Seifried et al., *Cogent Education* (2020), 7: 1719571
**https://doi.org/10.1080/2331186X.2020.1719571**

cogent education

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

# The computer-based assessment of domain-specific problem-solving competence—A three-step scoring procedure

Jürgen Seifried[1*], Steffen Brandt[2], Kristina Kögler[3] and Andreas Rausch[4]

Received: 16 July 2019
Accepted: 15 January 2020

*Corresponding author: Jürgen Seifried, Economic and Business Education – Professional Teaching and Learning, Universitat Mannheim, L 4,1, Mannheim 68161, Germany.
E-mail: seifried@bwl.uni-mannheim.de

Reviewing editor:
Timo Ehmke, Department of Education, Leuphana Universitat Luneburg, Germany

Additional information is available at the end of the article

**Abstract:** Problem-solving competence is an important requirement in back offices across different industries. Thus, the assessment of problem-solving competence has become an important issue in learning and instruction in vocational and professional contexts. We developed a computer-based low-stakes assessment of problem-solving competence in the domain of controlling. The assessment required participants to process three complex and authentic business-related scenarios within an office simulation. The low number of items and the open-ended response threaten the reliability of the scoring. We collected data from 780 German students in Vocational Education and Training (VET) and applied the following procedure to score the participants' solutions to the three scenarios: A fine-grained scoring resulted in a comprehensive set of response patterns for each scenario. Experts assigned partial credit scores for each of these response patterns and received feedback based on point-biserial correlations of WLE scores for the partial credit items with the total score of the respective subdimension. Finally, a multidimensional model was calibrated based on these partial credits and this model

## ABOUT THE AUTHORS

Jürgen Seifried is a full professor at the Business School at the University of Mannheim. His research interests include competence development of teachers and trainers, learning at the workplace, and competence assessment in vocational education and training (VET).

Steffen Brandt is a psychometrician with long-term experience in educational research, working as freelance consultant for numerous research projects. In his PhD he investigated a novel model for a unidimensional interpretation of multidimensional data.

Kristina Kögler is a full professor of Business Education at the University of Hohenheim. Her research examines antecedents and effects of emotion and motivation in learning and problem solving as well as the assessment of competences in VET.

Andreas Rausch is a full professor at the Business School at the University of Mannheim. His research interests include workplace learning, simulation-based learning in VET, computer-based assessment of competences, and problem solving in office work.

## PUBLIC INTEREST STATEMENT

Due to the increasing automatization and digitization of routine tasks in office work, domain-specific problem-solving competence has become an important goal in the education for office work. However, most competence assessments only measure domain knowledge by multiple-choice items. Authentic scenario-based assessments require the identification of information gaps, researching, evaluating and processing of information, decision-making and the appropriate communication of one's solution. Such performance assessments are supposed to be more valid concerning real-life requirements. However, since these scenarios take more time to process, fewer tasks can be presented. Furthermore, the open item format makes the scoring of solutions more difficult. In order to ensure reliability, we developed a scoring procedure that helps to transform the manifold behavior patterns into partial credit points for each dimension of problem solving. Hence, we demonstrate that complex workplace requirements can be measured validly and reliably.

was further developed. This procedure leads to satisfying EAP/PV reliabilities between .78 and .84 for the subdimensions and .89 for the overarching construct.

**Subjects: Work-based Learning; Assessment; Continuing Professional Development;**

**Keywords: professional competence; problem-solving competence; computer-based assessment; performance assessment; vocational education and training (VET)**

## 1. Introduction

Problem-solving competence has become an important requirement in workplaces across different industries. Automatization and outsourcing of routine tasks do not only affect blue-collar work in production lines but also white-collar work in the back offices of industry (e.g. Brynjolfsson & McAfee, 2014; Frey & Osborne, 2017), and the demands on workplaces have changed in the last decades. Thus, the assessment of problem-solving competence has become an important issue for learning and instruction in vocational and professional contexts. Two approaches to measure problem-solving competence can be distinguished:

(1) Problem solving can be modeled as a cross-curricular ability and assessed on the basis of a number of decontextualized and minimally complex problems as, for instance, applied in the Programme for International Student Assessment (PISA: OECD, 2013, 2014). The approach provides efficient measures for large-scale assessments but has been criticized for applying "simple forms of system behavior that are completely predictable and stable" (Dörner & Funke, 2017, p. 5).

(2) Problem solving can be seen as a complex phenomenon (Schoppek & Fischer, 2015) and assessed on the basis of authentic and complex problems (Gulikers, Bastiaens, & Kirschner, 2004). This approach seems more appropriate for assessing domain-specific problem-solving competence in vocational and professional contexts. However, authentic performance assessments often show poor psychometric qualities (Fischer, Greiff, & Funke, 2017). Usually, it is difficult to meet the requirements of reliability if the test is aimed at assessing complex skills such as problem solving (Kane, 2013).

We developed a computer-based assessment of domain-specific problem-solving competence to measure trainees' professional competences in the field of controlling. In this paper, however, we focus on quantitative aspects and reveal the scoring process for assessing problem-solving competence. First, we give an overview of the theoretical background concerning the assessment of problem-solving competence. Second, we describe our assessment and the scoring approach we used. Although we assume that problem solving comprises cognitive and non-cognitive facets, we omit the non-cognitive aspects here and focus on cognitive aspects of domain-specific problem solving and their corresponding results (for approaches on how to integrate non-cognitive facets of problem solving in the assessment of problem-solving competences and the corresponding results see Rausch, Seifried, Wuttke, Kögler, & Brandt, 2016).[1] Finally, in the discussion and conclusions section we highlight the benefits of providing authentic tasks to measure professional competences.

## 2. Theoretical background

### *2.1. Domain-specific problem-solving competence*

In the literature on problem solving, three aspects are usually taken into account, namely (1) the starting point of the problem, (2) a desired target state, and (3) an approach to move from (1) to (2). A problem exists when a person does not know how to achieve a goal (Duncker, 1945; Newell & Simon, 1972) and therefore experiences a need for action regarding this initial state of not knowing (Rausch et al., 2016; Rausch & Wuttke, 2016). In well-defined problems, the starting position, the goal, and a defined set of operators are given (Dörner, 1987; Mayer & Wittrock, 2006), but real-life problems are often ill-defined and complex. Typical problems in the business domain

are analytic or information problems (relevant information is available or can be derived by deductive reasoning, Brand-Gruwel, Wopereis, & Walraven, 2009) that can be complex in terms of the number and interconnectedness of variables (e.g. in pricing), conflicting goals (e.g. customer satisfaction versus profit maximization), and a lack of transparency (e.g. future pricing strategies of competitors in the relevant market). They often comprise algorithm problems (e.g. cost accounting schemes), diagnostic problems (e.g. identifying calculation errors), decision problems (e.g. supplier selection), and dilemma problems with regard to conflicting goals. We refer to these analytic and complex meta-problems when developing the business problem scenarios for the assessment.

