KAROLINE KOEPPEN[1], JOHANNES HARTIG[i], ECKHARD KLIEME[i]
AND DETLEV LEUTNER[2]

COMPETENCE MODELS FOR ASSESSING INDIVIDUAL LEARNING OUTCOMES AND EVALUATING EDUCATIONAL PROCESSES – A PRIORITY PROGRAM OF THE GERMAN RESEARCH FOUNDATION (DFG)[3]

INTRODUCTION

Social change, social cohesion, and opportunities for societal development are all dependent on the educational level of the members of a society. Current discussion in educational research emphasizes the importance of the products of educational processes, often referred to as educational *output* or *outcomes*, for human resources (Klieme & Leutner, 2006). The outcomes of education are the knowledge acquired, the abilities, skills, attitudes, and dispositions developed, and the qualifications attained.

Several large-scale international assessments of domain-specific competencies (e.g., reading literacy, science competencies) at the end of compulsory schooling (e.g., TIMMS, PISA) and in adulthood (e.g., IALS, ALL) have recently drawn increased public and scientific attention to educational outcomes and their assessment. The studies identified huge gaps between the competencies attained, on the one hand, and the goals of the education system, on the other. Clearly, effective quality development of educational processes is facilitated when the productivity of educational systems, the quality of educational institutions, and the learning gains of individuals are measurable. Thus, there has been an increasing focus within educational systems on defining and evaluating the goals to be attained by schools. In many cases, however, adequate assessment procedures are still lacking, as are procedures for analyzing and reporting the results.

The concept of competence is increasingly considered as an anchor point in this discussion. In this article, we focus on two sets of research questions that are central to the debate on the concept of competence. In the first part of the article, we discuss how competencies can be defined in general and in specific contexts. The new focus on competence has shifted attention from the measurement of general cognitive abilities to more complex ability constructs related to real world contexts. Sophisticated models of the structure and levels of these complex constructs need to be developed. Typical examples are different levels of reading literacy, mathematical modeling of real-world situations, planning and analyzing scientific experiments, or self-regulation and metacognition in domain-specific

S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), Modeling and Measuring Competencies in Higher Education: Tasks and Challenges, 171–192.

problem solving. Given the complexity of competencies, it is important that they be precisely defined in specific domains. The development of cognitive models of domain-specific performance is a central issue in this context.

The second set of research questions relates to the design and practical implementation of competence assessments. School policy and practice are moving toward *evidence-based policy and practice* (Slavin, 2002), where "evidence" often implies empirical assessments of competencies. Obviously, the assessment of competencies plays a key role in optimizing educational processes and advancing educational systems. At the same time, it is evident that assessments pursue various goals (i.e., the focus may be on individual learning outcomes, on program evaluation, or on system monitoring). Unfortunately, the difficulties and complexities of assessing learners' baseline competencies and learning gains are often underestimated in educational policy and practice. Developing appropriate measurement instruments that can be used for different purposes is a time- and resource-intensive undertaking that can only be achieved on the basis of theoretically and empirically founded cognitive models of competencies.

In this article, we give an overview of current issues in cognitive modeling and competence assessment. We first provide a working definition of the term competence and describe different goals of competence assessment. In the main part of the article, we outline the central research questions and the current state of research, identifying four main research areas. Finally, we present an interdisciplinary research program funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) to integrate, structure, and coordinate research activities relating to competence modeling and assessment.

The article focuses on current assessment practices in Germany, where the standardized assessment of student achievement does not have the same tradition as in the United States or Great Britain. In the past decade, however, the results of large-scale international assessments, particularly the PISA 2000 study (e.g. Baumert, Stanat, & Demmrich, 2001), have sparked intense discussion among both the public and educational policy makers. Extensive educational reforms have been initiated in response to the PISA findings (e.g., changes in mathematics and science instruction), and the results of the PISA 2006 assessment seem to indicate that these reforms have had positive effects on students' performance, especially in science (OECD, 2007b; Prenzel et al., 2007).

THE COMPETENCE CONCEPT AND THE CHALLENGES OF ITS ASSESSMENT

The competence concept is central to empirical studies dealing with the development of human resources and the productivity of education. Although it has been in use for decades, the term "competence" has enjoyed increasing currency in psychology and its neighboring disciplines in the last few years (e.g., Csapó, 2004; Klieme, Funke, Leutner, Reimann, & Wirth, 2001; Rychen & Salganik, 2001, 2003; Sternberg & Grigorenko, 2003; Weinert, 2001). Research uses the concept to characterize the changing demands of modern life and the working world, as well as the educational goals involved. However, its definition

remains fuzzy in educational research. It seems essential to narrow it down to specific contexts of abilities.

Drawing on Klieme and Leutner (2006; see also Klieme, Maag-Merki, & Hartig, 2007), we define competencies as context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains.

An essential element of competencies is their *context-specificity*. The concept of competence was introduced to psychology as an alternative to the focus in classical intelligence research on generalized, context-independent cognitive dispositions that are learnable only to a limited extent (e.g., McClelland, 1973; "Testing for competence rather than for 'intelligence'"). In contrast, competencies reflect a person's potential to meet cognitive demands in specific areas of learning and behavior. Competencies are, thus, more closely related to "real life". Connell, Sheridan, and Gardner (2003, p. 142) concisely characterize competencies as "realized abilities."

Having considered different theoretical and pragmatic arguments, Weinert (1999, 2001) proposed that the term competence be restricted to *cognitive* context-specific aspects, and that it should exclude motivational orientations or affective requirements for successful learning. Given that Weinert himself also discussed so-called *action competencies*, including motivation, attitudes, tendencies, and expectations in the context of competencies, this distinction was not self-evident. Nonetheless, Weinert proposed that cognitive and motivational aspects be assessed as separate constructs to allow the empirical analysis of their interaction. In this article, we focus on the cognitive aspects of competencies.

In psychological and educational practice and research, competencies often relate to specific content areas (e.g., Hartig & Klieme, 2006; Weinert, 2001). In the tradition of research on psychological expertise, we refer to these areas as *domains*. Typical domain-specific competencies in primary and secondary education include reading literacy, mathematical competence, and scientific competence.

Given their context-specificity, competencies have to be acquired by learning and experience in relevant, domain-specific situations. Consequently, they are amenable to external interventions (e.g., Baumert et al, 2001; Hartig & Klieme, 2006; Simonton, 2003). Basic cognitive abilities, in contrast, are much more difficult to train or learn (Weinert, 2001). In the construction of competence models, it is therefore important to consider and empirically examine the connections between specific competencies and basic cognitive abilities.

Valid measures of competence need to be based on theoretically sound and empirically tested competence models. These models have to (a) represent the internal structure of competencies in terms of specific basic skills and abilities, (b) describe different levels of competencies with reference to domain-specific performance, and (c) take into account changes occurring in learning and developmental processes.

