Blömeke, Sigrid; Gustafsson, Jan-Eric; Shavelson, Richard J. (2015). Beyond dichotomies: Competence viewed as a continuum. Zeitschrift für Psychologie, 223, 3-13.
DOI: http://dx.doi.org/10.1027/2151-2604/a000194

Post-print version for Green Open Access

Beyond dichotomies: Competence viewed as a continuum

Sigrid Blömeke,[1] Jan-Eric Gustafsson,[2] & Richard Shavelson[3]

[1] University of Oslo, Norway & Humboldt University of Berlin, Germany

[2]University of Gothenburg, Sweden & University of Oslo, Norway

[3]SK Partners, LLC & Stanford University, USA

Introduction

In our Call for Papers for this special issue of *Zeitschrift für Psychologie (ZfP)* we blithely said that the "assessment of competence development during the course of higher education presents a substantive and methodological challenge. The challenge is to define, and model competence—as the latent cognitive and affective-motivational underpinning of domain-specific performance in varying situations—in a reliable and valid way" (Blömeke, Gustafsson, & Shavelson, 2013, p. 202). We now say "blithely" because at the time we thought that "what was meant by the term 'competencies' seemed … to be an easier task [than measuring competencies]". We should have known better. For it is well known that definitions are important and contested. Moreover, once defined, constraints are placed on the measurement of competence—what constitutes a task eliciting competence and what doesn't, what is an allowable measurement procedure and what isn't, what is a reasonable approach to scaling and what isn't etc.? This reality was brought home not only in the diversity of definitions and measurement approaches represented in this special issue of *ZfP* but also in the debates and deliberations in the literature—and among the editors!

The Call actually set us up for controversy with a phrase so commonly seen in the measurement literature—"Competencies are conceptualized as complex ability constructs that are context-specific, ... and closely related to real life" (Koeppen, Hartig, Klieme & Leutner, 2008, p. 61)—that we editors left it unchallenged until we tried to unpack it and rephrased it as "the latent cognitive and affective-motivational underpinning of domain-specific performance in varying situations." To be sure, cognition and affect-motivation are latent traits (i.e., human constructions of unobserved processes); they cannot be directly observed but have to be inferred from observable behavior. However, this definition only provides a starting point to address the *conceptual* and *methodological* challenges involved in assessing competencies acquired in higher education.

Conceptually, in a first interpretation, the "complex ability" part of the definition is stressed and competence is analytically divided into several cognitive and affective-motivational traits (or resources), each to be measured reliably and validly. The validity of an interpretation that such a measurement taps competence could then be established by, for example, testing whether the trait structure was as

hypothesized, or whether the measurement predicted performance in a "criterion situation". The correlation between competence and performance might vary across different situations but we would expect it to be positive and of substantial magnitude. Many of the papers included in this special issue fit well within this analytic tradition. Förster et al. (this issue), for example, examine the content knowledge in micro- and macro-economics acquired during higher education in Germany and Japan with a paper-and-pencil test validated for cross-country comparisons.

A second interpretation focuses on the "real life" part of the definition and thus on observed behavior in context. Competence itself, then, is assumed to involve a multitude of cognitive abilities and affect-motivation *states* that are ever changing throughout the duration of the performance. In this case, the goal is to get measures as "closely related" to criterion performance as possible. Perhaps the closest a measurement can get to criterion performance is to sample real-world tasks and observe performance on them. What is to be measured, then, is behavior in real-life situations recognizing that no two people might use the exact same competence profile to carry out the behavior. Some of the papers included in this special issue fit within this more holistic tradition. Kulgemeyer, Tomczyszyn and Schecker (this issue), for example, developed a controlled assessment situation in which a pre-service physics teacher has to explain a physics phenomenon to a high-school student. The topic, the prompts given by the high-school student and the duration are standardized to facilitate the measurement but else the situation is natural and typical for classroom demands.

Methodologically, we note that the long-standing measurement traditions certainly provide useful tools to approach technical issues in assessment of competences. But factor analysis and other classical methods were developed to solve other measurement problems than those encountered when assessing domain-specific performance. Thus, with only few exceptions (e.g., generalizability theory [GT] as developed by Cronbach et al. (1972); see also Shavelson & Webb, 1991; Brennan, 2001) much of classical test-theory (CTT) focuses on reliable assessment of *individual differences* on *single* characteristics in *norm-referenced* contexts. But assessment of competences often requires *criterion-referenced* decisions, such as whether particular levels of competence have been reached (Berry, Clark & McClure, 2011).

Furthermore, as pointed out above, in competence assessments a multitude of characteristics is to be taken into account at the same time and the profile, how these characteristics are related to each other within a person often is of strong interest. For such purposes latent trait and mixed models–out of which item-response-theory (IRT) models are the most prominent ones–seem to hold promise, in particular because they make it possible to investigate the nature of scales (Rijmen et al., 2003).

If the latent variables are, in addition, categorical (mixture models; McLachlan & Peel, 2000), a person-oriented approach to competence profiles can be explored. The intent is to capture unobserved heterogeneity of profiles in subpopulations. These approaches open therefore up a wide range of possibilities in the field of educational measurement.

A specific methodological challenge in the context of competence assessments though is that reliability requirements typically imply a large number of items which leads to selected-response assessments that

can be quickly administered and scored. However, assessment of domain-specific competence in higher education does not necessarily lend itself to such approaches because validity considerations call for tapping "real-life" performance at some point. Achieving sufficient reliability and generalizability in the assessments is challenging given the complexities of higher-education competencies.