Apparently, whether a situation poses a problem or a task (the latter is also referred to as a "routine problem" by Mayer, 1994, p. 285) is, to a certain point, subjective, as it also depends on an individual's prior experience, knowledge, and skills in the respective domain (Beckmann & Goode, 2017; Dörner, 1987; Funke, Fischer, & Holt, 2018; Mayer, 1994). In the business domain for instance, calculating purchase prices from tender letters and arriving at an informed supplier selection is a routine task for an experienced employee in the purchasing department but may be a complex problem for a novice in the domain. The influence of prior domain-specific experience, knowledge, and skills points to the underlying concept of competence (seen as a latent construct or a bundle of latent constructs).

Competences have to meet complex demands in a particular context (Mulder, 2017; Rychen & Salganik, 2003; Weinert, 2001). A strongly developed competence in a particular domain is referred to as expertise (Evers & van der Heijden, 2017; Van Gog, 2012). Experts do not only have more knowledge but their knowledge is also organized in ways that enable them to quickly recognize chunks of domain-specific information, detect deep features of a problem, and effectively elaborate their initial representation (Anderson, 1993; Chi, Glaser, & Farr, 1988; Nokes, Schunn, & Chi, 2011)—all of which helps them to reduce the perceived complexity of a problem. To achieve such high levels of performance in any domain requires continued experience, which "indicates that problem solving expertise does not come from a superior problem solving ability but rather from domain learning" (Anderson, 1993, p. 40). Similarly, Weinert (2001, p. 53) emphasizes that "cognitive sciences have convincingly demonstrated that content-specific skills and knowledge play a crucial role in solving difficult tasks." Superior key competences cannot compensate for a lack of domain-specific competences. Instead, specific knowledge is required when solving specific problems (Weinert, 2001). In addition, problems in the workplace tend to be 'open' in the sense of Ackerman (2007), in other words, once an individual has mastered a particular situation, he or she can use that knowledge to master even more demanding situations and thus further extend his or her competence.

### *2.2. Assessment of professional competences*

In his literature review, Frederiksen (1984) highlighted the fact that the test format makes a difference when it comes to the measurement of higher-level processes. Multiple choice formats are not suitable if complex cognitive skills (e.g. scientific thinking) should be measured, and correlations between corresponding scores for different formats (multiple choice vs. free-response form) often were found to be very low (Frederiksen & Ward, 1978). This indicates that the measurement of higher vocational competences is subject to specific requirements with regard to task content and task format, and the concept of authenticity is of particular importance (Frey, Schmitt, & Allen, 2012). Authentic assessments require test-takers "to use the same competences, or combinations of knowledge, skills, and attitudes, that they need to apply in the criterion situation in professional life" (Gulikers et al., 2004, p. 69). For the assessment of professional skills, work sample tasks (e.g. roleplays, business simulations, case studies) are designed to achieve high content validity through an authentic assessment with a close relationship to the job task, and the tests require test persons to work on tasks or problems that are similar to the future tasks at the workplace. A prominent example is the in-basket test (Frederiksen, Saunders, & Wand, 1957), a test that aims at measuring skills such as the ability to organize and prioritize work or analytical skills. However, the hopes associated with work sample tests were only partially fulfilled. In their meta-analysis, Roth, Bobko, and McFarland (2005) reported a mean correlation between work

sample tests and measures of job performance of .26. This is lower than the values that were found in meta-analyses in the 1980s (e.g. Hunter & Hunter, 1984). Hence, a careful test construction is required to create a valid and reliable instrument for measuring professional competences. Thereby, especially computer simulations of workplaces are seen as a promising approach to assess professional competences (e.g. Williamson, Bejar, & Mislevy, 2006) and have been used in different domains (e.g. business education: Lohmann et al., 2019; Peng & Abdullah, 2018; medical training: Young et al., 2018; nursing and healthcare:; Padilla, Diallo, & Armstrong, 2018; Ramm, Thomson, & Jackson, 2015; teacher education: Kaufman & Ireland, 2016).

### *2.3. Validity issues in performance assessments*

When it comes to the assessment of complex capabilities, validity issues are of a particular relevance. Many authors (e.g. Borsboom, Mellenbergh, & van Heerden, 2004; Cizek, 2012; Kane, 2006, 2013; Lissitz & Samuelsen, 2007; Markus & Borsboom, 2013; Messick, 1989; Mislevy, 2007, 2009) have given detailed comments and reviews on the understanding of validity in educational assessments. In the 1970s, Cronbach (1971) argued that validity addresses what a test is measuring and what meaning can be drawn from the test scores. So, the issue of validation is not the test itself but rather the conclusions drawn from test scores and the decisions based on these (argument-based approach of validation). In the late 1980s, researchers such as Messick (1989) made a plea for a unified framework of validity (validity as a unitary though faceted concept) (see also Lissitz & Samuelsen, 2007; Mislevy, 2007). Messick (1989, p. 13) defines validity "as an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" and highlighted different sources of validity, namely content validity (from a theoretical and curricular perspective, often based on expert ratings on the relevance and representativeness of test items or observations for a job domain), cognitive validity (based on the theoretical and empirical analysis of test takers' response processes), structural validity (the fit between the scoring model and the structures of the theoretical construct), generalizability (generality of the interpretation over time and across groups and settings, e.g. based on differential item functioning analysis), external validity (the relationship between test scores and relevant criteria variables), and consequential validity (e.g. the consequences of high-stakes testing for individuals and institutions) (Messick, 1995; see also AERA, APA, and NCME Standards for Educational and Psychological Testing, 2014).

Kane (2006, 2013) proposes a strategy for validation by "highlighting key phases or inferences in planning and evaluating the validity argument" (Cook, Byrdges, Ginsburg, & Hatala, 2015, p. 561). Kane (2013, p. 9) outlined the following approach: "First, state the claims that are being made in a proposed interpretation or use (the IUA [interpretation/use argument, the authors]), and second, evaluate these claims (the validity argument).", He provides a holistic framework of validation and differentiates between four steps of arguments/references, namely (1) scoring (from observed performance to an observed score), (2) generalization (from observed score to universe score), (3) extrapolation (from universe score to the level of skill), and (4) decision (from conclusion about the level of skill to a consequence, e.g. placement in a study program). Kane's framework has often been adopted and used by different researchers such as Cook et al. (2015) or Tavares et al. (2018). Furthermore, Cook et al. (2015) provided an overview of how to apply Kane's framework on different types of assessments.