In addition, measurements of competence should build on psychometric models that link the empirical measurement operations with theoretical (cognitive) models of competencies. In short, the measurement of competencies should be based on a

solid theoretical and psychometric basis that allows the measurement result (e.g., quantity and quality of solved tasks) to be interpreted with reference to an underlying theoretical model of competencies.

Valid, model-based measures of competence can be used for different purposes. First, model-based measurement instruments can inform individual educational decisions (e.g., assignment to a certain track, conferral of qualifications, provision of educational interventions). In this context, assessment focuses on individual learning outcomes; it is "a process by which educators use students' responses to specially created or naturally occurring stimuli to draw inferences about the students' knowledge and skills" (Pellegrino, Chudowsky, & Glaser, 2001, p. 20). A second goal of competence assessment is to evaluate learning outcomes on the aggregated class, school, or even system levels, rather than the individual level (Leutner, Fleischer, Spoden, & Wirth, 2007). This ranges from classroom-based assessment to large-scale standardized assessment of competence levels across whole education systems (system monitoring; for example, the National Assessment of Educational Progress [NAEP] in the United States or the OECD PISA studies).

Assessments with a focus on individual learning outcomes, on the one hand, and different aggregated levels, on the other, make distinct demands on measurement instruments (e.g., in terms of reliability and testing time). Depending on the focus of the assessment and the level of aggregation, different measurement techniques and research designs may be more or less suitable. In many cases, however, different goals have to be accomplished within the same study (e.g., system monitoring and feedback on classroom level). Thus, another challenge in the research area of competencies is to develop competence measures and research designs that simultaneously satisfy different assessment goals.

CENTRAL RESEARCH QUESTIONS AND THE CURRENT STATE OF RESEARCH ON COMPETENCE MODELING AND ASSESSMENT

The development of adequate cognitive models for contextualized competence constructs is a challenging and multifaceted task. Theoretical models must provide a basis for describing the interaction between individual abilities and the environment, different levels of competence, and developmental processes. Furthermore, they must be related to advanced psychometric techniques and translated into appropriate empirical measurement procedures. As yet, neither cognitive research nor psychometrics meets these requirements; adequate measurement concepts and models are still lacking (Prenzel & Allolio-Näcke, 2006). Both disciplines need to contextualize their models in cooperation with representatives of other disciplines, such as educational researchers and domain experts. Indeed, the Committee on the Foundations of Assessment (Pellegrino at al., 2001), founded by the US National Research Council, has called for multidisciplinary research activities focusing on three different facets:

> "(1) development of cognitive models of learning that can serve as the basis for assessment design, (2) research on new statistical measurement models and their applicability, (3) research on assessment design" (p. 284).

As Pellegrino et al. (2001) accurately summarize: "Much work remains to focus psychometric model building on the critical features of models of cognition and learning and on observations that reveal meaningful cognitive processes in a particular domain (…). Therefore, having a broad array of models available does not mean that the measurement model problem has been solved. The long-standing tradition of leaving scientists, educators, task designers, and psychometricians each to their own realms represents perhaps the most serious barrier to progress" (p.6).

We identify four key areas in this research field: first and foremost, the development of theoretical models of competence (Area 1), complemented by the construction of psychometric models (Area 2). This leads onto the construction of measurement instruments for the empirical assessment of competencies (Area 3). Research on the use of diagnostic information (Area 4) rounds off the research field. In the following, we explicate the concrete questions and problems addressed within each of the four areas and outline the current state of research.

Area 1: Development of Cognitive Models of Competencies

As mentioned above, the shift toward the competence construct has prompted efforts to improve the assessment of these complex and contextualized constructs. The first question to arise here is which models provide a basis for developing measurement instruments and interpreting their results. In current educational research, only a limited number of competence models exist. Therefore, it is important to develop cognitive models that explain interindividual differences in domain-specific performance.

A first challenge in model development is the *contextualized character of competencies*, which means that both person- and situation-specific factors have to be taken into account. For example, when describing foreign language skills with reference to situational demands, the competencies required to read a text can be distinguished from those required to engage in conversation (e.g., by distinguishing written vs. spoken text, or text comprehension vs. text production). For individuals, knowledge structures relevant to different situations must be taken into account; for example, the available vocabulary, grammatical knowledge, and mastery of socio-pragmatic rules (Chen, 2004; Kobayashi, 2002). This simultaneous consideration of individual- and situation-specific components has consequences for the structure of competencies as well as for the description of competence levels. Hence, two groups of theoretical models devised to describe and explain competencies can be distinguished: *models of competence levels* and *models of competence structures* (Hartig & Klieme, 2006; Klieme et al., 2007). Models of competence levels define the specific situational demands that can be mastered by individuals with certain levels or profiles of competencies; levels of competencies are used to provide a criterion-referenced interpretation of measurement results. These models are particularly useful for assessing and evaluating educational

outcomes on an aggregated level. Models of competence structures deal with the relations between performances in different contexts and seek to identify common underlying dimensions. These models are especially interesting for explaining performance in specific domains in terms of underlying basic abilities, and can provide a basis for more differentiated measurement results of individual-centered assessments. The two kinds of models relate to different aspects of competence constructs. They are not mutually exclusive, but ideally complementary.

The aspect of *development* is also very relevant in the context of theoretical competence models. To date, only a few competence models have addressed the issue of competence development (primarily in the domain of science; e.g., Bybee, 1997; Prenzel et al., 2004, 2005). For the most part, these models have no empirical foundation, and their conceptualizations of competence development differ. Some models see competence development as a continuous progression, shifting successively from the lowest to the highest competence level (e.g. Prenzel et al., 2004, 2005). The level of elaboration and systematization increases with the competence level (as described by Bybee, 1997, for scientific literacy). Other models conceptualize competence development as a noncontinuous process characterized by qualitative leaps (e.g., conceptual change in science; Schnotz et al., 2004; Schnotz, Vosniadou, & Carretero, 1999). This process involves a fundamental reorganization of concepts and structures from everyday life to correspond with new science-based ideas (e.g., DiSessa, 2006; Vosniadou, Ioannides, Dimitrakopoulou, & Papademetriou, 2001; Wilson, 2008).

In addition, the design of cognitive models of competencies depends on the questions addressed or the decisions to be informed. A model fitting for some purposes (e.g., giving immediate feedback) may be totally ineffective for other purposes (e.g., comparative evaluation of educational institutions). A more detailed model of competencies is needed in the first case than in the second. In one case, precise estimates might be required on an individual level, in another case on an aggregated level. Switching between two purposes can cause a whole host of problems, as recent experiences in the United States have shown (Cheng, Wanatabe, & Curtis, 2004; Fuhrmann & Elmore, 2004).