Overview

We have purposively characterized the definition and measurement of competence by strong and opposing positions. Pragmatically, reality in both respects lies somewhere in between. At either extreme, there is a chance of forgetting either observable behavior or cognitive abilities. That is, our notion of competence includes "criterion behavior" as well as the knowledge, cognitive skills and affective-motivational dispositions that underlie that behavior. Statistically, we believe that both CTT, in particular GT and other approaches that are based on the decomposition of variance, and more recent latent trait, mixed and mixture models in the IRT tradition have a role to play in examining the quality of competence measurements.

This paper tries to tidy up "this messy construct". We do not intend to find "the" one definition and assessment-of-competence measurement. Rather by systematically sketching conceptual controversies and assessment approaches we attempt to clarify the construct and its measurement. Our discussion of "messy" challenges confronting the definition and measurement of competence begins with definitional issues. We unpack competing definitions and identify commonplaces where there seems to be a modicum of agreement. We also highlight disagreements and suggest how some might be resolved. We then provide examples of how competence is defined in several professions. Next we discuss methodological issues, focusing on how we can move beyond dichotomies by balancing and making the best use of both CTT and IRT. Finally, we conclude by tying key points and issues together.

Conceptual Framework: Definitions of Competence

The notion of competence was discussed first in the US during the 1970s (Grant et al., 1979). The discussion focused on performance on "criterion tasks" sampled from real-life situations. McClelland (1973) contrasted the "criterion-sampling" approach with testing for aptitude and intelligence. In McClelland's view, "intelligence and aptitude tests are used nearly everywhere by schools, colleges, and employers…. The games people are required to play on aptitude tests are similar to the games teachers require in the classroom…. So it is scarcely surprising that aptitude test scores are correlated highly with grades in school" (1971, p. 1). He argued that we instead should be testing for competence—successful behavior in real-life situations: "If someone wants to know who will make a good teacher, they will have to get videotapes of classrooms, as Kounin (1970) did, and find out how the behaviors of good and poor

teachers differ. To pick future businessmen, research scientists, political leaders, prospects for a happy marriage, they will have to make careful behavioral analyses of these outcomes and then find ways of sampling the adaptive behavior in advance" (p. 8).

A contrasting perspective stressed competence's dispositional, and in particular its cognitive nature; either generic competence, which is often synonymous with intelligence or information processing abilities, or domain-specific competence, often referred to as expertise. Boyatzis (1982) carried out one of the first empirical studies in this perspective. Based on top managers definitions of their competence he defined it as an "underlying characteristic of a person which results in effective and/or superior performance in a job" (p. 97). Spencer and Spencer (1993, p. 9) were more precise:

> A competency is an underlying characteristic of an individual that is causally related to criterion-referenced effective and/or superior performance in a job or situation. Underlying characteristic means the competency is a fairly deep and enduring part of a person's personality. [...] Causally related means that a competency causes or predicts behavior and performance. Criterion-referenced means that the competency actually predicts who does something well or poorly, as measured on a specific criterion or standard."

So, as we see, a variety of definitions has existed and still exists. The respective representatives mutually criticize each other fiercely for misconceiving the construct, reducing its complexity, ignoring important aspects and so on (e.g., McMullan et al., 2003). The value added by each of the perspectives is rarely acknowledged.

The dichotomy of a behavioral assessment in real-life situations versus an analytical assessment of dispositions underlying such behavior has much to do with the origins of these different models. The first approach stems from industrial/organizational psychology that has the selection of candidates best suited for a job as the main purpose in mind. Naturally, underlying dispositions are then not the focus because they are not as close as observed performance in context. Rather, predicting future job-performance by sampling typical job tasks and assessing how well a candidate does represents a reliable and valid approach to identify job-person fit (Arthur et al., 2003). Many large employers carry out such assessments as part of their recruitment process. It is not important how a candidate has come to his or her competence. What matters is that he or she shows it in situations relevant for the job (Sparrow & Bognanno, 1993). But also in the context of professional certification and licensure, performance criteria and their assessment according to the standards of a profession are foregrounded. Which opportunities to learn a candidate had during his or her training or which traits contribute to performance is not the focus. The license is only awarded if a teacher, nurse or psychologist is able to do what is required.

In contrast to this selection approach, the second approach stems from educational research and intends to find ways to foster the development of competence. Identifying a person's characteristics (resources) underlying her or his behavior and how these best can be developed are essential in this approach. An implicit assumption is that these characteristics are amenable to external interventions (Sternberg & Grigorenko, 2003; Koeppen et al., 2008) such as opportunities to learn and systematic training so that the relationship between educational inputs and competence outcomes is foregrounded

and a frequent research topic. In the long run, the purpose is not to identify job-person fit but to identify those opportunities to learn on the individual, classroom and system level best suited to foster competence development. The German research program, “Modeling and measuring competencies in higher education”, is an example (Blömeke et al., 2013). The program responds to the increasing discussion about instructional quality in higher education and the new wave of competence-based curricula as a result of the Bologna process’ requirements.

Overcoming disagreements due to oversimplified dichotomies

The industrial/organizational *selection* and the educational *training* approaches to the definition of competence and competence assessments are in some respects distinct. In the following, we unpack the disagreements and suggest how to overcome these. However, we also see substantial commonalities in the various notions of competence—a “framework” of sorts. We highlight these commonalities first.

Agreements in the definition of competence

There is some agreement in the two contrasting perspective laid out above that “competence” (plural “competences”) is the broader term whereas “competency” (competencies) refers to the different constituents of competence. The first term describes a complex characteristic from a holistic view point whereas the latter takes an analytic stance. The constituents (or resources) may be cognitive, conative, affective or motivational. In contrast to common views of intelligence as a less malleable trait, competence and competency are regarded as learnable and can thus be improved through deliberate practice (Weinert, 2001; Epstein & Hundert, 2002; Shavelson, 2010).

