First results based on preliminary data:

- Program year significant
- More variance explained by specific data (4 OTL scales)
- Particularly relevant „Utilizing informal settings to foster mathematics ability“, „Development of children‘s math ability“
  (Blömeke et al., in preparation)

Summary and Discussion

1) Valid conceptual model of pre-school teachers‘ competence including a multidimensional facet structure (research on beliefs will and has to come)
2) Reliable and valid standardized paper-pencil-assessment – covers characteristics underlying performance in real-world situations (initial validation but final one still has to come)
3) Data-based confirmation of the competence structure and qualitatively different levels (development during training and transition into pre-school will follow)
4) Identification of OTL predicting p-s teacher competence: math pedagogy classes specifically targeting the informal setting of pre-school and development of children‘s math ability (systematic classes less relevant but lack of OTL!)

Project-related references

Dunekacke, S. (in press). Erfassung mathematikdidaktischer Kompetenz von angehenden Erzieher/-innen: Theoretische Überlegungen und methodisches Vorgehen. *Tagungsband: Berliner und Brandenburger Beiträge zur Bildungsforschung*.

Dunekacke, S., Jenßen, L. & Blömeke, S. (in press). Validierung eines Leistungstests zur Erfassung mathematikdidaktischer Kompetenz angehender frühpädagogischer Fachkräfte durch die videogestützte Erhebung von Performanz. *Zeitschrift für Pädagogik – Beiheft 2015: Kompetenzen von Studierenden.*

Jenßen, L., Dunekacke, S. & Blömeke, S. (in press). Qualitätssicherung in der Kompetenzforschung: Empfehlungen für den Nachweis von Validität in Testentwicklung und Veröffentlichungspraxis. *Zeitschrift für Pädagogik – Beiheft 2015: Kompetenzen von Studierenden.*

Blömeke, S., Dunekacke, S., Jenßen, L., Koinzer, T., Baack, W., Grassmann, M., Tengler, M. & Wedekind, H. (in preparation). Effects of opportunities to learn mathematics on pre-school teachers' mathematics pedagogical content knowledge. In preparation for *Peabody Journal of Education. Special Issue „Modeling and Measuring Competencies in Higher Education".*

Silke Grafe, University of Wuerzburg, Germany

Andreas Breiter, University of Bremen, Germany

Modeling and Measuring Pedagogical Media Competencies of Pre-Service Teachers (M^{3}K)

Paper presented at 2014 AERA Annual Meeting as part of Session: *“Theoretical and Methodological Tasks and Challenges of Modeling and Measuring Competencies in Higher Education – Current State and Future Perspectives on Competence Assessment”*

April 2014

Objectives

This paper describes the deductive theory-based development of the areas of pedagogical media competencies and their differentiation into different levels. Furthermore, the approach to achieving content validity through semi-structured expert interviews and the results of the qualitative data content analysis are explained.

Theoretical framework

Drawing on prior research and concepts (cf. ISTE, 2008; Mathematica Policy Research, 2000; Mishra & Koehler, 2006, Schmidt et al. 2009) the project „Modeling and Measuring Pedagogical Media Competencies of Pre-Service Teachers“ started with the theoretical description of a model of pedagogical media competencies of pre-service teachers. In this theoretical model different aspects of competencies are described which refer to three areas:

- Media use for teaching and learning (media didactics): This area comprises the ability to analyze and assess given media education activities with regard to teaching and learning, and to analyze, prepare, give and assess exemplary lessons with the use of media.
- Teaching about media (media literacy education): This area focuses on the analysis, preparation, implementation and assessment of lessons in which the role of media in society is reflected to promote an appropriate, self-determined and creative use of media in a socially responsible way.
- Technology planning within school development: The third area of competencies is about the ability to shape school development processes, for example by understanding leadership, infrastructural, legal or organizational conditions for embedding educational media in schools (cf. Kozma 2003, Owston 2007).

The three areas of competencies are based on an understanding of the term “competencies” as learnable dispositions of achievement which comprise cognitive dimensions as well as attitudinal aspects and are directed to specific requirements and their accomplishment. In our case, this means the scientific foundation which is needed in order to cope with corresponding situations in the teaching profession and which should be acquired during university studies by future teachers.

Determination of aspects of competencies

The areas of competencies can be differentiated or specified with regard to the target group of pre-service teachers (cf. Tulodziecki 2006). They are supposed to acquire the scientific foundation for performing teaching, education and school development tasks:

- Understanding and assessing individual or social conditions for media education activities: e.g. in media didactics the ability to estimate how children's use of media outside school can affect their learning in school.
- Describing and evaluating theoretical approaches to media education activities: e.g. in media didactics, the ability to present empirical results on teaching and learning adequately using media.
- Analyzing and evaluating examples for media education activities: e.g. in media didactics, the ability to analyze exemplary lessons using media with regard to objectives or learning conditions.
- Developing personal examples of media education activities: e.g. the ability to make a theory-based assessment of media education activities with regard to a planned lesson.
- Testing and evaluating examples of media education activities: e.g. the ability to test and systematically evaluate one´s own lesson plans in classroom settings.

Methods and data sources

For the development and later the validation of the theoretical model of competencies, interviews with 10 German and 4 US experts in educational technology, media literacy education and technology planning within school development were conducted (Denzin & Lincoln 1994; Berg 2009). They were interviewed using semi-structured questions in conjunction with the critical incident technique, originally introduced by Flanagan (1954). The method was used to identify prototypical tasks and requirements for students on different levels including their learning context. At the same time, "critical incidents" were used to identify problem-solving strategies of experts in the respective fields. This was done along the dimensions, which were previously identified in the framework model. With the help of this methodology the attempt was made to reveal how experts act in specific contexts and which skills, knowledge, strategies, and beliefs they use to cope effectively with complex requirements for teaching and learning.

All interviews were recorded and transcribed. Based on qualitative methods of content analysis (Mayring 2000, Krippendorf 2008), the relevant aspects of media pedagogical competencies were extracted and paraphrased. The next step emphasized the link between the identified elements of the paraphrased text to the competencies dimensions previously identified deductively from literature research. This included a revision of phrases and categories. Based on these steps, a draft structured model of competencies is developed which describes the core elements students should achieve during their studies. This was again revised with the help of expert interviews in order to produce a comprehensive structural model of media pedagogical competencies.