In a next step, these theoretically founded models of competencies must be used as a basis for constructing psychometric models and measurement instruments (e.g., Hartig & Höhler, 2008; Wirth & Leutner, 2008). To date, however, these efforts have rarely proved successful, primarily because many of the existing cognitive models are insufficiently elaborated. Nevertheless, some research progress has been made in specific domains, such as foreign languages in the context of the Common European Framework of Reference (Alderson, 2005; Alderson et al., 2005; Gogolin, 2002), or in the area of mathematical modeling (e.g., Blum et al., 2004). Likewise, important contributions to the theory-based formulation of competence models have been made in specific areas of cognitive psychology (e.g., Frensch et al., 2003; Haider & Frensch, 1996, 1997, 2002; Hasselhorn & Grube, 2003; Oberauer, Schulze, Wilhelm, & Süß, 2005; Schneider, Lockl, & Fernandez, 2005; Spiel & Glück, 2008; Weinert & Schneider, 1995) and in research on personality and individual differences (e.g., Kröner, Plass, &

Leutner, 2005; Leutner, 2002; Leutner & Plass, 1998; Plass, Chun, Mayer, & Leutner, 1998; Wilhelm & Engle, 2005).

However, when it comes to developing actual test items, the level of abstraction of the competence models often turns out to be too high. As a consequence, test developers have to develop huge numbers of tasks that then have to be tested empirically. Those tasks that correspond to a (usually) relatively simple psychometric model (e.g., a unidimensional Rasch model) are retained. The scales and competence levels reported in the PISA study are an example for this procedure (e.g., in reading: Artelt, Schiefele, & Schneider, 2001; in mathematics: Klieme, Neubrand, & Lüdtke, 2001; in science: Prenzel, Rost, Senkbeil, Häußler, & Klopp, 2001; in cross-curricular problem solving: Dossey, Hartig, Klieme, & Wu, 2004). In these examples, levels of competence are not theoretically specified a priori, but defined post hoc after the inspection of model-conform leftover items. From a theoretical perspective, this procedure is less than satisfactory.

To summarize, in many domains where the need for well-founded competence assessments is evident, basic research concerning theoretically as well as empirically sound models of competence structures, competence levels, and competence development is still required. Although attempts have been made to interconnect cognitive competence models with psychometric models and measurement instruments, they have often failed to meet the demands of the current, more complex definition of competencies. There is a clear need for more integrative, interdisciplinary research activities.

Area 2: Psychometric Models

As Embretson (1983, p. 184) put it, psychometric models are about "modeling the encounter of a person with an item". Psychometric models are the link between theoretical constructs and the results of empirical assessments; they provide the measurement rules by which test scores are assigned based on performance in test situations. Given the contextualized nature and complexity of competence constructs, psychometric models have to meet certain requirements (Hartig & Klieme, 2007). On the one hand, they have to incorporate all relevant characteristics of the individuals whose competencies are to be evaluated. Because competencies refer to performance in complex domains, the models should take into account that multiple abilities may be required. At the same time, they have to take into account domain-specific situational demands. Because competencies are conceptualized as context-specific constructs, the results of competence assessments should be related to the mastery of specific, domain-relevant situations. Item response theory (IRT) has a long tradition in educational assessment, and many of its past and recent developments were made for specific needs in this area. IRT allows ability estimates and item difficulties to be compared (Embretson, 2006), thus providing a basis for models incorporating individual and situational characteristics. Several recent developments in IRT hold considerable promise for the modeling of competencies, namely, *explanatory IRT models,*

multidimensional IRT models (e.g., Hartig & Höhler, 2008), and *models for cognitive diagnosis.*

Explanatory IRT models (Wilson & De Boeck, 2004; Wilson, De Boeck, & Carstensen, 2008) incorporate predictors for successful interactions of a person with an item. These predictors can be either attributes of the person or features of the item ("person predictors" or "item predictors"). Specific item features can be used to represent certain situational demands. Incorporating effects of item features into the psychometric model is a highly suitable way of constructing a psychometric model of competence that takes the corresponding demands into account. Although models including item features have been in use for some time (e.g., the linear-logistic test model, LLTM; Fischer, 1973), recent developments such as the inclusion of random effects on the items side (e.g., Janssen, Schepers, & Peres, 2004; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000) make them more flexible to model empirical data from complex performance situations. Applications of models with item predictors have been presented by Janssen et al. (2000, 2004), Hartig and Frey (2005), and Wilson et al. (2008). Whereas the analysis of item predictors is highly promising for psychometric models of competence, the use of observed person predictors for latent abilities ("latent regression," e.g. Adams, Wilson, & Wu, 1997; van den Noortgate & Paak, 2004) is of less interest in this context. Observed person predictors may, however, be incorporated into models of competence in order to model interactions between certain person characteristics, abilities, and task demands (e.g., in differential item functioning).

Models with item features that allow situational demands to be incorporated (e.g., the LLTM) are typically unidimensional. To model performance in complex situations, it may be necessary to include more than one ability dimension in the model. A straightforward way of doing so is to apply *multidimensional IRT* (MIRT) models. These models are generalizations of unidimensional models such as the Rasch model, the two-parameter logistic model, and the normal-ogive model. Instead of one ability dimension, the probability of mastering a test item is modeled as a function of multiple basic abilities (e.g., McDonald, 2000; Reckase, 1997). The most frequently applied models are compensatory models, which model the probability of success on an item as a function of the sum of all abilities relevant for an item. In these models, low ability in one dimension can be compensated by high ability in a second dimension, and vice versa. However, non-compensatory models have also been described (e.g., Whitely, 1981). MIRT models differ in their complexity in terms of how many different abilities are required to solve each item. Models in which each item draws on a single ability are said to have between-item multidimensionality (Adams, Wilson, & Wang, 1997). In factoranalytic terms, these models have a simple-structure loading pattern. MIRT models with between-item multidimensionality have been applied to data from educational assessments to take into account relations between performance in different domains (e.g., in the PISA studies; Adams, 2005; Adams & Wu, 2002). Models with between-item multidimensionality consist of multiple scales that are, within themselves, unidimensional. In contrast, models that incorporate multiple abilities for each item

are said to have within-item multidimensionality. These models are more appealing for psychometric models of competencies, because within-item multidimensionality makes it possible to model successful performance as the result of a mixture of different abilities. Models with within-item multidimensionality have been applied relatively rarely, typically to account for "nuisance" dimensions (e.g., local item dependencies within testlets; Wang & Wilson, 2005) rather than for theoretically defined ability dimensions. Examples of simple MIRT models with meaningful within-item multidimensionality are given by Walker and Beratvas (2003), Stout (2007), and Hartig and Höhler (2008).