Furthermore, agreement exists in both perspectives that a competence framework recognizes the importance of real-world situations typical for performance demands in a field. The definition of competence has therefore to start from an analysis of authentic job or educational situations and enumerate the tasks and the cognition, conation, affect and motivation involved. And no matter whether one follows the behavioral or the dispositional perspective—such real-world situations should be sampled in measures of competence or in measures of criteria. In both cases, the underlying competencies inferred from such a framework do not necessarily have to be in line with those inferred from a curriculum in school or university.

Beyond dichotomies: Competence as a multi-dimensional construct

If we agree that competence ultimately refers to real-world performance, either as constituent of the construct or as a validity criterion, several disagreements are resolved. It is then no longer a question whether competence is a set of cognitive abilities only or is a *combination* of cognition, conation, affect and motivation. To the degree that conation, affect and motivation are involved in that performance besides cognition, so too should the definition of competence include them for that domain. Competence thus involves complex intellectual characteristics along with affect-motivation that underlies observable performance. Evidence exists that for long-term job success, such subjective

indicators have to be taken into account (Brief & Weiss, 2001). Job satisfaction predicts productivity and performance (Judge, Thorensen, Bono, & Patton, 2001). Work engagement also predicts performance and, in addition, organizational commitment (Bakker, 2011) and health (Hakanen & Schaufeli, 2012).

This argument leads back to Snow's (1994) idea of two pathways that contribute to achievement, namely a cognitive and a commitment pathway. Thus, he included motivational-conative processes in his new concept of aptitude. Lau and Roeser (2002) confirmed this framework empirically with respect to science achievement. Whereas students' cognitive abilities were the strongest predictors, it turned out that motivational characteristics increased the predictive validity and that these were also the strongest predictors for commitment.

A priori, it is impossible to specify which specific facets enter into a definition of competence. For example, what does a competent *physicist* know, believe, and is able to do? Only from detailed observation and other information particular profiles of cognition, motivation etc. can be specified. Not only is subject-matter knowledge required to solve force and motion problems, so too are problem solving strategies, analytic reasoning, critical thinking and the like. Moreover, if competent performance involves working successfully as a team member, this competency would be included in the definition of competence. Thus, any definition of competence should entertain the possibility that competence involves complex cognitive abilities along with affective and volitional dispositions to work in particular situations.

Beyond dichotomies: Competence as a horizontal continuum

Currently, the dichotomy of disposition versus performance comes down to and gets stuck with the question of whether (a) competence *is* performance in real-world situations, more specifically, whether behavior is the focus of competence, or (b) behavior is the *criterion* against which cognition and affect-motivation are validated as measures of competence. As we will see, such a dichotomy overlooks an essential question and this is how knowledge, skills and affect are put together to arrive at performance.

The first position (a) takes a holistic view in which cognition, affect-motivation and performance are complexly linked together, changing during the course of performance (Corno et al., 2002). A competence assessment, then, involves successfully carrying out concrete tasks in real-world criterion situations; a definition of competence, then, should be based on a thorough analysis of the demands of and variations in these situations. To be sure, knowledge, skill, and affective-motivation components underlie performance but they change during the in-situation performance as the situation moves along. Cognition, affect-motivation and performance are linked as a system, cobbled together in response to task demands, somewhat differently for each person. This observation is what Oser (2013) had in mind when he pointed out that competence involves a process dimension which he calls a "competence profile"—a set of resources enacted in practice. One important research question in this context is how precisely the different resources are cobbled together, what this interplay depends on and how the resources can be built up (i.e., how should they look like, e.g. at the end of higher education).

The second position (b) restricts the term “competence” to the sum of cognitive and motivational resources. This approach assumes that the whole *is* the sum of its (weighted) parts and divides competence into multiple constituents (latent abilities, skills) needed for competent performance. Competencies, then, are used to predict behavior in criterion situations (e.g., Spencer & Spencer, 1993). From this perspective, among others, measures of both declarative and procedural “knowing” tap underlying competencies such that they are applicable to multiple “real-world” situations in which doing is the end game. If this reasoning holds, we should seek a model of competence featuring cost-efficient selected-response measures of declarative and procedural knowledge in a domain. Note that this definition of competence would also lead to a measurement model that accounted for task/response sampling, in addition to scaling scores. Since real-world behavior is the core validity criterion in this case, again a careful analysis of the demands of and the variations of these situations would be crucial. One important research question is about the relation of competence and its constituents (Sadler, 2013): Is it possible to decompose competence exhaustively as it is often done in technology and science? The decomposition reduces complexity and aids understanding—but is it the same then?

In both perspectives, the behavioral and the dispositional, the question arises as to whether and how persons who possess all of the resources belonging to a competence construct are able to integrate them, such that the underlying competence emerges in performance. This might be an empirical question but would require assessments for each competency.

Conceptually, this question leads us to point out an important gap in the current dichotomized discussion: Which *processes* connect cognition and volition-affect-motivation on the one hand and performance on the other hand? Different facets have to be integrated, perhaps to be transformed and/or restructured through practical experience. Processes such as the perception and interpretation of a specific job situation together with decision-making may mediate between disposition and performance (see Figure 1).

Thus, instead of insisting on an unproductive dichotomy view of competence, in particular knowledge *or* performance, competence should be regarded as a process, a continuum with many steps in between. Thus, we suggest that *trait* approaches recognize the necessity to measure behaviorally, and that *behavioral* approaches recognize the role of cognitive, affective and conative resources. At this time, we encourage research on competence in higher education emanating from either perspective and paying attention particularly to the steps in between. Our model may help thinking about these.