Results and conclusions

We could identify the following core aspects, which were mentioned as significant competencies on different levels. For illustration, we present examples from each sub category. This exemplifies the developed model of competencies.

Level	Description	Examples
1	Understand and assess critical success factors for media pedagogical practices	From Media Didactics: Based on theoretical concepts and empirical results, students are able to describe the relevance of after-school (home) media use for teaching and learning with digital media.
2	Characterize and assess theoretical concepts for media pedagogical practices	From Media literacy education: Students are able to present approaches to media literacy education including relevant empirical evidence.
3	Analyze and assess examples of media pedagogical practices	From School development: Students are able to assess good practices for school technology planning and its relevant components from empirical, normative and implementation perspectives
4	Develop suggestions for media pedagogical practices based on theoretical concepts	From Media didactics: Students are able to design a lesson plan or a learning environment integrating digital media.
5	Test and systematically evaluate examples for media pedagogical practices based on theoretical concepts	From Media literacy education: Students are able to test and reflect a theory-based lesson plan integrating digital media in real or simulated classroom settings

It can be summarized that content validity through semi-structured expert interviews could be achieved and therefore, neither an addition nor a change was required. Based on this model, a range of items are currently being developed, tested, evaluated and re-tested. This is still work-in-progress.

Scientific or scholarly significance of the study of work

There is no empirical research concerning the necessary pedagogical media competencies of prospective teachers and school leaders, particularly in the field of school development. In this research project, we developed a first conceptual model on how to define, and how to measure competencies in this area. Our results will help to further improve teacher education programs as well as in-service trainings.

References

Berg, B. L. (2009). *Qualitative research methods for the social sciences* (7th ed.). Boston: Allyn & Bacon.

Denzin, N. K. & Lincoln, Y. S. (1994). *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage.

Flanagan, J. C. (1954). The Critical Incident Technique. *Psychological Bulletin*, Vol.54, No.4, July.

ISTE (International Society for Technology in Education) (2008): The ISTE NETS and Performance Indicators for Teachers (NETS-T). Available: http://www.iste.org/docs/pdfs/20-14_ISTE_Standards-T_PDF.pdf (March 13th, 2014).

Kozma, R. B. (2003). Technology, Innovation, and Educational Change. A Global Perspective. A Report of the Second Information Technology in Education Study Module 2. Eugene, OR: International Society for Technology in Education (ISTE).

Krippendorff, K. (2008). *Content analysis : an introduction to its methodology* (2. ed.). Thousand Oaks, CA: SAGE.

Mathematica Policy Research (2000). *Evaluating the technology proficiency of teacher preparation programs´graduates: assessment instruments and design issues. Preparing Tomorrow´s Teachers to Use Technology*. Final Report. Washington: US Department of Education.

Mayring, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research*, 1(2). Available: http://nbn-resolving.de/urn:nbn:de:0114-fqs0 (March 13th, 2014).

Mishra, P. & Koehler, M. J. (2006): Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record*, 108(6), 1017-1054.

Owston, R. (2007). Contextual factors that sustain innovative pedagogical practice using technology: an international study. *Journal of Educational Change*, 8(1), 61-77.

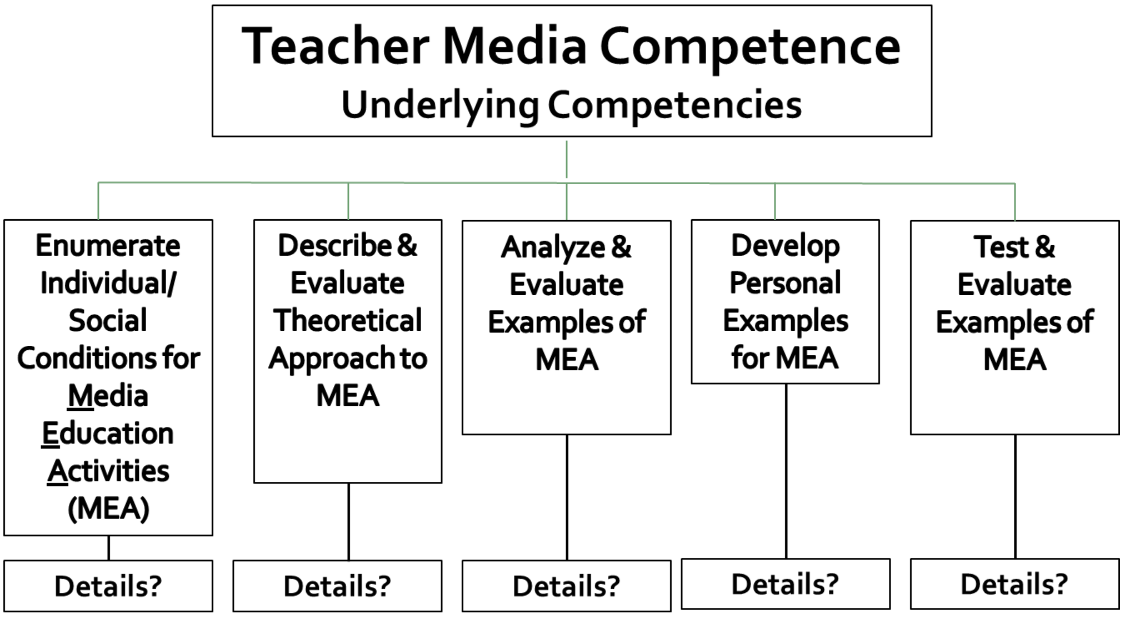
Schmidt, D. A. et al. (2009): Technological Pedagogical Content Knowledge (TPACK): The Development and Validation of an Assessment Instrument for Preservice Teachers. J*ournal of Research on Technology in Education*, 42(2), 123-149.

Tulodziecki, G. (2006): Zur Entwicklung lehr-lerntheoretisch basierter Kompetenzen in der Lehrerbildung. *Beiträge zur Kompetenzorientierung in der Lehrerbildung*. Paderborn: Schöningh, 137-154.