A third development of great relevance to psychometric models of competencies has been the emergence of *cognitive diagnostic models* or *multiple classification models* (DiBello & Stout, 2007; Maris, 1999). These models assume multiple latent ability variables that are modeled as latent categories instead of latent dimensions. They can be characterized as multidimensional latent class models with specific restrictions defining which items require which abilities, and how the abilities are combined for successful performance. A conceptual strength of cognitive diagnostic models is that they are explicitly designed to model mixtures of abilities within items, and that models and estimation methods for non-compensatory mixtures are available. However, the categorical nature of the ability variables (e.g., "does know" vs. "does not know") seems more appropriate for relatively fine-grained ability constructs than for cases in which latent variables represent a broader set of required abilities. Examples of empirical applications of cognitive diagnostic models are presented by von Davier (2005) and Gierl, Leighton, and Hunka (2007).

To summarize, recent years have seen a number of significant developments in psychometrics that hold great promise for the translation of theoretical models of competencies into measurement models. Models that succeed in taking both situational characteristics and individual abilities into account can do more than provide rules for measurement (i.e., generate test scores). They can also serve as empirically testable models of the interaction between individual abilities and situational demands. However, to realize the potential of the advanced psychometric methods recently developed, these techniques need to be combined with strong theoretical models.

Area 3: Measurement Concepts and Instruments

This section examines how competence models and psychometric models can be translated into concrete empirical measurement procedures, with a focus on computer-based assessment.

Competencies are assessed in different educational contexts: in large-scale assessments (e.g., TIMSS and PISA), in evaluations of specific programs or institutions, in basic research, and in the assessment of individual qualifications or learning outcomes. Researchers and stakeholders in educational processes assess student competencies for purposes of system monitoring, to test the effectiveness of specific forms of instruction, to give feedback about individual learning progress, or to describe developments in competencies. For the most part,

standardized tests are applied. However, nonstandardized tests and observations of educational processes (e.g., teachers' observations in direct interaction with learners) are also common ways of assessing competencies. Given the complexity of competence constructs, it is important to adapt and advance these measurement concepts and instruments. They should be parsimonious, have a firm theoretical foundation, and allow inferences to be drawn about the mastery of demands in real-life situations.

Research interest in measurement concepts and instruments for assessing competencies first emerged in the 1960s and 1970s. In Germany, this was a time of educational reform, with the introduction of new teaching and learning goals to the curriculum, as well as the establishment and evaluation of new educational curricula (e.g., in schools combining different academic tracks). In this context, there was a surge in interest in the area of educational assessment at the individual, diagnostic level: the traditional field of activity for competence assessment (Klauer, 1978). Assessment instruments were developed, based on the concept of goal-oriented and criterion-referenced testing and, usually, using the binomial test model or one of its derivates (Klauer, 1987). At the same time, IRT became increasingly popular and widespread in educational measurements. In some countries (e.g., the Netherlands and the United States), this tradition has continued unabated (van der Linden & Hambleton, 1996).

Given the complexity of competence constructs and the need to understand the different abilities and processes that lead to success in real-life situations, it has become increasingly important that assessment procedures are based on cognitive models of competence. An excellent example of empirical assessments based on theoretical models of competence is the Berkeley Evaluation & Assessment Research (BEAR) Center, which focuses on the model-based assessment of competencies in science education (Wilson, 2008; Wilson & Draney, 2004; Wilson & Sloane, 2000). In DESI, a large-scale assessment of language competencies in Germany, measurement instruments and measurement models were developed on the basis of cognitive and linguistic models of language competence and language acquisition (e.g., Beck & Klieme, 2007; Klieme et al., 2008; Nold, 2003; Nold & Rossa, 2007a, b).

In the context of developing new measurement concepts, it is important not to overlook innovative measurement procedures, many of which capitalize on new technologies. Technology-based assessment (TBA) are widely used in educational settings in the United States and some European countries (Hartig, Kröhne, & Jurecka, 2007). In Germany, however, TBA is used primarily in psychological research, and is not yet well established in educational practice. During the 1990s, the use of TBA in psychological and educational competence assessment became increasingly widespread. This kind of assessment has numerous advantages: It allows complex stimuli and response formats, interactive testing procedures, real-time assessment of cognitive processes (Wirth, 2004; Wirth & Klieme, 2003), and automatized analysis and feedback procedures (Chung, O'Neil, & Baker, 2008; Ordinate, 2004; Reeff, 2007). In addition, TBA offers the possibility of computerized adaptive testing (CAT; e.g., van der Linden, 2005), in which the items presented are selected to fit the individual ability level of the test-taker.

Computer-adaptive testing allows for dynamic testing. The concept of dynamic testing is already well-established in the domain of intelligence assessment, but it can also be transferred to other performance domains. Dynamic testing focuses on the potential for intellectual development and is applicable in areas where the assessment of the status quo of a certain competence is unsatisfactory.

Furthermore, technology-based assessment permits the construction of complex and interactive stimuli that would be very costly or impossible to realize without the use of computers. It, thus, affords the possibility to empirically assess new competence domains that were not assessable with traditional measurement procedures. Because TBA allows the *simulation of complex and dynamic situations*, assessment designs can be more valid with respect to the demands of real-life, complex situations (Drasgow, 2002). For example, virtual patients can be used for competence assessment in medical education (Jung, Ahad, & Weber, 2005). Virtual environments can be used to examine individual navigation skills or, in networked environments, interactions between different individuals (Frey, Hartig, Ketzel, Zinkernagel & Moosbrugger, 2007). The PISA 2009 study includes a new component assessing the competence to read electronic texts (OECD, 2007a), which can be conceptually distinguished from the competence to read printed text. Thanks to technology-based procedures, the real-life situation of reading a hypertext can be simulated for these assessments. Test-takers' behavior can be recorded in log files, allowing their navigation within electronic tests to be analyzed and process indicators to be constructed that supplement responses to the test items (e.g., Wirth, 2004; Wirth & Leuter, 2008; Künsting, Thillmann, Wirth, Leutner, & Fischer, 2008).

In addition, technology-based testing allows parsimonious assessment and data administration. In particular, computers networked in local area networks or via the internet make it possible to test and provide feedback independently of time and of the test-takers' location, and to simultaneously administer tests to large samples (ETS, 2005; Groot, de Sonneville, & Stins, 2004; Jude & Wirth, 2007).

The new possibilities afforded by TBA have been used in numerous contexts. However, many of these applications are driven by the rapid development of computer technology rather than by well-founded theories. Much empirical and theoretical work is still needed to link complex computerized measurement procedures to cognitive and psychometric models.