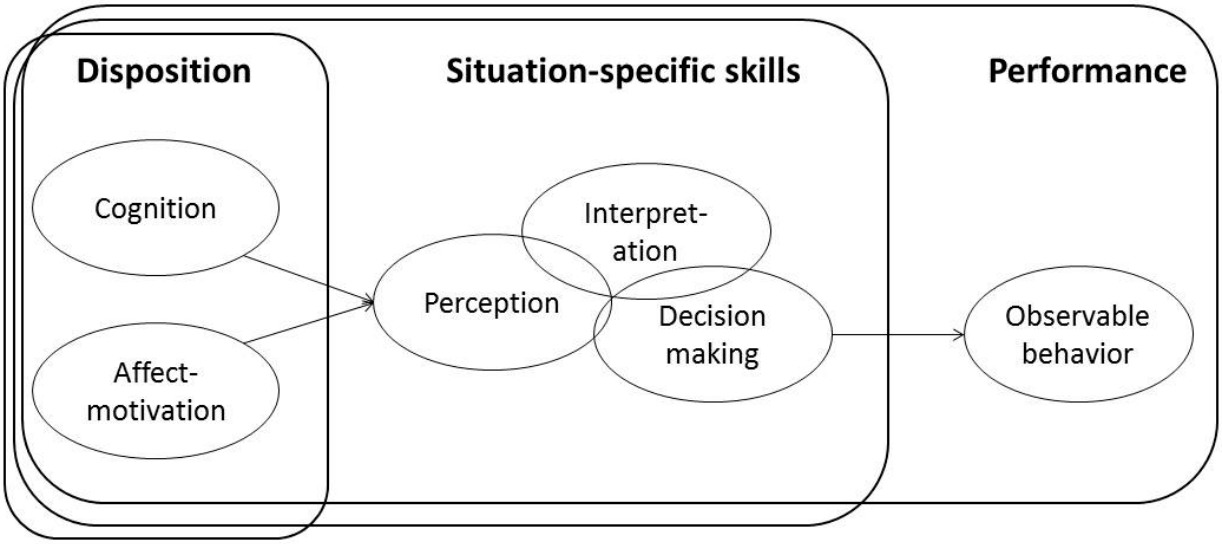

Figure 1: Modeling competence as a continuum

Competence is also a continuum in other respects

Before we consider particular fields of research on competence, it is worth noting at least briefly that competence is also a vertical continuum in terms performance levels and of (developmental) stages. More specifically, one interpretation is that competence is a continuous characteristic with higher and lower levels (more or less competent). Additionally, as competence is a multi-dimensional construct, a person's profile might include stronger estimates in one dimension and weaker ones in another. So, the definition of competence includes the notion of how much is enough to be called "competent". Furthermore, an important research question is whether the different dimensions of competence can compensate for each other (i.e., are additive by nature) or if strength on one cannot compensate for weakness on another dimension (i.e., multiplicative nature of competence dimensions; Koeppen et al., 2008). In the latter case, an interesting follow-up research question would be which minimum threshold has to be in place before someone can show a certain behavior.

Taking a longitudinal, developmental perspective on competence adds complexity. The model might be similar to figure 1 in that, firstly, some dispositions have to be in place before situation-specific skills can be acquired. Many higher education programs (e.g., teaching or medicine) are built on such an implicit assumption by delivering basic knowledge first before students undergo practical training. But it might as well be that a developmental model would look completely different in that growth or loss continuously happens on all dimensions at the same time (Cattell, 1971; Baltes, 1980). An interesting research question is whether competence changes are then best characterized by linear increase (or decrease), by differentiation processes from more general and basic expressions to more specialized one, or by qualitative changes as it is assumed in the novice-expert paradigm. In the two latter cases, developmental trajectories would imply structural changes in the nature of competence.

Particular Fields of Research on Competence

Many professions are concerned about the nature and assessment of competence. They have to train the next generations of professionals on the one hand and to award licenses or to select candidates on the other hand. Here we look specifically at medicine, teaching and vocational education to see how each deals with competence and its measurement.

In medicine, the debate about the meaning of competence has a long tradition and included from the beginning both perspectives: competence development through medical training but also selection at its end in terms of licensing. The debates resulted in Miller's (1990) widely used pyramid of clinical competence. The pyramid provides a framework for how to think about the different transformation processes that link factual knowledge and job-related behavior by distinguishing between knowledge, competence, performance and behavior. The level of each category and the relation between categories (e.g., knowledge is regarded an antecedent to competence) are assumed to be influenced by other characteristics such as beliefs, opportunities to learn, practical experiences or situational affordances. The Accreditation Council for Graduate Medical Education (http://www.acgme.org) strives to include both, the educational and the selection perspectives in their accreditation procedure for medical programs in the U.S. by requesting assessments of different dimensions of clinical competence. Epstein and Hundert (2002) summarize these as a cognitive function—knowledge to solve real-life problems; an integrative function—using biomedical and psychosocial information in clinical reasoning; a relational function—communication with patients and colleagues; and an affective/moral function—the willingness, patience, and emotional awareness to use these skills judiciously and humanely. For each category, specific assessment formats have been developed (Wass et al., 2001): traditional paper-and-pencil tests, standardized performance assessments using laboratory experiments or simulations and unstandardized performance assessments at the workplace.

Teaching is another field with extensive research on what it means to be competent. Outstanding teacher performance is regarded to involve different types of resources, in particular knowledge, skills, beliefs, values, motivation and metacognition (Shulman, 1987). The corresponding research is mostly driven by the objective of long-run improvement of teacher education. A study by Blömeke et al. (in press), for example, found that mathematics teachers' perception accuracy of classroom situations and speedy recognition of students' errors are influenced by their knowledge acquired during teacher education (see also König et al., 2014). Kersting et al. (2012) demonstrated in addition that higher perceptual ability is not only positively correlated with teacher knowledge but also with higher student achievement in mathematics. Gold, Förster and Holodynski (2013) showed that it is possible to train perception abilities with respect to classroom management through guided video analysis. Correspondingly, Stürmer, Könings and Seidel (2012) confirmed a positive effect of classes in teaching and learning on professional vision. However, selection also plays an important role in teacher education and is addressed differently. The German teacher education system, for example, requires two comprehensive exams before a license is awarded: a first one after university with typical written and oral knowledge tests and a second one on-site in schools where student teachers have to demonstrate their teaching skills.