Rich Shavelson, SK Partners & Stanford University, USA

Comment on - Modeling and Measuring Pedagogical Media Competencies of Pre-Service Teachers (M^{3}K)

Is My Representation Of Your Conceptual Framework Accurate?


Teacher Media Competence
Underlying Competencies
Enumerate Individual/ Social Conditions for Media Education Activities (MEA)
Describe & Evaluate Theoretical Approach to MEA
Analyze & Evaluate Examples of MEA
Develop Personal Examples for MEA
Test & Evaluate Examples of MEA
Details?
Details?
Details?
Details?
Details?

Questions for Research Team

- □ How did data from experts inform the schema above?
  - ◘ Did experts agree?
  - ◘ Isn't it interesting that their data not change framework from what originally was expected?
  - ◘ I wonder just what experts said.
- □ How do levels and entries in your table map onto this schematic?
- □ While did you say there is no available conceptual frameworks when you cite such frameworks (e.g., Mishra & Koehler (2006)) provide an interesting example?

Elena Bender, Niclas Schaper, Melanie Margaritis, Laura Ohrndorf and Sigrid Schubert, University of Paderborn, Germany

Modeling Competences of Teaching Computer Science in German Schools at High School Level - Theoretical Framework, Curriculum Analysis and Critical Incident Based Expert Interviews (KUI)

Abstract

This article aims to outline the first results of the development of a competence model for teaching computer science at secondary school level. Therefore, three main methodological steps are conducted and described. To put the competence model on a strong theoretical basis a variety of theoretical and normative oriented documents with regard to teacher education is analyzed in a first step. Second, a broad curriculum analysis is undertaken to validate the theoretical derived categories and to refine them. The third step implies expert interviews based on the critical incident technique with the purpose of further differentiation of the competence categories and deriving specific descriptions of relevant competences and competence relevant facets. Concrete formulations regarding teachers´ beliefs within the subject computer science are suggested. All methodological steps are analyzed by a qualitative content analysis. Finally, the results are discussed.

Keywords: competence model; teacher education; computer science; teachers´ beliefs

Introduction

Research on teachers´ competences seeks to determine and describe what professional teachers have to learn and how they are able to adapt to their changing environment. Mainstream approaches on this topic comprehend teaching competence on a wide theoretical basis integrating cognitive, affective, and motivational factors (Kunter & Pohlmann, 2009).

If you compare computer science teachers to teachers of other subjects they are in an inconvenient situation. They are a minor group within the teaching community and many schools are provided with only one computer science teacher. Furthermore they have to cope with rapid changes in technologies (Diethelm, Hildebrandt & Krekeler, 2009). Hubwieser, Mühling and Brinda (2010) report in their study that computer science teachers in the Bavarian region are not satisfied with their own planning and success in teaching. The study has shown that despite existing challenging teaching concepts the implementation in classroom seems to be difficult. The situation depicted

shows the need for a systematic derivation of computer science teachers´ competences. The article aims to describe the way getting from job requirements to concrete competence formulations. Research questions are: How can relevant competence facets of computer science teachers be derived? How can curricular analyses and expert interviews be used to specify the theoretically considered competence areas? And how can competences be concretely formulated? The research is embedded in the project KUI ("Competences for Teaching Computer Science"; Schaper et al., 2013) funded by the German Federal Ministry of Education and Research (BMBF).

Approach of Deriving Competence Categories for a Competence Model

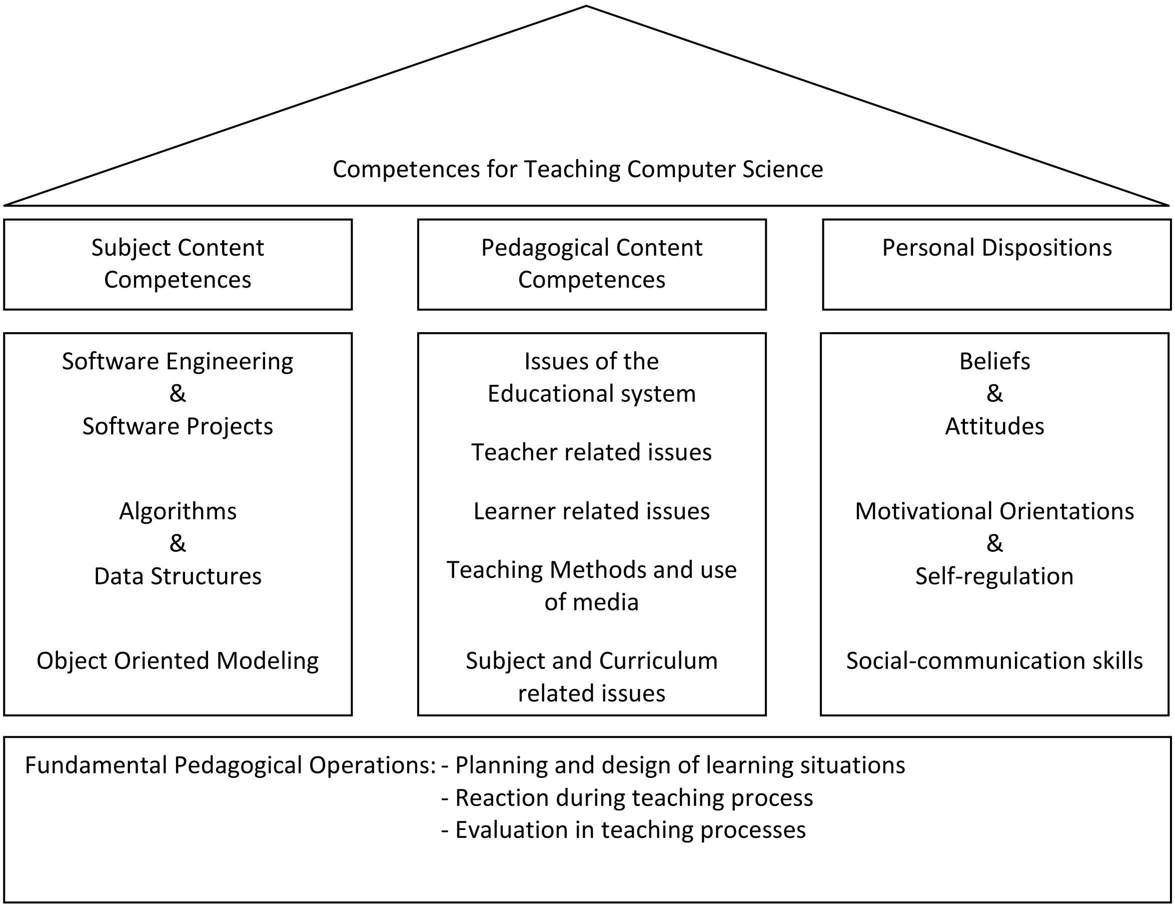
Several ways for the development of a competence model exist in educational research. Especially approaches using systematic competence modeling strategies enhance measurement as they allow a systematic linkage between theoretical constructs and results of empirical assessments (Klieme, Hartig & Rauch, 2008). As the adequate way of modeling is depending on the context of analysis we had to develop a suitable approach for teaching computer science. In our case a mixture of theoretical and normative oriented approaches as well as empirical methods seems to be adequate. Therefore three main methodological steps are conducted (Schaper, 2009). All steps aim at the refinement of the relevant competence facets and each step fulfils a different purpose building the basis for the next one (shown in figure 1).