Area 4: Reception and Usage of Assessment Results

The success of many educational decisions and interventions hinges on accurate assessments of learners' baseline competencies and learning outcomes. Assessments may have different practical goals: On an individual level, they allow educators to select appropriate interventions for individual cases (i.e., to further individual learning). The results of individual assessments may also inform decisions on the admittance to secondary tracks or to higher education. In contrast, assessment programs that are designed to report achievement on an aggregated level serve to evaluate educational programs, institutions, or systems, as well as to inform decision makers on the administrative and political levels. As Pellegrino et al. (2001)

concluded, "one size of assessment does not fit all" (p. 222). Depending on the goal of the assessment, different measurement instruments are needed and different research questions arise. Thus a continuing challenge facing researchers in this area is thus to determine which models, measurement rules, and measurement procedures provide the appropriate information for various goals of assessment.

Assessment to further individual learning can be regarded as *formative evaluation* on an individual level. It should allow precise conclusions to be drawn about individual learning processes and learners' strengths and weaknesses with respect to specific curricular units. These conclusions can help to support individual instruction and learning, and ideally offer considerable potential to enhance teaching. Teachers make observations of students' understanding and performance in a variety of ways: In classroom dialog, homework assignments, and formal tests (Pellegrino et al., 2001). These procedures should permit diagnosis on an individual level, in terms of understanding students' individual solution paths, misconceptions, etc. (Seger, Dochy, & Cascallar, 2003; Wilson, 2008). Appropriate individual feedback is crucial to support the subsequent learning process. A number of research questions arise in this context: What kind of diagnostic information is best understood by students, and what kind by teachers? How well can teachers evaluate individual learning processes? What factors influence teachers' grading decisions? What models of competence do teachers rely on – implicitly or explicitly? How well founded and how helpful is the feedback provided by the teacher to the individual students?

The assessment of individual achievement may also entail the *summative evaluation* of an individual's competencies. These evaluations help to determine whether a student has attained a certain level of competence after completing a particular phase of education (e.g., in end-of-unit tests or the letter grades assigned at the end of a course; Pellegrino et al., 2001). These performance measurements are often *high stakes*, meaning that their outcomes have significant consequences. Students who fail to attain certain standards (e.g., passing their final school exams) may be refused access to the next level. An important question for research on assessment in this field is how tests can be constructed to reflect educational goals and how results can be interpreted with reference to curricula (e.g., Cizek, Bunch, & Koons, 2004; Haertel & Lorié, 2004; Klauer & Leutner, 2007; Klieme et al., 2003). A related research question is how the content of educational assessments affects the methods and content of instruction ("washback effects"; e.g., Cheng et al, 2004; Cizek, 2001; Fuhrman & Elmore, 2004; Nichols, Glass, & Berliner, 2006; as well as Pellegrino et al. 2001, p. 212 ff). Individual data aggregated on the classroom level can be used by teachers to evaluate their own instruction and to identify their students' specific instructional needs (e.g., Leutner et al., 2007). Teachers need detailed, contextualized information about their students' learning progress to efficiently adjust their instructional focus – like the information needed to inform decisions on an individual level. However, there is a marked gap between the information teachers need and the information they are given.

Huff and Goodman (2007) report that although most teachers in the United States receive assessment results from state mandated or commercial large-scale

assessments, 20–30% of them almost never use these results to reflect on their instruction. Moreover, 30–38% of the teachers state that the diagnostic information provided by large-scale assessments is not detailed enough to use.

The results of competence assessments on an *aggregated level* provide information about classrooms and schools. Aggregated data usually serves evaluation purposes and supports the quality development of educational processes. School principals can use classroom-level data as a basis for evaluating teachers' performance and as indicators for the need for professional development. Educational administrations can use the results of competence assessments aggregated at the school level to inform budget decisions concerning individual schools. Aggregated data from educational assessments can also be used to guide and control whole education systems on the political level, that is, for purposes of *system monitoring* (Leutner et al., 2007). Policy makers can use information aggregated on a district or country level to gauge the effectiveness of educational systems and to make decisions about measures to improve their effectiveness. Of course, the information required for such decisions differs markedly from that required by students and teachers to further learning processes. Aggregated information about the overall achievement concerning curricular targets is more functional than is detailed contextualized information. The *proficiency levels or levels of competence* used to report the results of large-scale assessments (e.g., Adams, 2005; Adams & Wu, 2002) constitute one well-known technique for facilitating the understanding of assessment results among administrators, policy makers, and the public. However, there is little empirical research on which kind of information is actually best suited to guide administrative and political decisions.

In Germany, standardized assessments of students' competencies were previously a relatively rare occurrence, but this is now changing. For example, there is a marked trend towards the use of standardized achievement tests to control admittance to higher education (e.g., Amelang & Funke, 2005; Gold & Souvignier, 2005; Köller, 2004). Several of the newly introduced Bachelors and Masters programs have been designed to communicate specific competencies that are generally verifiable. More importantly, educational standards have been developed to describe the goals of primary and secondary education (Klieme et al., 2003). Based on these standards, a system has been developed to assess students' competencies. New evaluation agencies have been founded as part of these ongoing educational reforms. These agencies assess learning outcomes on both the classroom and the school level, and provide information for policy makers.

DESCRIPTION OF THE DFG PRIORITY PROGRAM ON THE ASSESSMENT OF COMPETENCIES

As outlined above, the accurate empirical assessment of competencies is essential for the enhancement of educational processes and the development of educational systems. Yet devising and implementing such assessments entails numerous theoretical and methodological challenges.

To facilitate this task, the German Research Foundation, DFG, has funded the priority program "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (Klieme & Leutner, 2006). The program, which is scheduled to run for six years, involves a network of currently 20 individual research projects covering different areas of competence assessment.

The program unites experts in different domains of study with cognitive psychologists and experts in educational measurement. Its objective is to develop theoretically and empirically grounded models of competencies as a basis for constructing valid and fair instruments for the assessment of student competencies in terms of both individual learning outcomes (thus promoting individual learning processes) and the output of educational institutions and systems on an aggregated level. Research dealing with the reception of assessment results and their application in pedagogical decisions rounds off the research program. The program extends ongoing research on existing models of competence (e.g., the competence levels used in large-scale assessments) and initiates research in qualitatively new areas (e.g., development of competence models in new content areas; application of innovative psychometric models).

Based on our working definition of competence, the DFG priority program defines competencies as domain-specific cognitive dispositions that are required to successfully cope with certain situations or tasks, and that are acquired by learning processes. Thus, a specific competence is understood as the potential to meet the cognitive demands of a certain domain of learning, or vocational demands. In line with the four areas of research presented above, the program has four specific objectives:

1. To develop cognitive competence models that reflect the contextualized and domain-specific nature of competencies and that enable the theory-based development of instruments for their assessment. These models will focus on cognitive processes, characterize different levels of competence, and describe and explain the quantitative and qualitative development of competence. Ten of the program's projects deal with questions related to this area.