Finally, in the field of vocational education and training (VET) competence is discussed intensely. Although many different definitions exist here as well (Biemann et al., 2004), some agreement exists with respect to core concepts (Mulder, Gulikers, Biemans & Wesselink, 2009). Competence is regarded as an integrated set of knowledge, skills, and attitudes. It is regarded as a necessary condition for task performance and for being able to function effectively in a certain situation. Shavelson (2010) presented a definition of competence in VET from the holistic behavioral perspective; included in the assessment are also probes of knowledge and skills though. The German dual VET system intended to combine dispositional and behavioral approaches by partly taking place in school—delivering traditional knowledge and completed with theoretical exams—and partly at the work place—delivering practical experience in an occupation and completed with exams in which students were supposed to master real-world challenges.

Importantly, all three fields are heavily affected by the debate about the theory-practice gap between schools or universities and work place. A major research focus is therefore on competence-based school and university education which has become increasingly popular in Western Europe. Instead of following a disciplinary curriculum, defining cognitive outcomes related to typical situations in an occupation and examining them in a performance-based way is regarded a promising way to raise the quality of the workforce (Handley, 2003). After some euphoria, implementation turned out to be more difficult and less related to higher quality than expected (Eraut, 2003).

Methodological framework: Assessing competence

Assessments developed to measure competence, by nature, have to differ from traditional knowledge tests (Benett, 1993; Birenbaum, 2007). For example, frequent or central real-world situations typical for performance demands in a domain play a crucial role either for determining ~~as~~ constituents of competence or as validity criteria. Thus, the sampling of these situations is crucial and their representativeness for the universe of tasks has to be ensured (Shavelson, 2012). Moreover, whereas reliability and (construct) validity as classical criteria of test quality remain important, the range of quality criteria has been expanded to address specific characteristics of competence assessments such as authenticity, fairness, transparency, consequences for student achievement and motivation, and cost efficiency (Messick, 1995; Kane, 2013). These requirements impose challenges for competence assessments which currently often are given too limited attention.

Challenges and issues

The analytic view of competence assessment focuses on measuring different latent traits (cognitive, conative, affective, motivational) with different instruments. Assessing the resources one-by-one has the advantage that it identifies specific preconditions for performing well in real life. The approach also has the advantage of diagnostic accuracy because what is measured within reasonable time and cost

constraints by a particular scale is a constituent of the broader competence, thereby pinpointing particular strengths and limitations. Because such measures include large numbers of observations, the approach often leads to high reliability. Nevertheless, serious validity concerns exist, most notably construct underrepresentation.

From the holistic view of competence (performance in complex, messy real-life situations), assessments have been developed to estimate real-life performance without accounting for the contribution of specific dispositional resources. Assuming the whole is greater than the sum of its parts, it is argued that assessing them one-by-one might distort the actual underlying traits needed for successful performance. The Collegiate Learning Assessment provides an example, sampling tasks from newspapers and other common sources and constructing an assessment around them to tap critical thinking, analytic reasoning, problem solving and communication (Benjamin, 2013; Shavelson, 2010). However, there are several challenges in this approach, too (Kane, 1992). The first is that there is a tradeoff between testing time and the number of independent samples of behavior that can be collected. Performance tasks are complex and take considerably more time than selected-response or short-answer tasks. Hence, only a limited sample of behavior can be collected in a given amount of time which imposes limits on generalizability. A second issue is that assessment of the complex student responses which typically are produced in performance assessments introduces considerable amounts of measurement error because it is harder to define and assess quality of responses in complex situations than with respect to clearly-defined items. Yet another issue is that different components of extended performance tasks tend to depend on one another, thereby violating the assumption of local independence which is central in most measurement models. This raises questions about how to model the item responses appropriately. In some situations solutions may be found by creating testlets (Wainer et al., 2007), but development of specialized models to deal with this issue may also be needed.

Overcoming disagreements due to oversimplified dichotomies

Thus both the analytic and the holistic approaches to assessment are afflicted by issues of validity and reliability. These issues need attention in further work on modeling and measuring competence. The issues space is not primarily a matter of dichotomy and choice between the analytical or holistic approaches. Rather the space involves *how the different approaches may be developed and combined in fruitful ways to improve the reliability and validity of competence assessments*. This involves many conceptual and empirical questions, and data rather than opinion is needed to inform future measurement methods. Below we discuss future work in three areas which we see as promising for methodological development, namely assessment formats, conceptual frameworks and dimensionality, as well as modeling techniques.

Beyond dichotomies: Tapping into a broader range of assessment formats

One gets the impression that the unproductive dichotomy of dispositions (analytic) versus performance (holistic) in assessments translates into the use of a limited range of assessment formats, with either multiple-choice items or very complex tasks dominating. It is obvious that knowledge and personality tests as well as performance assessments have important functions to fulfill in a competence assessment. The limitation to either-or should be of concern because they each only tap into parts of the construct definition.

Using combinations of approaches, we may also be able to cover the *processes* mediating the transformation of dispositions into performance. Wass et al. (2001) demonstrated the richness of available formats in building competence measurements in medicine that capture different levels of proximity to real-life situations: Besides multiple-choice and constructed-response items or performance assessments in real-life or laboratories, they suggested video-based assessments using representative job situations so that the perception of real-life, i.e. unstructured situations, can be included. Also the speed of performance which provides information not available with accuracy (Stanovich, 2009) has increasingly been examined with the advent of computer-based testing. Blömeke et al. (in press) developed different assessment formats to capture teacher competence in terms of different knowledge facets as well as perceptual, interpretation and decision-making skills as well as their speedy reaction to student errors. And the Comparative Judgment procedure, based on Thurstone's early work, represents an interesting implementation in assessments of authentic Design and Technology tasks (Kimbell, 2006).