Figure 1. *Methodological steps for the development of the competence model*.

Methodological Step	**Purpose**
Theoretical and normative oriented development of competence categories	Develop a theoretical Framework (i.e. Weinert, 2001, Shulman,1986/87, Kunter et al., 2013, Blömeke et al., 2008, ACM/IEEE, 2012)
Application to curricula (a. universities, b. schools)	First step of specification - by document analysis
Conduction of expert interviews	Second step of specification and refinement – by experts

First, several theoretical and normative oriented documents are analyzed in order to support the competence model with a strong theoretical basis. Our framework is based on the competence notion of Weinert (2001) leading to a competence model being divided into three dimensions: subject content competences, competences on pedagogical content knowledge and personal dispositions (shown in figure 2). The chosen and analyzed documents for the differentiation and specification of these dimensions are referring to three areas. First, existing models and standards of teacher education presented in large scale studies and expert papers (i.e. Terhart, 2002; Oser, 2002) are taken into account as well as normative oriented documents, i.e. developed by the German minister conference on educational and cultural affairs ("Kultusministerkonferenz; KMK"). These standards are analyzed with regard to their relevance for teaching computer science and how they have to be adapted and specified for this domain. Second, recent empirical studies from related fields of research like mathematics (i.e. Kunter et al., 2013; Blömeke, Kaiser & Lehmann, 2008) and the natural sciences (i.e. Riese, 2009, Riese & Reinhold, 2008) serve as additional relevant references. As a third source documents from the field of computer science are taken into account, particularly developed by the Joint Taskforce on Computer Science Curricula ACM and IEEE (2012) or the standards of computer science for lower secondary schools by the German Informatics Society (GI, 2008).

Figure 2. Structural competence model for teaching computer science.


Competences for Teaching Computer Science
Subject Content Competences
Pedagogical Content Competences
Personal Dispositions
Software Engineering & Software Projects
Algorithms & Data Structures
Object Oriented Modeling
Issues of the Educational system
Teacher related issues
Learner related issues
Teaching Methods and use of media
Subject and Curriculum related issues
Beliefs & Attitudes
Motivational Orientations & Self-regulation
Social-communication skills
Fundamental Pedagogical Operations: - Planning and design of learning situations
- Reaction during teaching process
- Evaluation in teaching processes

The next two methodological steps for the development of the competence model are a curricular analysis and an expert interview study explained in the following chapters.

Application of Competence Categories to Curricula

In the first empirical step the theoretically derived competence categories are deductively applied to curricula in computer science education. All available curricula (43) for computer science teacher education from German universities and school curricula of six federal states were collected. This analysis aims at investigating which theoretically derived categories are represented by the curricula. For the text analysis of the curricula the structural qualitative content analysis technique by Mayring (2010) was used. Text parts are systematically extracted and allocated to the existing categories. The coders agreed on a coding manual. To take care of reliability issues 20 % of the curricula were analyzed by two raters. The ratio of agreements between the coders relative to the sample size as a basic coefficient (von Eye, 2006) range from 70 % to more than 80 % according to the different competence dimensions. Regarding subject content competences the curriculum of the ACM/IEEE (2012) is used as a category system. It splits up the three main categories shown in figure 2 into 18 subcategories. Nine of them are covered in more than 75 % of the curricula (for example algorithms and complexity, programming languages and software engineering) and lead to a balanced representation of our main categories. For the refinement of the pedagogical content competences we identified 15 subcategories on a theoretical basis belonging to the five content-oriented categories shown in figure 2. Furthermore three process-oriented categories called “Fundamental Pedagogical Operations (FPO)” are identified. The FPO contain the phases "Planning and design of learning situations" (FPO1), "Reacting on student's demands during teaching processes" (FPO2) and "Evaluation of teaching processes" (FPO3) and build a fundamental basis for all competence areas in the model. Results show a clear emphasis of curricular contents. The four content-oriented subcategories are represented in more than 75 % of the curricula and they all refer to the first category “Issues of the educational system” (learning content, curricula and standards, science, school subject). As expected, the planning aspect of FPO1 had the highest coding frequency with 71%, followed by FPO3 (51%) and FPO2 (21%) (Hubwieser et al., 2013). With reference to personal dispositions the overall dimensions (shown in figure 2) contain several theoretically derived subcategories, three in the area of social-communication skills (i.e. aspects of empathy or cooperation), seven in the area of motivational orientations (i.e. teaching efficacy) and seven in the beliefs area (i.e. beliefs about the subject computer science). 94 % of the curricula include elements or contents regarding social-communication skills. Regarding beliefs and attitudes, 63 % of the curricula contain relevant aspects. 55 % also consider motivational oriented aspects. We also figured out that the competence categories concerning personal dispositions are formulated on a rather high degree of abstraction. Research and theoretical background focus mainly on teachers´ motivational orientations and beliefs, curricula in contrast to that underline especially social-communication aspects. Summing up, the curricula do not address competences the literature-based model could not cover. They rather have deficits concerning the consideration of personal dispositions compared to the theoretical model so that relevant aspects of teachers´ beliefs and motivational orientations are not given enough importance in the curricula. As those categories prove to be relevant for professional proficiency according to diverse empirical studies curricula should consider the implementation of those aspects with more scrutiny. To overcome this discrepancy, competence descriptions are suggested using an expert interview study, described in the next chapter.