2. To develop and empirically examine appropriate psychometric models on the basis of these theoretical models. The psychometric models will take into account the contextualized character of competencies and to incorporate interindividual differences in underlying abilities, as well as situational demands of performance in complex tasks. Four projects have their main focus in this research area.

3. To develop instruments for the empirical assessment of specific competencies. These instruments permit the empirical examination of the theoretical and psychometric models of competencies, and are essential for basic research on these models. They are also required for basic research that needs measures of competence as outcome variables (e.g., research on the prerequisites of competence development or on educational processes). In applied contexts, these instruments allow individual and institutional learning outcomes to be monitored and provide feedback for learners and educators. Five of the program's projects focus on the development of measurement concepts and instruments.

4. To examine how model-based assessments of competencies are used to inform educational decisions on the individual level, as well as political and administrative decisions concerning educational systems and institutions at the aggregated level. It is crucial to know how different stakeholders in educational processes and decision making understand and use the information provided by empirical assessments of competencies, depending on the underlying theoretical and psychometric models and the measurement methods employed, and how this information impacts the subsequent teaching and learning process. One project addresses these aspects.

Although the program's projects are all assigned to one of the four research areas, most of them are not restricted to a single area. As outlined in this article, research on competency assessment is interdependent, and different areas build on each other. Theoretical models are needed as a starting point; psychometric models translate theoretical models into rules of measurement. Empirical measurement procedures apply the theoretical and psychometric models of competencies and provide data that can be used to inform educational processes and decision making. It is in the nature of the field of research that the majority of the individual projects will initially focus on theoretical and psychometric models. Most projects will then move on to the development of measurement instruments and, in some cases, to research on the reception of the assessment results.

The majority of the projects relate to subjects in primary and secondary education (e.g., mathematics or reading), which is not surprising, given that there has been more research on competencies and competence assessment in these areas and that the corresponding theories are already better developed. However, some projects address competencies in vocational domains; for example, there is a particular focus on well-defined aspects of teachers' professional competence, such as diagnostic competence and specific aspects of pedagogical content knowledge (Weinert, Helmke, & Schrader, 1992). Other projects examine the competencies required in specific non-professional areas of life, such as health or attitudes to environment protection. Altogether the projects cover a wide range of competence domains.

NOTES

[1] German Institute for International Educational Research

[2] Duisburg-Essen University

[3] Reprint of Koeppen, K., Hartig, J., Klieme, E. & Leutner, D. (2008). Current issues in competence modeling and assessment. Zeitschrift für Psychologie/ Journal of Psychology, 216, 61–73 (with permission of Hogrefe & Huber Publishers). The preparation of this paper was supported by grants KL 1057/9-1 and DL 645/11-1 from the German Research Foundation (DFG) in the Priority Program "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293). The paper is an extension of an article by Klieme & Leutner (2006).

REFERENCES

Adams, R.. (2005). *PISA 2003 technical report.* Paris: OECD.

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

Adams, R., Wilson, M., & Wu, M.L. (1997). Multilevel item response modelling: An approach to errors in variables regression, *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Adams, R., & Wu, M.. (2002). *PISA 2000 technical report.* Paris: OECD.

Alderson, C. (2005). Diagnosing foreign language proficiency: the interface between learning and assessment. London: Continuum.

Alderson, C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2005). *The Dutch CEF grid reading/ listening (revised internet version available for test development and analysis).* http://www.lancs.ac.uk/fss/projects/grid/.

Amelang, M., & Funke, J. (2005). Entwicklung und Implementierung eines kombinierten Beratungs- und Auswahlverfahrens für die wichtigsten Studiengänge an der Universität Heidelberg [Development and implementation of a combinded instrument for counseling and selection for the most important courses at the University of Heidelberg]. *Psychologische Rundschau, 56,* 135–137.

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*, 363–383.

Baumert, J., Stanat, P., & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie [Subject, theoretical background and implementation of the study]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 15–68). Opladen: Leske & Budrich.

Beck, B., & Klieme, E.. (2007). *Sprachliche Kompetenzen – Konzepte und Messung* [Language competencies – concepts and measurement]. Weinheim: Beltz.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, H.C. (2004). Mathematische Kompetenz [Mathematical literacy]. In PISA-Konsortium Deutschland (Eds.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (pp. 47–92). Münster: Waxmann.

Bybee, R.W. (1997). Toward an understanding of scientific literacy. In W. Gräber & C. Bolte (Eds.), *Scientific Literacy, an international Symposium* (pp. 37–68). Kiel: IPN.

Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: a reply to Kobayashi, 2002. *Language Testing, 21,* 228–234.

Cheng, L., Watanabe, Y., & Curtis, A.. (2004). *Washback in language testing. research contexts and methods.* Mahwah: Lawrence Erlbaum.

Chung, G.K.W.K., O´Neil, H.F., & Baker, E.L. (2008). Computer-based assessments to support distance learning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe & Huber.

Cizek, G.J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement, Issues and Practice, 20(4)*, 19–28.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). A NCME Instructional Module on Setting Performance Standards: Contemporary Methods. E*ducational Measurement: Issues and Practice, 23*, 31–50.

Connell, M.W., Sheridan, K., & Gardner, H. (2003). On abilities and domains. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 126–155). Cambridge: Cambridge University Press.

Csapó, B. (2004). Knowledge and competencies. In J. Letschert (Ed.), *The integrated person. How curriculum development relates to new competencies* (pp. 35–49). Enschede: CIDREE/SLO.

von Davier, M. (2005). A general diagnostic model applied to language testing data. ETS Research Report 0x-2005.

DiBello, L.V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44,* 285–291.

DiSessa, A. (2006). A history of conceptual change research. In K. Sawyer (Ed.), *The Cambridge Handbook of the learning Sciences* (pp. 265–281). Cambridge: Cambridge University Press.

Dossey, J., Hartig, J., Klieme, E., & Wu, M. (2004). Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003. Paris: OECD Publications.

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive testing. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Wards (Eds.), *Computer-based testing. Building the foundation for future assessments* (pp. 1–35). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S.E. (1983). Construct validity: construct representation vs. nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S.E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61,* 50–55.

ETS (2005). TOEFL iBT at a glance. Retrieved September 25, 2005, from http://www.ets.org/Media/Test/TOEFL/pdf/TOEFL_at_a_Glance.pdf.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Frensch, P. A., Haider, H., Rünger, D., Neugebauer, U., Voigt, S., & Werg, D. (2003). The route from implicit learning to awareness of what has been learned. In L. Jiménez (Ed.), *Attention and implicit learning* (pp. 335–366). New York: John Benjamins Publishing Company.