Finally, it is helpful to have a closer look at the framework of Pellegrino, DiBello, and Goldman (2016) with its specific focus on classroom assessment. It builds on the validity concepts that were discussed above and focuses on three aspects of validity, namely (1) cognitive validity ("the extent to which an assessment taps important forms of domain knowledge and skill in ways that are not confounded with other aspects of cognition such as language or working memory load"), (2) instructional validity ("the extent to which an assessment is aligned with curriculum and instruction, including students' opportunities to learn, as well as how it supports teaching practice by providing valuable and timely instruction-related information", and (3) inferential validity ("the extent to which an assessment reliably and accurately yields model-based information about

student's performance, especially for diagnostic purposes") (Pellegrino et al., 2016, p. 4). Pellegrino and colleagues provided their framework with the intention to support classroom teaching and learning. In doing so, they used a computer-based assessment for analyzing student's patterns of correct and incorrect answers to learn more about the conceptual understanding of mathematical functions. We used a similar approach to analyze the domain problem-solving competence in the business domain (see below).

### *2.4. The validity-reliability-dilemma in performance assessments*

In the context of performance assessments, two facets of the validity-reliability-dilemma have to be mentioned. First, Wilson (2004) discussed different assessment modes against the background of educational accountability. He identified two crucial aspects of educational assessments (control over judgment and control over task specification) with regard to validity and reliability. In line with that, Biggs (1996, see also Biggs & Tang, 2011) as well as Trigwell and Prosser (2014) discussed the idea of constructive alignment between the intended learning outcomes (ILO), teaching, and assessment. When the ILO relies on factual knowledge, the assessment focus is on frequent, separate, and memory-based assessment formats, whereas for more complex forms of knowledge (e.g. relations between constructs, justification and evaluation of knowledge, creation of new knowledge) more complex, integrated and flexible formats of assessments are needed. Figure 1 illustrates the trade-off between reliability and validity with regard to item format and the precision of scoring. Thus, it also reflects the current discussion of assessing complex problem solving (see Section 2.1). A combination of complex and authentic performances with a reliable and automated machine scoring would then constitute the "gold standard" of a performance assessment.

Second, the reliability of a test (in the sense of the stability of test results) only increases when a small range of more or less similar (homogeneous) items is provided. In contrast, the validity increases when a test comprises a wide range of different contents and formats. In other words: An increase of reliability goes hand in hand with a decrease of validity. From our perspective, the assessment of domain-specific problem-solving competence should be based on the performance within authentic problem scenarios. However, with regard to test time, test motivation, and test efficiency, only a limited number of authentic problem scenarios can be presented in an assessment. This second facet of the validity-reliability-dilemma refers to the trade-off between the depth and breadth of an assessment with regard to the content and is crucial for the assessment of professional competences. As Kane (2013, p. 30) argues, "the standardization of test format, content, and procedure creates a universe of generalization that is distinct from (and usually narrower than) the target

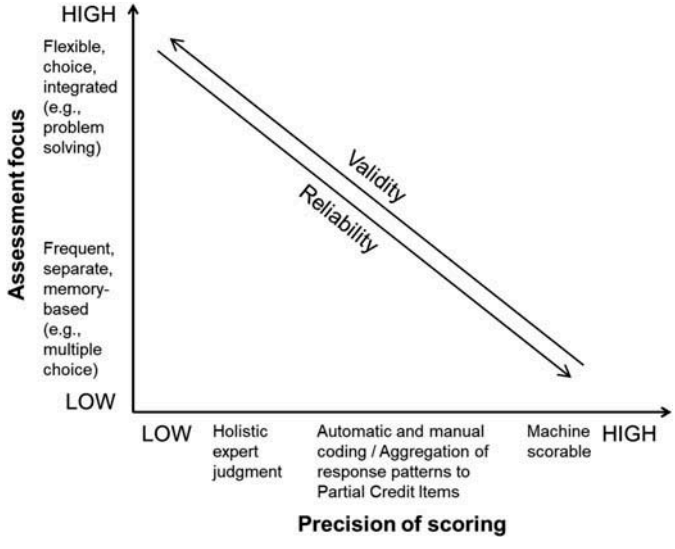


**Figure 1. Assessment modes, validity, and reliability (based on Wilson, 2004; Trigwell & Prosser, 2014).**

domain." However, this loss of authenticity should be kept as small as possible according to Gulikers et al. (2004). Therefore, following a construct-centered approach in the sense of Messick (1994), the test development should be based on a thorough domain analysis (Mislevy & Riconscente, 2006).

### *2.5. Competence framework for the controlling domain*

The assessment that is presented here aims at measuring problem-solving skills in the domain of controlling (i.e. support of managerial decisions, cost planning, cost control, cost accounting). Nowadays, controllers are more than technicians processing transactions (see Brink & Stoel, 2019). From a more strategic perspective, controlling (from an anglophone viewpoint the term management accounting is more suitable; Sheridan, 1995) is seen as a basic management function including a planning, controlling, and reporting function. In this comprehensive understanding, controllers need a deep understanding of entire business operations, best practices, and corporate strategies (Bragg, 2011; Brewer, 2008; Langfield-Smith, Smith, Andon, Hilton, & Thorne, 2018). Globally, in accounting education, there is a trend towards competency-based education with more complex learning and assessment formats (e.g. Abbasi, 2013; Abraham & Jones, 2016).[2]

Our assessment was located in initial vocational education and training in Germany in general and aimed to measure vocational competences of industrial clerks (a three-year training program combining workplace learning and instruction in vocational schools at a non-academic level). We focused on controlling, which is an important part of the curriculum in the apprenticeship, as well as being a relevant domain of business administration in general. At the core of our assessment was the question of how the participants obtained data needed for the assigned management decision, processed and evaluated it, and communicated the results properly, so that they could be used for well-founded decisions.

The cognitive facets of problem-solving competence in the controlling domain were defined by using an action-oriented approach which emanates from the analysis of key processes when working on a controlling problem. Based on a literature review as well as a comprehensive domain analysis and in line with the idea of Evidence-Centered Design (ECD, see Mislevy, 2011; Mislevy, Almond, & Lukas, 2003), four knowledge-related facets of problem-solving competence were identified: (1) Identifying Needs for Action and Information Gaps, (2) Processing Information, (3) Coming to Well-Founded Decisions, and (4) Communicating Decisions Appropriately. Even though these facets can be applied to many domains, we do not argue for a domain-general competence but instead want to emphasize that these competences are highly domain-specific. For instance, being competent in processing information in terms of cost accounting schemes is very different from being competent in processing information when diagnosing an engine failure in a car. With regard to the business domain, different groups of stakeholders confirmed the significance of the proposed competence facets (Rausch & Wuttke, 2016).