This challenges us to make productive, integrative use of performance assessments, traditional discrete items and other innovative formats in competence measurement. One consequence of combining formats is that when selected-response and performance tasks are scaled together, unless specific weights are assigned to performance data, selected-response data "swamps" the signal provided by the performance tasks. Additional challenges now arise, among others is the (multi) trait-(multi) method issue of distinguishing constructs from methods (Campbell & Fiske, 1959). Do differences between the results of analytic and holistic instruments reflect differences in the methods used or are we talking about different constructs which should consequently then be labeled differently? This problem can be thought of as a sampling problem, that is of defining the sampling frame for constructing performance assessments—does the frame involve sampling of assessment methods in addition to the sampling of items, raters and test takers?

Beyond dichotomies: Essential unidimensionality/multi-dimensionality

One of the fundamental challenges in competence assessment is to reduce a large amount of observational complexity into scores which maintain meaningfulness and interpretability. To this end one of the classic principles upon which measurement is based is the principle of unidimensionality. This principle fundamentally states that the different components (e.g., items, tasks, ratings) of an assessment should reflect one and the same underlying dimension. It should be noted that the principle of unidimensionality does not imply any requirement that the different components should in themselves be simple; they can, for example, be complex authentic tasks as used in performance assessments (Gustafsson & Åberg-Bengtsson, 2010). However, a strict application of the principle of unidimensionality rarely is possible in competence assessments because conceptually it typically is not expected (e.g., analytic approach with cognition and affect) and empirically it is violated by the presence of method variance and multiple-expected dimensions. Such challenges have typically been met by splitting the construct into more narrow sub-constructs, each of which satisfies the assumption of unidimensionality. While such approaches typically are successful in the sense that statistical criteria of unidimensionality are met, the approach in itself is self-defeating because the construct itself is splintered into pieces (Gustafsson, 2002).

An alternative approach is to focus instead on "essential unidimensionality" which preserves the construct while allowing for additional minor dimensions and different sources of method variance. Models for essential unidimensionality can be implemented in different ways, for example with so-called bi-factor models or hierarchical measurement models (Gustafsson & Åberg-Bengtsson, 2010; Reise, 2012). Such models identify a general factor hypothesized to represent the construct but also allowing for minor dimensions. Given that competence dimensions may be assumed to be multidimensional while at the same time a common underlying dimension is expected, this approach may be particularly useful in developing and understanding competence assessments. An extended version of this approach is multidimensional item response theory (MIRT) which is able to model several latent traits simultaneously and thus provides a promising approach to competence assessments. However, it may be argued that the ideas of essential unidimensionality or multi-dimensionality still do not solve the fundamental dimensionality issue, because there are limits to how far these approaches may be stretched.

Beyond dichotomies: Psychometric pluralism

Way too long CTT and IRT have been regarded as another allegedly incompatible dichotomy. We see a continuum from linear CTT models to nonlinear IRT models and beyond. Each theory has something to contribute to our understanding of competence measurement with respect to item/task functioning, scalability, reliability and validity of assessment scores. Of course, different models have been developed to solve different problems so models should be carefully selected to suit the particular problem at hand.

For example, IRT is useful for forming scales, examining the dimensionality of competence (as pointed out above), estimating persons' scores, and typifying levels of competence to provide criterion-referenced interpretation. IRT makes two important contributions, especially within the context of criterion-referenced testing: First, IRT produces interval-scale measurements and, second, it links individual performance to levels of performance that can be exemplified by items an individual at a particular ability (theta) has some (e.g., .5) probability of performing—anchoring the interpretation of the score in the items and not in rank order. This link between performance on items and scale levels is one of the main approaches for investigating the meaning and characteristics of a scale.

CTT, in particular GT, is in contrast useful for assessing the impact of inconsistencies due to tasks, raters and their combinations with persons, on the basis of which an optimal assessment design can be set forth. This strength is particularly important in the field of competence assessments because rater effects and temporal instability tend to be large in more complex studies. GT can thus be helpful in estimating the extent of measurement error as a first step and then to estimate the effects of re-designing a study by using more or better trained raters or more tasks.

For example Shavelson (2012) suggests an assessment approach based on a criterion-sampling approach and shows the close link to GT—a mixed model sampling theory of measurement. The variance of a score is split up so that the error variance resulting from inconsistencies between raters, task difficulty and their interactions with each other and test takers can be partialed out and only the variance of interest remains. This approach can be extended by taking measurement methods into account because

a particular competence test can be regarded as one instrument out of a broad range of possible instruments.

However, we also believe that the psychometric theories can and should be used in combination, as is sometimes done. For example, GT provides an initial step in that once reliable scores are produced, they can be IRT scaled with a number of different approaches such as partial credit or rater models. Vice versa, generalized linear mixed models and generalized latent variable modelling (e.g. Muthén, 2002; Skrondal & Rabe-Hesketh, 2004) provide ways to analyse typical "GT questions" by explicitly stating hypothesis and testing statistical models, and by offering flexible frameworks in which to deal with measurements from virtually any assessment format, data structures and a multitude of fixed or random effects (e.g. time, rater, classrooms).