Expert Interview Study

In this second empirical step the competence model is refined and specified by expert interviews based on the critical incident technique. Modifying the method, we developed challenging situations the experts are confronted with. The leading questions are: How does competent behavior in these situations look like? Which personal dispositions do experts assume as useful and conducive? The study is combining scenarios aiming on the one hand at the exploration of pedagogical content competences and on the other hand at the elaboration of personal dispositions embedded into a concrete computer science teaching context. The interviews are conducted with seventeen experts, nine of them are experts from computer science teacher education and eight are experienced computer science teachers in secondary school education. They are from different federal states in Germany to get a broad overview in congruence with the curricular analysis. Every interview lasts about one hour, is recorded and systematically transcribed. The interviews are analyzed by the structural qualitative content analysis technique (Mayring, 2010). Main goal of this analysis is to determine additional aspects being relevant in computer science teacher education. As a main result of the content analysis competence formulations are suggested following three analytical steps. First, core competence statements are derived from the interview texts. Then the statements are clustered with similar passages to determine more general categories of competence descriptions. Third, competence descriptions are formulated for every cluster. With regard to personal dispositions 82 % of the experts mention relevant social-communication aspects. 76 % of the experts’ statements are considered with teachers´ beliefs and 47 % with aspects of motivational orientations. As an example for a competence description, beliefs about the subject “computer science” are outlined in table 1.

Table 1. Competence descriptions for beliefs about the subject computer science.

Beliefs about the subject computer science
Students are convinced that superordinate strategies and principles make up the subject computer science and are relevant to all sections of subject
The students are convinced that the core of computer science consists of processes that can always be traced back to relationships between information and data.

Beliefs about the subject computer science refer to the concept of epistemological beliefs referring to the nature of knowing and the process of knowing (Hofer, 2001, Hofer & Pintrich, 1997). The subject “computer science” is in practice often seen as a place for learning office applications or programming languages. As the recent curricula do not recommend how beliefs about the subject "computer science" should be, experts underlined that a complex view on the subject is necessary for students learning processes. The complex view on the subject is mainly expressed by two aspects. First, the subject should be conceived as a discipline which is guided by superordinate strategies and principles. Second, a process view on the subject is strongly conducive. By refining the whole structural competence model by formulations in the described manner it is considered to build the basis for developing competence measurement instruments.

Conclusion and Further Steps

After having identified relevant competence categories for teaching computer science they are refined and specified in two empirical steps, a broad curricular analysis and an expert interview study. This approach led to a first structural competence model or framework model for teaching computer science. The curricular analysis has shown deficits in comparison to our derived theoretical model. Especially in the investigated areas of motivational orientations and teachers´ beliefs relevant aspects are not formulated explicitly. As a result, relevant competences have to be identified and formulated more clearly. Results from the expert interviews give important hints concerning missing elements and lead to suggestions of concrete competence formulations with an example given from the area of teachers´ beliefs. The structural model builds the basis for further necessary research steps. Valid measurement instruments need to be developed to determine the prevailing characteristics among computer science student teachers. They will be empirically tested on large scale during this project KUI. Finally, concrete implications for the practical use of the developed instruments in teacher education should result from this research.

References

Blömeke, S., Kaiser, G. & Lehmann, R. (2008). Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematik-Studierender und -referendare - Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung. [Professional competence of prospective teachers. Knowledge, beliefs, and learning opportunities of German mathematic students and trainees – First results of the effectiveness of teacher education]. Münster: Waxmann.

Diethelm, I., Hildebrandt, C. & Krekeler, L. (2009). Implementation of computer science in context - a research perspective regarding teacher-training. In Koli Calling 2009, p. 97-100, Koli, 2009.

Eye, A. von (2006). An Alternative to Cohen's κ. European Psychologist, 11, 1, pp. 12–24.

Hofer, B. K. & Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. Review of Educational Research, 67, pp. 88–140.

Hofer, B. K. (2001). Personal epistemology research: Implications for learning and teaching. Journal of Educational Psychology Review, 13(4), pp. 353–383.

German Informatics Society (GI) (2008). Grundsätze und Standards für die Informatik in der Schule. Bildungsstandards für die Sekundarstufe I. [Principles and standards for computer science in school. Educational standards for secondary I]. Available: http://www.informatikstandards.de/ (September 3th, 2013).

Hubwieser, P., Berges, M., Magenheim J., Schaper, N., Bröker, K., Margaritis, M., Schubert, S. & Ohrndorf, L. (2013). Pedagogical Content Knowledge for Computer Science in German Teacher Education Curricula, The 8th Workshop in Primary and Secondary Computing Education, Aarhus, DK.

Hubwieser, P., Mühling, A. &, Brinda, T. (2010). Erste Ergebnisse einer Lehrerbefragung zum bayerischen Schulfach Informatik. [First results of a teacher survey on the Bavarian school subject of computer science]. In I. Diethelm, C. Dörge, C. Hildebrandt & C. Schulte (eds.). Didaktik der Informatik - Möglichkeiten empirischer Forschungsmethoden und Perspektiven der Fachdidaktik. 6. Workshop der GI-Fachgruppe "Didaktik der Informatik", 16. - 17. September 2010 in Oldenburg, pp. 45–55. GI, Bonn.

Klieme, E., Hartig, J. & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme & D. Leutner (Eds.), Assessment of competencies in educational contexts, pp. 3–22. Göttingen, Germany: Hogrefe & Huber.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S. & Neubrand, M. (2013). Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project. New York, NY: Springer.