### *2.6. Scoring of performance assessments*

The assessment of problem-solving competences in the context of vocational education can be seen as a performance assessment (see Lane & Stone, 2006) which "involves a sample of performance from some domain of performances, with the resulting scores interpreted in terms of typical, or expected, performance in this domain" (Kane, Crooks, & Cohen, 1999, p. 7). However, with regard to scoring issues and the given aim of the assessment to inform stakeholders on a school or program level, it is a crucial question how to provide a scoring solution that is suitable for complex assessments, which (1) implies allowing for a reliable rating of responses, and (2) meets the restrictive psychometric requirements.

*(1) Reliable rating of responses*: In the last decades, a debate started on how scoring performance tasks can be rated automatically to increase the reliability of scoring. In the past, complex performance tasks were usually individually reviewed and rated by experts. Therefore, ways of scoring performance tasks as reliably as multiple-choice tasks were sought (for an overview see Williamson et al., 2006). For Clauser et al. (1995), (1997); Margolis & Clauser, 2006) worked on

automated scoring algorithms for a performance assessment of physicians' patient management skills. They provided computer-based case simulations with free-text entries for test persons and compared the relevance of predictor variables (e. g. the counts of actions for different beneficial and risky actions, time interval in which the test person completed the most important beneficial action) with rule-based automated scorings (based on expert advices). The results show that both approaches are helpful for the approximation of expert judgments, but they found evidence that the regression approach seems to be favorable.

*(2) Psychometric requirements*: Item response theory (IRT) is, in particular, the state-of-the-art method to calibrate achievement data in educational large-scale assessment scenarios. However, some requirements for applying IRT are in conflict with the constraints of assessing complex performances in authentic problem scenarios. In particular, the assumption of local item independence (LID; cf., e.g. Yen, 1993) is crucial in this context. IRT requires that the probability of a correct answer to an item in a test solely depends (except for item individual measurement error) on the construct to be measured. Unwanted effects of LID are—among others—a biased estimation of the difficulty parameters or an overestimation of item discrimination and realibility (Monseur, Baye, Lafontaine, & Quittre, 2011; Tuerlinckx & De Boeck, 2001; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005; Yen, 1984). Considering LID, Bennett, Jenkins, Persky, and Weiss (2003, p. 354) state that "problem solving tasks are particularly susceptible to such effects, which may emanate from chance familiarity with a particular topic, personal interests, fatigue or other sources." The most common source of LID is probably due to a shared stimulus for a set of items and the corresponding familiarity or non-familiarity of respondents with the stimulus, and a scenario-based assessment is therefore in particular susceptible to such LID. An approach to avoid this is to combine the scoring of LID items by, for example, combining different dichotomous items into a single partial credit item. Such a score-based approach has been used and investigated by various researchers (e.g. Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Yen, 1993).

Complex problem-solving performance derives from competences in different facets, resulting in several different competence scores for each participant. From a psychometric point of view, this results in a dual interpretation of a single response data set: (1) a multidimensional interpretation, where each facet loads on a separate dimension, and (2) a unidimensional interpretation, where all facets load on a common dimension. In the context of IRT analyses, however, these interpretations are not as trivial as they might seem since the dimensionality of a given data set is a fundamental assumption in IRT: A model that fits the data set is either unidimensional or multidimensional but not both at the same time. Even more important, the same holds true for the construction of a test: A test can be constructed to be unidimensional (i.e. the item selection is conducted in order to fit a unidimensional model) or multidimensional (i.e. the item selection is conducted in order to fit the given multidimensional model), but not both at the same time. Current large-scale assessments apply two different approaches to deal with this conflict (Brandt, 2015): (a) the calibration of a multidimensional IRT model for the multidimensional achievement scores and the calculation of a weighted mean score for the unidimensional achievement scores (approach used in the National Assessment of Educational Progress) and (b) the calibration of a multidimensional IRT model for the multidimensional achievement scores and the additional calculation of a unidimensional IRT model for the unidimensional achievement scores (approach used in the Program for International Student Assessment and the Trends in Mathematics and Science Study). The first approach uses a multidimensional test development and avoids the possible negative impacts due to LID by constructing separate scales (cf. Yen, 1984) and calculating a weighted mean score based on the calculated scale scores. However, by falling back on classical test theory and leaving the IRT framework, it is not possible to calculate IRT scores for the unidimensional achievements anymore, such as the Weighted Likelihood Estimates (WLE) (Warm, 1989) or Expected a Posteriori (EAP) Estimates (von Davier, Gonzalez, & Mislevy, 2009). The second approach has the advantage of being completely based on IRT, in particular for the score estimation. However, the problem is that it ignores LID due to the multidimensionality of the data set and, thus, there is the possibility of resulting biases for the estimated parameters and the possibility of overestimating the reliability.

In their study on the assessment of complex problem-solving competences, Bennett et al. (2003) use a three-step procedure that includes a scoring based on response patterns instead of sum scores. Their three steps are as follows: (1) obtaining detailed response patterns based on process data and product data, (2) generating partial credit items based on these response patterns, and (3) accumulating evidence. Similarly, Lane and Stone (2006) outline an approach to avoid LID by combining the presumably locally dependent items using scoring rubrics. Our approach presented in the following section will be based on the approach used by Bennett et al. (2003) and the idea of Evidence-Centered Design (ECD, see Mislevy, 2011; Mislevy et al., 2003).

## 3. Development and implementation of an authentic problem-solving assessment

### *3.1. Aim of the assessment*

The aim of the assessment is to measure the problem-solving competence of apprentices in the German dual vocational and educational training (VET) program for industrial clerks. The project for the development of the assessment was part of the German research initiative "Technology-based Assessment of Skills and Competences (ASCOT)". The overarching goal of the ASCOT research initiative was the development and testing of computer-based assessment instruments for "measuring vocational competences which are required for working in high-quality jobs in a changing working world" (Bundesministerium für Bildung und Forschung (BMBF) [German Federal Ministry of Education and Research], 2012). Hence, it is, in particular, a matter of providing stakeholders in vocational education and training with information about the performance level of the trainees. The aim of the assessment is therefore to inform e.g. politicians or enterprises about the current general competence level of trainees, in particular considering the application of their knowledge in realistic situations (Weber & Achtenhagen, 2017). A further aim was also to provide teachers with individualized, formative information about the performance of their students.

The ASCOT initiative focused on selected dual training programs[3] covering different industries (industrial clerks, car mechatronics, electronics technicians, medical assistants, and caring occupations). Our assessment focuses on industrial clerks, as this is one of the most popular training programs in Germany. Certified industrial clerks usually work in back-office departments of industrial or service companies. Further education and professional development can lead to lower or middle management positions. Although routine tasks are still an important part of office work, many of those repetitive processes have been automated or outsourced in recent decades. Thus, employees in back offices of industrial and service companies are increasingly confronted with non-recurrent problems.