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A., & Moosbrugger, H. (2007). Usability and internal validity of a modification of the computer game Quake III Arena® for the use in psychological experiments. *Computers in Human Behavior, 23,* 2026–2039.

Fuhrmann, S.H., & Elmore, R.F.. (2004). *Redesigning accountability systems for education.* New York: Teachers College Press.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2007). Using the attribute hierarchy method to make inferences about examinees' cognitive skills. In M.J. Gierl & J.P. Leighton (Eds.), *Cognitive diagnostic assessment for education* (pp. 242–274). Cambridge: Cambridge University Press.

Gogolin, I. (2002). Linguistic and cultural diversity in Europe: a challenge for educational research and practice. *European Educational Research Journal, 1*, 123–138.

Gold, A., & Souvignier, E. (2005) Prognose der Studierfaehigkeit. Ergebnisse aus Laengsschnittanalysen [Prediction of college graduation. Results from longitudinal studies]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37*, 214–222.

Groot, A.S., de Sonneville, M.J., & Stins, J.F. (2004). Familial influences on sustained attention and inhibition in preschoolers. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 45,* 306–314.

Haertel, E.H., & Lorié, W.A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary research and perspectives, 2*, 61–103.

Haider, H., & Frensch, P.A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30*, 304–337.

Haider, H., & Frensch, P.A. (1997). Lernmechanismen des kognitiven Fertigkeitserwerbs [Learning mechanisms in cognitive skill acquisition]. *Zeitschrift für Experimentelle Psychologie, 44*, 521–560.

Haider, H., & Frensch, P.A. (2002). Why individual learning does not follow the power law of practice but aggregated learning does: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 392–406.

Hartig, J., & Frey, A. (2005). *Application of different explanatory item response models for model based proficiency scaling.* Paper presented at the 70th Annual Meeting of the Psychometric Society in Tilburg, July 5–8, 2005.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within- and between-item multidimensionality. *Zeitschrift für Psychologie / Journal of Psychology, 216,* 88–100.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik [Competence and competence diagnosis]. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 127–143). Berlin: Springer.

Hartig, J., & Klieme, E. (2007). *From theoretical notions of competence to adequate psychometric models*. Paper presented at the 12th Biennial EARLI Conference, Budapest, August 28-September 1, 2007.

Hartig, J., Kröhne, U., & Jurecka, A. (2007). Anforderungen an Computer- und Netzwerkbasiertes Assessment [Requirements for computer- and network based assessments]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 57–67). Berlin: Federal Ministry of Education and Research (available at http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Hasselhorn, M., & Grube, D. (2003). The phonological similarity effect on memory span in children: Does it depend on age, speech rate, and articulatory suppression? *International Journal of Behavioral Development*, *27*, 145–152.

Huff, K., & Goodmann, D.P. (2007). The demand for cognitive diagnostic assessment. In M.J. Gierl & J.P. Leighton (Eds.), *Cognitive diagnostic assessment for education* (pp. 19–61). Cambridge: Cambridge University Press.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck, & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Jude, N., & Wirth, J. (2007). Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen [New opportunities of technology based assessment of competencies]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 81–91). Berlin: Federal Ministry of Education and Research (available at URL: http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Jung, B., Ahad, A., & Weber, M. (2005). The Affective Virtual Patient: An e-learning tool for social interaction training within medical field (*Proceeding TESI 2005 – Training Education & Education International Conference).* Kent, UK: Nexus Media (available at http://isnm.de/aahad/Downloads/AVP_TESI.pdf).

Klauer, K.J. (1978). Perspektiven pädagogischer Diagnostik [Perspectives of educational assessment at the individual level]. In K.J. Klauer (Ed.), *Handbuch der Pädagogischen Diagnostik* (pp. 3–4). Düsseldorf: Schwann.

Klauer, K.J. (1987). *Kriteriumsorientierte Tests* [Criterion-referenced tests]. Göttingen: Hogrefe.

Klauer, K.J., & Leutner, D. (2007). *Lehren und Lernen. Einführung in die Instruktionspsychologie* [Teaching and learning. Introduction to instructional psychology]. Weinheim: Beltz-PVU.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. Reiss, K., Riquarts, K., Rost, J., Tenorth, H.E., & Vollmer, J. (2003). *The development of national educational standards. An expertise.* Berlin: Federal Ministry of Education and Research (avaliable at http://www.bmbf.de/pub/the_development_of_national_educationel_standards.pdf).

Klieme, E., Eichler, W., Helmke, A., Lehmann, R. H., Nold, G., Rolff, H.-G., Schröder, K., Thomé, G., & Willenberg, H. (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* [Instruction and competence development in German and English. Results of the DESI study]. Weinheim: Beltz.

Klieme, E., Funke, J., Leutner, D., Reimann, P., & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz? Konzeption und erste Resultate aus einer Schulleistungsstudie [Problem solving as cross-curricular competence? Concepts and first results from an educational assessment]. *Zeitschrift für Pädagogik*, *47,* 179–200.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms bei der DFG [Competence models for assessing individual learning outcomes and evaluating educational processes. Description of a new priority program of the German Research Foundation, DFG]. *Zeitschrift für Pädagogik, 52,* 876–903.

Klieme, E., Maag-Merki, K., & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen [The concept and relevance of competencies in education]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 5–16). Berlin: Federal Ministry of Education and Research (available at http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse [Mathematical literacy: assessment framework and results]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 139–190). Opladen: Leske & Budrich.

Koeller, O. (2004). *Konsequenzen von Leistungsgruppierungen* [Consequences of homogenous groups with regard to school performance]. Muenster: Waxmann.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing 19,* 193–220.

Kröner, S., Plass, J.L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33,* 347–368.

Künsting, J., Thillmann, H., Wirth, J., Fischer, H.E., & Leutner, D. (2008). Strategisches Experimentieren im naturwissenschaftlichen Unterricht [Strategic experimentation in science lessons]. *Psychologie in Erziehung und Unterricht, 55*, 1–15.

Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior, 18*, 685–697.

Leutner, D., & Plass, J. L. (1998). Measuring learning styles with questionnaires versus direct observation of preferential choice behavior: Development of the Visualizer/Verbalizer Behavior Observation Scale (VV-BOS). *Computers in Human Behavior, 14,* 543–557.

Leutner, D., Fleischer, J., Spoden, C., & Wirth, J. (2007). Landesweite Lernstandserhebung zwischen Bildungsmonitoring und Individualdiagnostik [State-wide standardized assessments of learning between educational monitoring and individual diagnostics]. *Zeitschrift für Erziehungswissenschaft, Sonderheft, 8,* 149–167.

van der Linden, W. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42,* 283–302.

van der Linden, W., & Hambleton, R.K. (1996). Item response theory: brief history, common models, and extensions. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item-response theory* (pp. 1–28). Berlin: Springer.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187–212.