Kunter, M. & Pohlmann, B. (2009). Lehrer. [Teacher]. In J. Möller & E. Wild (Eds.), Einführung in die Pädagogische Psychologie, pp. 261–282. Berlin: Springer.

Mayring, P. (2010). Qualitative Inhaltsanalyse. [Qualitative content analysis]. Beltz: Weinheim.

Oser, F. (2002). Standards in der Lehrerbildung. Entwurf einer Theorie kompetenzbezogener Professionalisierung. [Standards in teacher education. Outline of a theory of competence-based professionalization]. Journal für Lehrerinnen- und Lehrerbildung, 2 (1), pp. 7 – 19.

Riese, J. (2009). Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften. [Professional knowledge and professional competence of (prospective) physic teachers]. Dissertation. Berlin: Logos Verlag.

Riese, J. & Reinhold, P. (2008). Entwicklung und Validierung eines Instruments zur Messung professioneller Handlungskompetenz bei (angehenden) Physiklehrkräften. [Development and validation of an instrument to measure professional competence for (prospective) physic teachers]. Lehrerbildung auf dem Prüfstand, 1 (2), pp. 625–640.

Schaper, N., Magenheim, J., Schubert, S., Hubwieser, P., Bender, E., Margaritis, M., Ohrndorf, L. & Berges, M. (2013). Competences for Teaching Computer Science. In KoKoHs Working Papers, [3]. Berlin & Mainz: Humboldt University & Johannes Gutenberg University. Available: http://www.kompetenzen-im-hochschulsektor.de/Dateien/KoKoHs_WP3_Bloemeke_Zlatkin-Troitschanskaia_2013.pdf (March 17th, 2014).

Schaper, N. (2009). Aufgabenfelder und Perspektiven bei der Kompetenzmodellierung und -messung in der Lehrerbildung. [Areas of responsibility and competence perspectives in modeling and measurement]. Lehrerbildung auf dem Prüfstand, 2 (1), pp. 166-199.

Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Eds.) (2008). Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. [Common guidelines of content requirements for subject areas and subject didactics in teacher education]. Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 16.09.2010 Berlin.

Terhart, E. (2002). Standards für die Lehrerbildung. [Standards for teacher education]. Eine Expertise für die Kultusministerkonferenz. Münster: Universität Münster.

The Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) (2012). Computer Science Curricula 2013 - Strawman Draft. Stanford University.

Weinert, F. E. (2001). Concept of Competence: A conceptual clarification. In Defining and Selecting Key Competencies, D. S. Rychen and L. Salganik, (Eds.), Hogrefe and Huber, Seattle.

Fritz Oser, University Freiburg, Switzerland

Comment on- Modeling Competences of Teaching Computer Science in German Schools at High School Level - Theoretical Framework, Curriculum Analysis and Critical Incident Based Expert Interviews (KUI)

Goals of the paper

1. Identifying relevant competence categories
2. Realizing a curricular analysis
3. Proceeding an expert interview study (presenting critical incidents to the expert)
4. Developing a first frame work model or a first competence model

Importance of the study 1

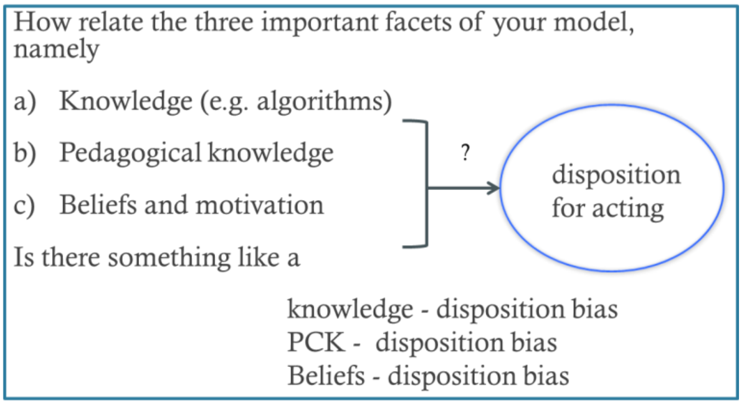
- ICT is the fifth element of educational basics (math, MINT, language, foreign language) (see Baumert, 2001)
- Most of contemporary studies show that there is no systematic training of teaching ICT (and in addition no real time window within the instruction plans)
- The new curriculum that presents the intended student competences are mostly trivial (f. i. the student should know how to handle PC technics) (see Plan 21 of harmonization of the school system)
- ICT is necessary for each profession, for each human interaction, for each systematic analysis etc.

Importance of the study 2

There is a huge research movement with respect to the issues of ICT in education

- ICILS, an international large scale comparison study from the IEA, involving teacher and learner relatedness, societal necessities of computer use, moral aspects etc.(Schulz, 2014).
- Studies on the demystification of computer use with respect to political apathy, computer loneliness, writing deficiencies, intellectual dependency, obesity and overweight etc. (Appel & Schreiner, 2014; Hattie, 2009 etc.)

Question 1



How relate the three important facets of your model, namely

a) Knowledge (e.g. algorithms)