Within the broad range of competences of industrial clerks, our assessment focuses on operative controlling, which is one of the most important content domains for industrial clerks in particular and for business administration in general. Operative controlling comprises manifold activities such as support of managerial decisions, cost planning, cost control, and cost accounting. Below, we provide information on the problem scenarios and the office simulation, and information on the data sample of the main study. For further information see Rausch et al. (2016).

### *3.2. Development of authentic problem scenarios*

A valid measurement of domain-specific competence builds on the requirements of a particular domain. To give a valid representation of the knowledge needed in real work scenarios, the construction of the scenarios was based on extensive curriculum, textbook, and workplace analyses as well as interview and diary studies with practitioners. To ensure authenticity, all problem scenarios were embedded in a model company, which is based on a real-life medium-sized German bicycle manufacturer in the premium market sector. Based on real business cases, we developed three complex and authentic problem scenarios, each of which demanded various steps of researching, evaluating and processing information, decision making, and communicating a proposed solution within thirty minutes. The built-in complexity of the scenarios was designed to reflect typical characteristics of complex problems and requirements in the workplace.

Scenario 1 requires a deviation analysis of budget and actual costs. Test takers receive a spreadsheet file of cost accounting and have to calculate budget costs, as well as absolute and relative deviations. Furthermore, they must identify relevant deviations and investigate the diverse reasons of these deviations in a large number of business documents. Finally, test takers have to explain the reasons for cost deviations and propose adjustments for future budgeting in an email to their imaginary supervisor. In Scenario 2, test takers are asked to carry out a supplier selection. They must calculate acquisition prices from offers from different possible suppliers. Besides the quantitative comparison of acquisition prices, they have to judge several qualitative aspects of the suppliers, which are presented in a variety of documents of varying credibility. All derived information must be integrated in a value analysis, and, finally, they have to send a well-founded decision proposal to their supervisor via email. Scenario 3 concerns a make-or-buy decision with regard to a new product in the product line-up of the company. Besides the quantitative comparison of the total costs of each alternative (including customs duty, machine hour assessment costs, personnel costs, etc.), test takers have to balance the chances and risks of each decision (e.g. delivery reliability, load factor of the machine park, environmental aspects). Again, they have to communicate their proposed decision to their supervisor.

The problem scenarios can be classified as authentic and complex analytical meta-problems (Dörner, 1997; Dörner & Funke, 2017; Funke, 2003; Gulikers et al., 2004; Jonassen, 2000; Leutner, Funke, Klieme, & Wirth, 2005). Although all scenarios fall into the scope of operative controlling, the scenarios cover a wide field of domain-specific knowledge.

### *3.3. Development of an office simulation*

The assessment was administered in a computer-based office simulation that enabled a holistic processing of the problem scenarios without any artificial fragmentation. The simulation provided typical features such as a file system, an email client, a spreadsheet software, a notepad, and a calculator. It also offered a large variety of archives with information commonly found in enterprises (e.g. invoices, letters, bids, notes) and a comprehensive archive containing short explanations of relevant and irrelevant technical terms. In total, the test takers had access to more than 150 documents (of which 17 % were technical product information, e.g. how to assemble a certain bike type; 17 % general information on the company, e.g. on its history or the current earnings; 37 % lookup information in a Wiki, e.g. to have a look at particular formulas or definitions; and 30% supplier and invoice documents). By providing such a vast amount of information, we reduced the need to know every detail by heart. However, none of the documents contained a complete solution to the problem. Altogether, as within real-life problem solving, the participants could look up information but could also be distracted or overwhelmed by documents that provided irrelevant, conflicting, or misleading information. Each problem scenario started with an email from a supervisor who assigned a problem and requested an email response within 30 minutes.

## 4. Method

### *4.1. Sample*

The study took place between April and September 2014. A total of $N$ = 786 vocational students participated in the test. They came from 18 different vocational schools, which were located in seven of the 16 German federal states. The participation was voluntary, both on the school level as well as on the individual student level. Six students were excluded due to missing data, resulting in 780 (50.1 % female) participants that were included in the analyses. All students were in the second or third year of a three-year commercial apprenticeship program and showed a typical right skewed age distribution ($M$ = 21.3 years; $SD$ = 2.69; $Min$ = 17; $Max$ = 44). 537 of the students who took part in an apprenticeship program were apprentices to become industrial clerks, 106 were apprentices to become IT-system management assistants, and another 137 were apprentices to become merchants in wholesale and foreign trade. The main focus group for the study was the group of industrial clerks, the two remaining groups were included for validation purposes due to the respective differences in the curricula of their apprenticeship programs. Taking curricular validity into account, it was expected that the industrial clerks outperform both other

groups. Rausch et al. (2016), showed this. All results reported in the following refer to the full sample of 780 students.

### *4.2. Scoring procedure*

For scoring issues, we used a three-step approach inspired by Bennett et al. (2003) and combined a qualitative analysis of test persons' response patterns with a quantitative analysis for the scoring of the assessment. The coding approach comprised the following steps: (1) a fine-grained coding that led to response patterns, (2) an assignment of each response pattern to a partial credit score based on expert ratings, and (3) an IRT calibration using a partial credit model.

First, coding guidelines were developed for each of the three scenarios. These coding guidelines were used to obtain meaningful variables from the students' open-ended email answers (manual coding), the included spreadsheet application (manual and automated coding), and the logging data of information-seeking activities (automated coding). Each variable (in the following referred to as "item") was assigned to exactly one of the four above-defined facets of problem-solving competence. The manual coding was done by nine raters. The raters were trained by using a training sample of 50 test persons. Each rater was assigned to a subset of items he or she specializes in and therefore coded the answers of a large number of students. In doing so, possible biases were reduced since only a subset of items was coded by the same rater and multiple raters were involved in coding a single student's answer in every case. Cohen's kappa (Cohen, 1968) was calculated for each rater. Since the values of the first attempt were not consistently satisfactory, a second coder rating was necessary. After that, the values were at a satisfactory level. Considering the automated coding, the correctness of the codings was proven by an additional verification sample, comparing the automated machine codings with human codings. In some cases, the automated coding had to be readjusted, but all in all the automated coding worked satisfactorily. In total, between 12 and 41 items were coded for each competence facet representing a sub-dimension in the IRT model. Thus, we used the term subdimension as the equivalent of competence facets within the IRT framework.