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

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99–114.

Nichols, S.L., Glass, G.V., & Berliner, D.C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14* (available at http://epaa.asu.edu/epaa/v14n1/).

Nold, G. (2003). DESI – a language assessment project in Germany and the pros and cons of large-scale testing. *Empirische Pädagogik, 17,* 368–379.

Nold, G., & Rossa, H. (2007a). Hörverstehen [Listening comprehension]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 178–196). Weinheim: Beltz.

Nold, G., & Rossa, H. (2007b). Leseverstehen [Reading comprehension]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (pp. 197–211). Weinheim: Beltz.

Van Den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: a generalized linear and nonlinear approach* (pp. 167–187). New York: Springer.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence – their correlation and their relation: A comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 61–65.

OECD (2007a). *PISA – Programme for International Student Assessment* (available at http://www.oecd.org/dataoecd/51/27/37474503.pdf).

OECD (2007b). *PISA 2006 Science competencies for tomorrow's world* (volume 1: analysis). Paris: OECD.

Ordinate (2004). *SET-10 test description & validation summary*. Menlo Park, CA: Ordinate.

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

Plass, J.L., Chun, D., Mayer, R.E., & Leutner, D. (1998). Supporting visualizer and verbalizer learning preferences in a second language multimedia learning environment. *Journal of Educational Psychology, 90*, 25–36.

Prenzel, M., & Allolio-Näcke, L.. (2006). *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* [Research on educational quality of schools. Final report of the DFG priority program]. Münster: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R.. (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006. Results of the third international study]. Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U.. (2004). *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* [PISA 2003: Educational outcomes of German students – results of the second international study]. Münster: Waxmann.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U.. (2005). *PISA 2003: Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?* [The second comparison of the German laender – What do students know?]. Münster: Waxmann.

Prenzel, M., Rost, J., Senkbeil, M., Häußler, P., & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse [Scientific literacy: assessment framework and results]. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 192–248). Opladen: Leske & Budrich.

Prenzel, M., Schöps, K., Rönnebeck, S., Senkbeil, M., Walter, O., Carstensen, C.H., & Hammann, M. (2007). Naturwissenschaftliche Kompetenzen im internationalen Vergleich [Science competencies in international comparison]. In PISA-Konsortium Deutschland (Eds.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 63–105). Münster: Waxmann.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.

Reeff, J.-P. (2007). Technische Lösungen für ein computer- und internetbasiertes Assessment-System [Technical solutions for computer and internet based assessment systems]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 81–91). Berlin: Federal Ministry of Education and Research (available at http://www.bmbf.de/pub/band_zwanzig_bildungsforschung.pdf).

Rychen, D.S., & Salganik, L.H.. (2001). *Defining and selecting key competencies.* Seattle: Hogrefe & Huber Publishers.

Rychen, D.S., & Salganik, L.H.. (2003). *Key competencies for a successful life and a well-functioning society.* Washington: Hogrefe & Huber Publishers.

Schneider, W., Lockl, K., & Fernandez, O. (2005). Interrelationships among theory of mind, executive control, language development, and working memory in young children: A longitudinal analysis. In W. Schneider, R. Schumann-Hengsteler & B. Sodian (Eds.), *Young children's cognitive development: Interrelationships among executive functioning, working memory, verbal ability, and theory of mind* (pp. 259–284). Mahwah, NJ: Lawrence Erlbaum.

Schnotz, W., Eckhardt, A., Molz, M., Niegemann, H., Hochscheid-Mauel, D., & Hessel, S. (2004). Deconstructing instructional design models towards an integrative conceptual framework for

instructional design research. In H. Niegemann, R. Brünken & D. Leutner (Eds.), *Instructional design and multimedia learning* (pp. 71–89). Münster: Waxmann.

Schnotz, W., Vosniadou, S., & Carretero, M.. (1999). *New perspectives on conceptual change*. Oxford: Elsevier.

Segers, M., Dochy, F., & Cascallar, E.. (2003). *Optimising new modes of assessment: in search of quality and standards*. Dordrecht: Kluwer.

Simonton, K. (2003). Expertise, competence, and creative ability: The perplexing complexities. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 213–239). Cambridge: Cambridge University Press.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31,* 15–21.

Spiel, C., & Glück, J. (in press). A model based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe.

Sternberg, R.J., & Grigorenko, E.. (2003). *The psychology of abilities, competencies, and expertise.* New York: Cambridge University Press.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*, 313–324.

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction, 15,* 317–419.

Walker, C.M., & Beratvas, S.N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255–275.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29,* 126–149.

Weinert, F.E. (1999). *Konzepte der Kompetenz* [Concepts of competence]. Paris: OECD.

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

Weinert, F. E., Helmke, A., & Schrader, F.-W. (1992). Research on the model teacher and the teaching model: Theoretical contradiction or conglutination? In F. Oser, A. Dick & J.L. Patry (Eds.), *Effective and responsible teaching: The new synthesis* (pp. 249–260). San Francisco: Jossey-Bass Publishers.

Weinert, F.E., & Schneider, W.. (1995). *Memory performance and competencies: Issues in growth and development.* Hillsdale, NJ: Lawrence Erlbaum.

Whitely, S.E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement, 18,* 67–84.

Wilhelm, O., & Engle, R.. (2005). *Understanding and measuring intelligence*. London: Sage.

Wilson, M. (2008). Cognitive Diagnosis using Item Response Models. *Zeitschrift für Psychologie/ Journal of Psychology, 216, 73–87.*

Wilson, M., de Boeck, P., & Carstensen, C. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: State of the art and future prospects.* Göttingen: Hogrefe & Huber.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson, (Ed.), *Towards coherence between classroom assessment and accountability (*103rd Yearbook of the National Society for the Study of Education, Part II). Chicago: University of Chicago Press.

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

Wilson, M., & DeBoek, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models. A generalized linear and nonlinear approach* (pp. 43–74). New York: Springer.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts: state of the art and future prospects.* Göttingen: Hogrefe & Huber.

Wirth, J. (2004). *Selbstregulation von Lernprozessen* [Self-regulation in learning processes]. Münster: Waxmann.

Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem-solving competence. *Assessment in Education: Principles, Policy & Practice, 10,* 329–345.

Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence. Implications of theoretical models for assessment methods. *Zeitschrift für Psychologie / Journal of Psychology, 216,* 101–109.

Detlev Leutner
Department of Instructional Psychology,
Duisburg-Essen University, Germany

Karoline Koeppen, Johannes Hartig & Eckhard Klieme
German Institute for International Educational Research