b) Pedagogical knowledge

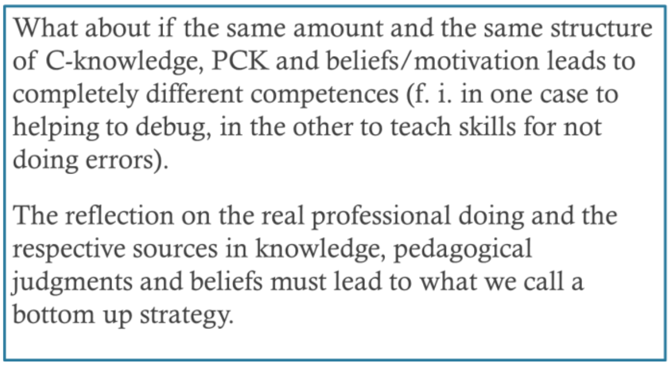
c) Beliefs and motivation

?
disposition for acting

Is there something like a

knowledge - disposition bias
PCK - disposition bias
Beliefs - disposition bias

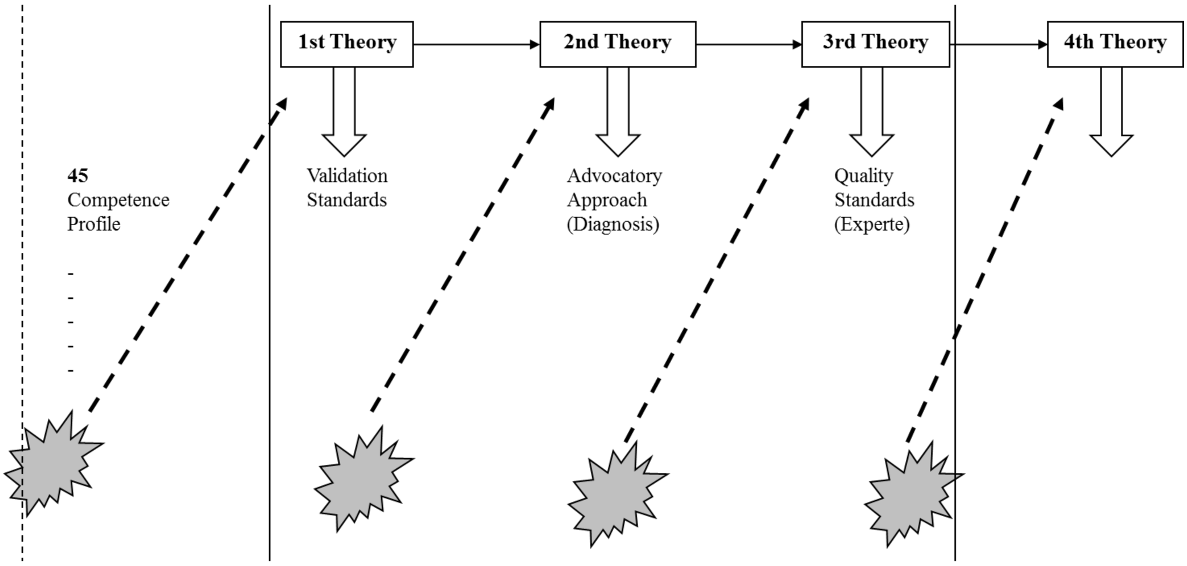
Question 1 (cont)



What about if the same amount and the same structure of C-knowledge, PCK and beliefs/motivation leads to completely different competences (f. i. in one case to helping to debug, in the other to teach skills for not doing errors).

The reflection on the real professional doing and the respective sources in knowledge, pedagogical judgments and beliefs must lead to what we call a bottom up strategy.

Theoretical element: Expert studies



Question 2

It seems to me that we need three types of modeling:

(1) A domain specific model of professionally necessary actions (collection of necessary actions)

(2) A competence specific model (going beyond situation specificity, and clustering sub-actions)

(3) A measuring model (with intensive specific validity check)

Question 3

- How can we distil the specificities of digital learning from e.g. mathematical learning (e.g. software has nearly always a service function and has less epistemic value than mathematics)
- These specificities must have a convincing face
- The concept of "Signature pedagogy" could help to develop it (It e.g. typical that most students do not learn from skill oriented mistakes; they try and try until they overcome the bug often without consciousness

Application device

- Schools and teacher education have not yet a clear concept for teaching and learning computer technics and computer sciences as a teaching matter
- It should become evident that competence orientation in this ICT field means also structuring and directing the development of such competences, and thus lounging a program with precise and highly controllable targets
- To make modeling computer sconce as pedagogical content knowledge (PCK) oriented academic clear structured topic is a necessary claim.

A final message to the research group

- Since we think that your work is important and will lead to a new specific net of competences
- Please tell us more:
  - about the critical incidents you are using
  - about the specific competence formulation and the respective net of them
  - about what you did see looking into the teacher training situation
  - about your ethnographic experiences
  - and especially about your critical stance towards what is happen now

We love the critical point at the end of your paper. This is the right direct.

Sigrid Blömeke, Humboldt University Berlin, Germany

Concluding Commentary

AERA symposium "Assessment of competencies in higher education" (Philadelphia, April 4, 2014)

Concluding commentary on the KoKoHs session at AERA 2014 and the accomplishments of the funding program "Modeling and Measuring competencies in higher education"

Introduction

It seems to be a fair evaluation to sum up that the assessment of competencies in higher education has come a long way but that a lot of work is still ahead. In the following, the overall discussion of this AERA symposium on the assessment of higher-education competencies and the conclusions based on its results are, for analytical reasons, distinguished into a theoretical and a methodological section although this distinction does not mean that both perspectives are independent from each other. The summary will be linked to Li Cao's introduction to this Working Paper on the one hand and to a paper that deals with the assessment of competencies in higher education in more detail on the other hand (Blömeke, Gustafsson & Shavelson, in press).

Theoretical perspectives on the assessment of competencies

Blömeke, Gustafsson and Shavelson (in press) identify two extreme approaches to the assessment of competencies. One is focusing on assessing performance holistically in real-world situations. This approach is more prevalent in English-speaking countries. In this extreme, a risk exists to forget about the resources necessary to be able to perform successfully in these situations. Another approach is focusing on these resources, namely the different latent dispositions, in particular the cognitive ones, underlying performance. This analytical approach is more prevalent in Germany, the context where most of the studies presented today took place. Like the first approach, also the second one is a highly valuable approach which has been convincingly demonstrated today. However, it is in high risk to forget about the need to integrate the dispositions and to validate them against success in real world. The "ecological validity" (Ciao, in this Working Paper) is to be questioned then.

So, both approaches have their obvious shortcomings. Blömeke, Gustafsson and Shavelson (in press) propose therefore to overcome such an unfruitful dichotomy by "viewing competence as a continuum" (p. 1) and including a process dimension which points out that situated skills to perceive, interpret and make decisions are necessary to be able to apply personal resources to different situations and thus to transform dispositions into performance. Each person can then be characterized by a profile of how the different dispositions are linked to each other (Oser, 2013).