In the second step, each combination of level-one items of a particular subdimension and within a particular scenario was considered a unique response pattern. These identified response patterns for each scenario and each subdimension were then summarized and presorted according to the respective sum score. Thereafter, an expert group assigned credit points to the different response patterns. Table 1 shows an example of such an assignation table used by the experts for scenario 1 (deviation analyses of budget and actual costs) and subdimension 1 (Identifying Needs for Action). The sum score was merely used as an orientation for the experts. As Table 1 shows, there were cases where the sum scores are equal but the experts assign differing partial credits and cases where the sum scores are unequal but the assigned partial credits are equal. When condensing the response patterns of various variables into partial credit scores, two main approaches were identified: First, following a hierarchical-sequential approach, some particular level-one items were defined as a sine qua non. To achieve the next higher score of the level-two item, it was essential to have a certain score (e.g. code '1') in the respective level-one item; for instance, a very basic calculation had to be done correctly. Otherwise, the next partial credit score cannot be awarded, irrespective of the scores in the other level-one items of the respective subdimension. Second, following a compensation approach, low scores in one level-one item can be compensated by higher scores in another level-one item. For instance, a particular partial credit score was awarded when at least one out of three calculations is done correctly, irrespective of which particular calculation that is. For most competence facets, both approaches as well as combinations of the approaches were implemented across the various level-two items.

In a third step, unidimensional analyses for each of the subdimensions were conducted providing the mean WLE scores for each of the response patterns as well as point-biserial correlations of the respective partial credit item with the total score in the subdimension. This approach corresponded to the item selection and item definition process in large scale assessments, such as PISA (cf. OECD, 2014), and aimed at helping the content experts in identifying items with potential

**Table 1. Table for the assignation of partial credit scores for the subdimension "Identifying Needs for Action (A1)" to answer patterns of scenario 1 (S1)**

| Items | | | | | | | | | Sum Score | Number of Students With This Pattern | Assigned Partial Credit Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A1S1I1a | A1S1I1b | A1S1I1c | A1S1I1e | A1S1I1f | A1S1I2a | A1S1I2b | A1S1I2c | A1S1I2d | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 57 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 29 | 1 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 38 | 1 |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 11 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 26 | 1 |
| 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 4 | 12 | 2 |
| 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 14 | 1 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 27 | 2 |
| 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 4 | 91 | 2 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 5 | 34 | 2 |
| 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 5 | 123 | 3 |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 6 | 12 | 3 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 6 | 24 | 3 |

Level-one items A1S1I1a to A1S1I2d are derived from the fine-grained codings.

weaknesses. This assignment of the level-one items to partial credit scores were denoted as level-two coding and the results, correspondingly, as level-two items.

After completion of the coding process, three partial credit items (one for each scenario) with differing numbers of scoring categories were obtained for each subdimension of problem-solving competence. The Generalized Subdimension Model (GSM), which provides the weighted mean scores via a restriction of the multidimensional Rasch model, was used here (Brandt, 2012, 2015). When calculating a unidimensional score via weighted mean scores, the question of the significance of the different subdimensions arises (cf. Herl et al., 1999). A conducted survey study with different stakeholders in the area of VET confirmed that all suggested subdimensions are considered equally important for solving typical problems at work (Rausch & Wuttke, 2016). We therefore used an equal weighting of the subdimensions, which is also supported by Baldwin (2015), who suggests a weighting of the components of a composite score based on expert judgments. The calibration of the GSM as well as of the multidimensional Rasch model for the achievement scores in the subdimensions were conducted using the package TAM in R (Kiefer, Robitzsch, & Wu, 2015). Figure 2 outlines the iterative three-step procedure from fine-grained response patterns over partial credits to IRT calibration.

### *4.3. Findings*

Table 2 shows the average scores achieved on each item in the assessment and thereby provides a first impression of the difficulty of the different items as well as of the difficulties of the different scenarios and facets. While the facet "Communicating Decisions Appropriately" seems to be the easiest, the "Coming to Well-Founded Decisions" facet seems to be the most difficult. Further, Scenario 2 has shown to be the easiest while Scenario 3 was the hardest in most of the facets. For Scenario 1, the "Coming to Well-Founded Decisions" facet was very hard while the "Processing Information" facet was rather easy.

Table 3 shows the reliabilities obtained for each subdimension. Besides Cronbach's Alpha, reliabilities of the Expected a Posteriori (EAP) and Plausible Value (PV) scores are provided (Mislevy, Beaton,

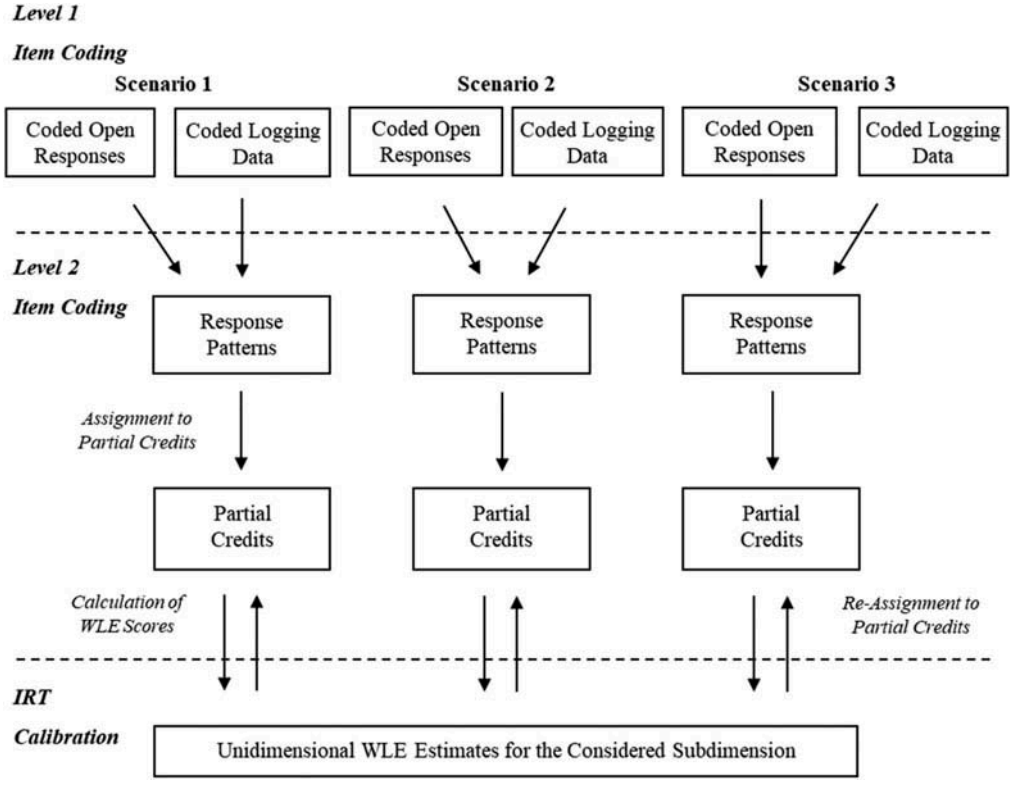


**Figure 2. Visualization of the three-step procedure from problem-solving performance to problem-solving competence.**

**Table 2. Average scores achieved for each of the constructed partial credit items (Level 2 Items) and maximum scores achievable per item**

| Subdimension | Scenario 1 Level 2 Item | Scenario 2 Level 2 Item | Scenario 3 Level 2 Item |
|---|---|---|---|
| Identifying Needs for Action | 1.7 (4) | 2.2 (4) | 1.3 (4) |
| Processing Information | 2.8 (6) | 1.4 (4) | 0.7 (5) |
| Coming to Well-Founded Decisions | 0.8 (6) | 1.7 (4) | 1.4 (5) |
| Communicating Decisions Appropriately | 1.9 (3) | 2.1 (3) | 1.9 (3) |

The maximum score for each item is given in parentheses. A partial credit item with a maximum score of 4 has five scoring categories, with the scores ranging from 0 to 4 points.