Particular Applications of Interesting Assessment Approaches

Combinations of GT and IRT have been successfully applied and their usefulness demonstrated. Raudenbush, Martinez, Bloom, Zhu and Lin (2010) integrated GT and IRT in the assessment of group-level quality measures. Characteristics assumed to influence competence development such as classroom quality or opportunities to learn can be measured then in a reliable and valid way. Based on quantifying various sources of error, for example rater inconsistency, temporal instability and item inconsistencies, Raudenbush et al. (2010) developed a six-step paradigm that systematically integrates GT and IRT for the design of measurements of social settings that minimizes measurement error and thus maximizes statistical power.

We also encourage use of specialized models to approach specific research questions, such as, for example, the stability-change issue of competence. Performance can be regarded as an interaction of competence (latent abilities and dispositions) and situation. A person has to integrate several cognitive and motivational resources in order to master situational demands. Latent state-trait theory (LST) has been developed to deal with this challenge. LST is methodologically similar to the (multi) trait-(multi) method. It emphasizes that besides the person's characteristics also effects of the situation and the interaction of person and situation contribute to the variance of a variable (Steyer, Schmitt & Eid, 1999). Situational aspects can be distinguished into systematic variation of the context such as teaching different classes and into similar contexts but differential situational reactions due to working memory or exhaustion (for more details see Eid & Diener, 1999; Jenßen et al., this issue).

Beyond Dichotomies

This paper tried to tidy up the "messy construct," competence, that has been plagued by misleading dichotomies (e.g., analytic vs. holistic, IRT vs. GT, trait vs. behavior). We did not expect to find "the" one definition and statistical model for competence assessment. Rather by systematically sketching conceptual and statistical controversies and assessment approaches we attempted to clarify the construct and its measurement.

We unpacked competing competence definitions (analytic/traits vs. holistic/real-world performance) and identified commonplaces. This led to the construction of a framework for moving beyond dichotomies to show how the analytic vs. holistic approaches complemented one another (Figure 1). The measurement of competence, then, may be viewed along a continuum from traits (cognitive, affective, motivational) that underlie the perception, interpretation and decision making that give rise to observed behavior in a particular real-world situation. Dichotomies arise because one position looks at only one part of the continuum (e.g., underlying traits) while another position looks only a different part (behavior in criterion situation). We hope that the proposed integrated perspective moves us beyond dichotomies.

We unpacked competing statistical approaches to modeling competence-assessment scores, namely IRT (latent trait) vs. GT (sampling error variance). Once again we viewed these models not as dichotomies but as arraying along a continuum of linear to non-linear models. Rather than competing, the various statistical models serve different purposes. IRT models may be used for scaling item responses and modeling structural relations and GT models for pinpointing sources of measurement error variance and thereby enabling the design of reliable measurements.

Finally, we would like to point out that the proposed framework (Figure 1) is not only heuristic in suggesting multiple new research studies but also in viewing it as a "grand" structural model. The analytic (latent trait) side of the model (left-side of Figure 1) includes indicators for cognitive, affective and motivational traits demanded in particular contexts/situations. Such competencies are structurally related to real-world performance (right-side) through a set of perceptual, interpretive and decision-making processes (middle). Research on competence measurement, then, might work on various parts of the model and even attempt to test the entire model conceptually and statistically.

Viewing competence as a continuum and applying a corresponding range of assessment formats required by the framework is conceptually and methodologically challenging. But we believe that solutions exist or can be developed to deal with these challenges and we tried to sketch out possible approaches to trustworthy competence assessments that overcome the risk of forgetting either observable behavior or cognitive abilities. If our reasoning holds, it opens up for a great range of research questions.

With the proposed integrated approach and the improvement of measurement of competence, the field of higher education will be in a position to address important, substantive questions. For example, we should be able to examine the developmental trajectories of competence, identify groups of students with differential developmental patterns, and determine effective educational strategies for development. We should be able to go beyond immediate measurement of behavior in situ to longer-term measurements of life outcomes beyond earning and including health, family, and civic and social engagement. We should also be able to study the interaction of perception, interpretation and decision making in the education and training of students for particular life outcomes.

Higher education is certainly a field with huge research gaps. By providing this overview and by editing this special ZfP issue, we hope to inspire and encourage many colleagues to look into this field and to take up the challenge what it means to define and assess competence acquired in higher education.

References

Arthur, W., Day, E. A., McNelly, T. L. & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154.

Bakker, A. B. (2011). An Evidence-Based Model of Work Engagement. *Current Directions in Psychological Science, 20,* 265–269.

Baltes, P. B.*,* Reese, H. W. *&* Lipsitt, L. P. (1980). Life-span developmental psychology*. Annual Review of Psychology, 31,* 65-110*.*

Bennett, Y. (1993). The validity and reliability of assessments and self-assessments of workbased learning. *Assessment & Evaluation in Higher Education, 18 (*2), 83–94.

Benjamin, R. (2013). The principles and logic of competency testing in higher education (pp. 127-136). In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges.* Boston: Sense.

Berry, C. M., Clark, M. A. & McClure, T. (2011). Black-White differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96,* 881-906.

Biemans, H., Nieuwenhuis, L., Poell, R., Mulder, M. & Wesselink, R. (2004). Competence-based VET in The Netherlands: Backgrounds and pitfalls. *Journal of Vocational Education and Training, 56,* 523-538.

Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation, 33,* 29–49.

Blömeke, S., Busse, A., Suhl, U., Kaiser, G., Benthien, J., Döhrmann, M. & König, J. (in press). Entwicklung von Lehrpersonen in den ersten Berufsjahren: Längsschnittliche Vorhersage von Unterrichtswahrnehmung und Lehrerreaktionen durch Ausbildungsergebnisse. *Zeitschrift für Erziehungswissenschaft.*

Blömeke, S., Gustafsson, J-E. & Shavelson, R. (2013). Assessment of competencies in higher education: A topical issue of the Zeitschrift fur Psychologie. *Zeitschrift für Psychologie, 221*, 202.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, Ch. & Fege, J. (Eds.) (2013). *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges* (= Professional and VET Learning; 1). Rotterdam, The Netherlands: Sense Publishers.