Against such a theoretical framework, different research needs can be identified:

a) basic research on the nature of the different constructs and how precisely they are connected – we cannot expect immediate practical outcomes from such type of research because it is tedious work but some of the presentations today addressed this issue already (see, e.g., Blömeke et al., in this Working Paper)
b) more applied research on the generalizability of competence across different higher-education programs or professional situations (Shavelson & Webb, 1991) with the objective to learn more about the potential domain-specifity of the transformation process; some of the presentations today addressed this issue (see, e.g., Schmidt, Förster & Zlatkin-Troitschanskaia, in this Working Paper)

Methodological perspectives on the assessment of competencies

The research needs pointed out include complex methodological demands because they require different assessment approaches, more analytical multiple-choice based ones but also holistic performance-based approaches – and in particular new, innovative approaches in-between that on the one hand provide the chance to come close to real-world performance but on the other hand still allow for standardized assessments. Only then, it is possible to cover the whole process from disposition to performance, to make inferences above cases and to model the person-situation interaction. The sampling framework will be of utmost importance in such studies, given that the selection of situations has to be validated very carefully with respect to their frequency and centrality (Kane, 1992).

In analyzing the data, classical theory has a lot to say – and may be more than the studies presented in this symposium today acknowledged. Generalizability theory (Brennan, 2001) or latent-state trait theory (Eid & Diener, 1999) are two approaches particularly valuable and recommended in this context. Approaches based on the item-response theory, of course, have their say, too. However, instead of routinely applying them, they need to be related to theoretical and methodological considerations in a better way than presented today. In contrast, to classical-theory based approaches which essentially try to identify sources of error, IRT-based approaches are more suited for scaling purposes and for examining the nature of constructs. For example, the issue of uni- versus multidimensionality can then be resolved by focusing on “essential” unidimensionality (Gustafsson & Åberg-Bengtsson, 2010).

Thus, also in this respect different research needs exist:

a) We are in urgent need of methodological research on the benefits and limits of different approaches based on classical theory and item-response theory; some of the presentations today addressed this issue (see, e.g., Gräfe & Frey, in this Working Paper).

b) In addition, more applied research is needed that includes careful, theory-driven development of instruments and their validation that are able to assess the full process of dispositions, skills and performance and how they are transformed into each other (see, e.g., Zahner & Wolf, in this Working Paper).

These research needs can only be addressed if different communities are brought together, substantive experts as well as methodological experts. The studies presented today demonstrated the fruitfulness of such interdisciplinarity in a convincing way. Hopefully, the research can be continued so that not only the field of assessment of competencies in higher education can be advanced but also assessment approaches in general.

References

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. (in press). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie.*

Brennan, R.L. (2001). *Generalizability theory.* NY: Springer.

Eid, M. & Diener, E. (1999). Intraindividual Variability in Affect: Reliability, Validity, and Personality Correlates. *Journal of Personality and Social Psychology, 76*, 662-676.

Gustafsson, J-E. & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. I Embretson, S. E. (Ed.). *Measuring Psychological Constructs: Advances in Model-Based Approaches.* Washington: American Psychological Association.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Oser, F. (2013). “I know how to do it, but I can’t do it”: Modeling competence profiles for future teachers and trainers. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges* (pp. 45-60). Rotterdam, The Netherlands: Sense Publishers.

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

Summary

Alicia C. Alonzo, Michigan State University, USA

Closing Remarks -

KoKoHs - Theoretical and Methodological Tasks and Challenges of Modeling and Measuring Competencies in Higher Education: Current State and Future Perspectives on Competence Assessment

A Significant Accomplishment

- Systematic efforts to assess higher education competencies across a wide range of domains
- Careful attention to the development of competency models
- Innovative efforts to elicit hard-to-assess competencies
- Incredible amount of effort devoted to validation work

A Next Step: Pooling Expertise

- While taking into account the domain-specific nature of competencies, what can be learned across efforts?
    - Methods for articulating theoretical models of competencies
    - Actual theoretical models of competencies
    - Tasks
    - Methods for validation
    - Factors affecting performance
    - Psychometric and statistical methods

A Next Step: Sharing Results with a Global Audience

- To what extent can the instruments developed as part of the KoKoHs project be used in other countries?
    - What contextual features complicate the process of adapting (the KoKoHs-developed) instruments for use in other countries?
- What “lessons learned” from these projects could be used by others?
    - Methodological audiences
    - Domain-specific audiences
- How can the capacity developed through these projects be “disseminated”?

A Next Step: The (More) Fun Part!

- How can the instruments that have been developed be used to answer important questions about learning in higher education?
    - What conditions support (domain-specific) learning in higher education?
    - What instructional models support learning in higher education?

Previously published:

KoKoHs Working Papers, 1

Blömeke, S. & Zlatkin-Troitschanskaia, O. (2013). Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor: Ziele, theoretischer Rahmen, Design und Herausforderungen des BMBF-Forschungsprogramms KoKoHs [Modeling and Measuring Competencies in Higher Education: Aims, theoretical framework, design, and challenges of the BMB-fFunded research program KoKoHs] (KoKoHs Working Papers, 1). Berlin & Mainz: Humboldt-Universität & Johannes Gutenberg-Universität.

KoKoHs Working Papers, 2

Blömeke, S. (2013). Validierung als Aufgabe im Forschungsprogramm “Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor” [The task of validation in the research program „Modeling and Measuring Competencies in Higher Education“] (KoKoHs Working Papers, 2). Berlin & Mainz: Humboldt University & Johannes Gutenberg-University.

KoKoHs Working Papers, 3

Blömeke, S. & Zlatkin-Troitschanskaia, O. (Eds.) (2013). The German funding initiative “Modeling and Measuring Competencies in Higher Education”: 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students (KoKoHs Working Papers, 3). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

KoKoHs Working Papers, 4

Berger, S., Hammer, S., Hartmann, S., Joachim, C. & Lösch, T. (2013). Causal Inference in Educational Research. Approaches, Assumptions and Limitations. (KoKoHs Working Papers, 4). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

KoKoHs Working Papers, 5

Toepper, M., Zlatkin-Troitschanskaia, O., Kuhn, C., Schmidt, S. & Brückner, S. (2014). Advancement of Young Researchers in the Field of Academic Competency Assessment – Report from the International Colloquium for Young Researchers from November 14-16, 2013 in Mainz (KoKoHs Working Papers, 5). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.