**Table 3. Obtained test reliabilities**

| (Sub-) Dimension | Cronbach's Alpha | EAP/PV-Reliability Separate unidimensional models | EAP/PV-Reliability Multidimensional model | EAP/PV-Reliability Multidimensional model with regression data |
|---|---|---|---|---|
| Identifying Needs for Action | .57 | .56 | .78 | .84 |
| Processing Information | .54 | .54 | .74 | .79 |
| Coming to Well-Founded Decisions | .40 | .38 | .65 | .82 |
| Communicating Decisions Appropriately | .54 | .49 | .64 | .78 |
| Problem Solving | .75 | — | .78 | .89 |

Kaplan, & Sheehan, 1992).[4] Both score estimation methods take advantage of the additional correlational information provided by other dimensions in a multidimensional model or by added regression data (often also referred to as a regression model or background model). This additional information is based on a background questionnaire that the students completed. By using this information, they can provide score estimates with less measurement error, i.e. higher reliability. While EAPs are point estimates, assigning each person to a single score, PVs (typically five to ten PVs for each person) represent a score distribution for each person and are particularly useful for the estimation of group mean scores (Von Davier et al., 2009). The multidimensional calibrations showing EAP/PV reliabilities for the four facets as well as for the overarching general problem-solving competence (third and fourth column) are based on the GSM and show how the added information leads to an increase in the estimated reliability in comparison to the estimation of four separate unidimensional models without additional information (second column), which shows a much closer relationship to the Cronbach's Alpha included in the first column. Finally, Table 4 displays the latent correlations between the four subdimensions, which range from .15 to .64. More general findings with regard to the results of the developed assessment have been reported by Rausch et al. (2016).

## 5. Discussion and conclusions

### *5.1. Discussion*

The presented performance assessment aims at informing different stakeholders about students' problem-solving competence in the domain of operative controlling, which is a major curricular

**Table 4. Latent correlations between the subdimensions**

| Subdimension | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (1) Identifying Needs for Action | — | | | |
| (2) Processing Information | .64 | — | | |
| (3) Coming to Well-Founded Decisions | .61 | .31 | — | |
| (4) Communicating Decisions Appropriately | .43 | .15 | .51 | — |

content in business administration and commercial vocational education and training in Germany. We defined four facets of problem-solving competence: (1) identifying needs for action and information gaps, (2) processing information, (3) coming to well-founded decisions, and (4) communicating decisions appropriately. These competence facets are to be assessed on the basis of domain-specific performances which should correspond to real-life situations in the workplace as far as possible. Thus, we developed authentic and complex problem scenarios in the domain of operative controlling based on a thorough domain analysis in order to attain validity. Furthermore, we developed an authentic office simulation that included typical domain-specific tools and, thus, allowed for an open-ended constructed response (Rausch et al., 2016).

Working on complex scenarios and allowing for very individual responses poses several threats to the psychometric qualities of an assessment. A common validity-reliability-dilemma arises because complex responses usually lead to a reduced precision of scoring. We implemented a scoring approach which was based on a combination of manual and automatic codings. We broke down the complexity of the scoring into concise codings of the level-one items and condensed the resulting response patterns to level-two items. Hence, deviations in the partial credit scores on level two due to different scores on level one are much less probable. The applied scoring procedure appears as laborious as experts also had to assign each of the response patterns to a certain partial credit score, which for some subdimensions and scenarios resulted in large amounts of patterns that had to be assigned. However, the involvement of the experts further supports validity, as they were aware of an item's significance and were able to implement different strategies to summarize the level-one items to level-two partial credit items. Moreover, once each response pattern was assigned to a partial credit score, this step could be fully automated, that is if the response pattern occurred again, it could be scored without any further involvement of the experts. Since we analyzed the response patterns of almost 800 participants, future applications of our assessment will probably require no expert involvement.

Considered as separate unidimensional constructs, the obtained EAP/PV reliabilities estimated for the four subdimensions ranged between .38 and .56, which would be very low for a typical achievement test with ten items or more. In the given case, however, with just three items, the results were influenced by the content-specific effects of each scenario. It was therefore not plausible to assume that a calibration of the students' achievements with a reliability of, for example, .80 is possible if it is just based on the information of these three items. In order to increase the reliability, we therefore calibrated the data using a multidimensional IRT model, and finally used a multidimensional model with regression data from a background questionnaire that the students completed additionally. In doing so, we were able to increase the EAP/PV reliabilities to values which then ranged between .78 and .84 for the subdimensions and to an EAP/PV reliability of .89 for the overarching construct (domain-specific problem-solving competence), which are typically considered to be sufficient.

The presented study has some further limitations. We only addressed an isolated part of professional problem-solving competence in the business domain. Although the four facets of domain-specific problem-solving competence can be easily adapted to other domains, further research is needed to uncover possible shortcomings when applying the model to new contexts. Finally and most important, the current degree of automated codings should be developed further in order to reduce the effort involved in human coding. This would also increase the opportunities for disseminating these findings into research and practice.

### *5.2. Conclusions*

Based on the experience from the development of our assessment, we draw the following main conclusions.

(1) The computer-based assessment of domain-specific problem-solving competence in the field of commercial vocational education was focused on a broad understanding of validity with regard to the competence that is required in the real working life. Therefore, authentic problem scenarios on the basis of extensive domain analyses were implemented. We also provided an open-ended problem space for working on these problems within an authentic office environment instead of applying highly structured items in terms of multiple choice. Expanding the problem space for the test-takers (i.e. reducing experimental control) resulted in heterogeneous behaviors and solution patterns and complicated the application of item response theory (IRT). Nevertheless, statistical tests and indices based on IRT demonstrate the reliability of the measurement of cognitive competence facets.