Boyatzis, R.E. (1982). *The competent manager.* New York: Wiley.

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

Brief, A. P. & Weiss, H. M. (2001). Organizational behavior: Affect in the workplace. *Annual Review of Psychology, 53,* 279–307.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* New York: Houghton Mifflin.

Corno, L., Cronbach, L. J.,Kupermintz, H., Lohman, D., Mandinach, E. B., Porteus, A.W. et al. (2002). *Remaking the concept of aptitude: Extending the legacy of R. E. Snow.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Eid, M. & Diener, E. (1999). Intraindividual Variability in Affect: Reliability, Validity, and Personality Correlates. *Journal of Personality and Social Psychology, 76*(4), 662-676.

Epstein, R. M. & Hundert, E. M. (2002). Defining and Assessing Professional Competence. *JAMA, 287*, 226-235.

Eraut, M. (2003) National vocational qualifications in England: Description and analysis of an alternative qualification system. In: G. Straka (Ed.), *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen.* Münster: Waxmann.

Förster, F., Zlatkin-Troitschanskaia, O., Brückner, S., Happ, R., Ronald K. Hambleton2, William B. Walstad3, Tadayoshi Asano4 & Michio Yamaoka5 Validating Test Score Interpretations by Comparing the Results of Students from the United States, Japan and Germany on a Test of Economic Knowledge in Higher Education

Gold, B., Förster, St. & Holodynski, M. (2013). Evaluation eines videobasierten Trainingsseminars zur Förderung der professionallen Wahrnehmung von Klassenführung im Grundschulunterricht. *Zeitschrift für Pädagogische Psychologie, 27,* 141-155.

Grant, G., Elbow, P. & Ewens, T. (1979). *On Competence: A critical analysis of competence-based reforms in higher education.* San Francisco: Jossey-Bass.

Gustafsson, J-E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.) *The role of constructs in psychological and educational measurement* (pp. 73-95). London: Lawrence Erlbaum Associates, Publishers.

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.

Hakanen, J. J. & Schaufeli, W. B. (2012). Do burnout and work engagement predict depressive symptoms and life satisfaction? A three-wave seven-year prospective study. *Journal of Affective Disorders, 141,* 415-24.

Handley, D. (2003) Assessment of competencies in England's National Vocational Qualification system. In: G. Straka (Ed.), *Zertifizierung non-formell und informell erworbener beruflicher Kompetenzen.* Münster: Waxmann.

Jenßen, L., Dunekacke, S. & Blömeke, S. (this issue). The Relationship of Mathematical Competence and Mathematics Anxiety in prospective Kindergarten Teachers: An Application of Latent State-Trait Theory. *Zeitschrift für Psychologie.*

Judge, T. A., Thorensen, C. J., Bono, J. E. & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin, 127,* 376–407.

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

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R. & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49,* 568-589.

Kimbell, R. A. (2006). Innovative Technological Performance (pp. 159-179). In Dakers, J. (Ed), *Defining Technological Literacy: Towards an Epistemological Framework*. Palgrave Press.

König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A. & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education, 38,* 76-88.

Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Zeitschrift für Psychologie, 216,* 61–73.

Kounin, J. S. (1970). *Discipline and group management in classrooms.* New York: Holt, Rinehart, & Winston.

Kulgemeyer, Ch., Tomczyszyn, E. & Schecker, H. (this issue). Assessing teacher trainees' competencies for explaining physics: A process-oriented methodological approach. *Zeitschrift für Psychologie.*

Lau, S. & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment, 8*(2), 139–162.

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

McLachlan, G. & Peel, D. A. (2000). *Finite mixture models*. New York: Wiley.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Miller, G. E. (1990). The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine: Journal of the Association of American Medical Colleges, 65,* 63–67.

Mulder, M., Gulikers, J., Biemans, H. & Wesselink, R. (2009). The new competence concept in higher education: error or enrichment? *Journal of European Industrial Training, 33,* 755-770.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81-117.

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

Raudenbush, S. W., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2010). Studying the reliability of group-level measures with implications for statistical power: A six-step paradigm. University of Chicago Working Paper.

Reise, S. P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research, 47*(5), 667-696.

Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185-205.

Sadler, R. (2013). Making competent judgments of competence. In Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, Ch. & Fege, J. (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges* (pp. 13-27). Rotterdam, The Netherlands: Sense Publishers.

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

Shavelson, R. J. (2012). An approach to testing and modeling competencies. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges.* Boston: Sense.

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

Shulman. L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57,* 1-22.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Chapman and Hall/CRC.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence.* New York: Cambridge University Press.

Sparrow, P. R. & Bognanno, M. (1993). Competency requirement forecasting: Issues for international selection and assessment. *International Journal of Selection and Assessment, 1,* 50-58.

Spencer, L. M., Jr. & Spencer, S. M. (1993). *Competence at work: Models for superior performance.* New York: Wiley.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought.* New Haven: Yale University Press.

Sternberg, R. J. & Grigorenko, E. L. (Eds.), The Psychology of Abilities, Competencies, and Expertise. Cambridge, MA: Cambridge University Press.

Steyer, R., Schmitt, M. & Eid, M. (1999). Latent State-Trait Theory and Research in Personality and Individual Differences. *European Journal of Personality, 13,* 389-408.

Stürmer, K., Könings, K. D. & Seidel, T. (2012). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology, 83,* 467–483.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge University Press.

Wass, V., Van der Vluten, C., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. *Lancet, 357,* 945–949.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45-66). Göttingen: Hogrefe.