(2) In recent years, the emphasis on psychometric qualities has led to a loss of variety concerning assessment designs (Dörner & Funke, 2017; Funke, 2014). In our assessment—despite its mentioned limitations—the authentic open-ended problem space is a distinctive feature. Many assessments in vocational and professional contexts do present authentic problems but then limit the response formats to multiple choice. Indeed, "few students end up with jobs where they get paid to fill out multiple-choice test bubble sheets" (Frey et al., 2012, p. 12). Allowing for complex open-ended responses is a strength with regard to validity (or authenticity) but it is also the biggest threat to test efficiency because the scoring of the answers usually requires human coders. But we are convinced it is worth the effort, particularly as many scoring steps can potentially be conducted more efficiently in the future. For instance, considering the initial assignment of answer patterns to partial credit scores, this might, in the future, be supported via a scoring tool that provides all necessary information in one place and supports the above described strategies to improve the scoring of the response patterns. Considering the coding efficiency itself, for assessments with existing coding and scoring instructions, we have already managed to code many of the low-level items automatically on the basis of logging data and also have identified many possibilities for enlarging the automated scoring in subsequent developments, in particular by using machine learning techniques (see, e.g. Mao et al., 2018; Riordan, Horbach, Cahill, Zesch, & Lee, 2017).

In summary, it can be said that with the chosen approach it is possible to grasp the complex problem-solving ability. The assessment of domain-specific problem-solving competence by means of complex and authentic tasks proves to be challenging, but feasible. In the future, however, the procedures will have to be further systematized and automated.

**Funding**

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the research initiative "Technology-based Assessment of Skills and competences (ASCOT)" under Grant number [01DB1119-23]. The publication of this article was funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the University of Mannheim.

**Author details**

Jürgen Seifried[1]
E-mail: seifried@bwl.uni-mannheim.de
ORCID ID: http://orcid.org/0000-0002-9460-7721
Steffen Brandt[2]
E-mail: steffen@opencampus.sh
ORCID ID: http://orcid.org/0000-0001-5410-7186
Kristina Kögler[3]

E-mail: kristina.koegler@uni-hohenheim.de
ORCID ID: http://orcid.org/0000-0002-6766-1685
Andreas Rausch[4]
E-mail: rausch@uni-mannheim.de
ORCID ID: http://orcid.org/0000-0002-0749-2496
[1] Economic and Business Education – Professional Teaching and Learning, University of Mannheim, Business School, L 4,1, 68161, Mannheim, Germany.
[2] Art of Reduction, Wissenschaftszentrum Kiel, Fraunhoferstraße 13, 24118, Kiel, Germany.
[3] Economic and Business Education, Universität Hohenheim, Fruwirthstr. 47, 70593, Stuttgart, Germany.
[4] Economic and Business Education – Workplace Learning, University of Mannheim, Business School, L 4,1, 68161, Mannheim, Germany.

**Citation information**
Cite this article as: The computer-based assessment of domain-specific problem-solving competence—A three-step scoring procedure, Jürgen Seifried, Steffen Brandt, Kristina Kögler & Andreas Rausch, *Cogent Education* (2020), 7: 1719571.

**Notes**

1. Although broad definitions of competence emphasize the interplay of cognitive as well as non-cognitive prerequisites (e.g. Weinert, 2001), research on problem-solving competence usually focuses on cognitive prerequisites (for a critique, see Sembill, Rausch, & Kögler, 2013), particularly the acquisition and application of knowledge in problem situations (Fischer et al., 2017).
2. In the US, the American Institute of Certified Public Accountants (AICPA) provides a competency framework of the skills tested on the Certified Public Accountant Exam. They advocate a skills-based curriculum and arrange the competences under the following three categories: (1) functional competences (e.g. research relevant literature; analyze and interpret business information; communicate business information); (2) personal competences (render judgment based on available business information as well as individual attributes and values); and (3) broad business perspective competences (e.g. understanding of key business terms, facts, and processes) (Boritz & Carnaghan, 2017).
3. The German dual system of VET comprises about 330 state-recognized apprenticeship programs. It is characterized by a combination of workplace learning in the training company and classroom-based learning in state-run vocational schools and usually takes three years.
4. EAP and PV estimates take advantage of the covariances of the latent variables to be measured with other observed characteristics, such as, for example, the achieved scores of the other latent dimensions in a multidimensional model. By doing so, it allows to provide estimates with reduced measurement error (Embretson & Reise, 2000).

**Data availability**
The data set collected and used for this study is available under
https://www.iqb.hu-berlin.de/fdz/studies/01DB1119_DomPL-IK.

**References**

Abbasi, N. (2013). Competency approach to accounting education: A global view. *Journal of Finance and Accountancy*, *13*. Retrieved from http://www.aabri.com/manuscripts/131566.pdf

Abraham, A., & Jones, H. (2016). Facilitating student learning in accounting through scaffolded assessment. *Issues in Accounting Education*, *31*(1), 29–49. doi:10.2308/iace-51320

Ackerman, P. L. (2007). New developments in understanding skilled performance. *Current Directions in Psychological Science*, *16*, 235–239. doi:10.1111/j.1467-8721.2007.00511.x

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*, 35–44. doi:10.1037/0003-066X.48.1.35

Baldwin, P. (2015). Weighting components of a composite score using naïve expert judgments about their relative importance. *Applied Psychological Measurement*, *39*, 1–12. doi:10.1177/0146621615584703

Beckmann, J. F., & Goode, N. (2017). Missing the wood for the wrong trees: On the difficulty of defining the complexity of complex problem solving scenarios. *Journal of Intelligence*, *5*, 15. doi:10.3390/jintelligence5020015

Bennett, R. E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem solving performances. *Assessment in Education: Principles, Policy & Practice*, *10*, 347–359.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347–364. doi:10.1007/BF00138871

Biggs, J. B., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Maidenhead: McGraw Hill Education & Open University Press.

Boritz, J. E., & Carnaghan, C. (2017). Competence-based education and assessment in the accounting profession in Canada and the United States. In M. Mulder (Ed.), *Competence-based vocational and professional education bridging the world of work and education* (pp. 273–296). Cham: Springer International.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061

Bragg, S. M. (2011). *The controller's function: The work of the managerial accountant* (4th ed.). Hoboken, NJ: Wiley.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education*, *53*, 1207–1217. doi:10.1016/j.compedu.2009.06.004

Brandt, S. (2012). *Definition and classification of a generalized subdimension model. Presented at the 2012 annual conference of the National Council on Measurement in Education (NCME)*, Vancouver, BC. doi:10.1094/PDIS-11-11-0999-PDN

Brandt, S. (2015). *Unidimensional interpretation of multidimensional tests* (Doctoral dissertation). Kiel: Christian-Albrechts-Universität zu Kiel. Retrieved from http://macau.uni-kiel.de/receive/dissertation_diss_00018271

Brewer, P. (2008). Redefining management accounting. *Strategic Finance*, *89*, 27–34.

Brink, W. D., & Stoel, M. D. (2019). Analytics knowledge, skills, and abilities for accounting graduates. In T. G. Calderon (Ed.), *Advances in accounting education: Teaching and curriculum innovations* (Vol. 22, pp. 23–43). Bingley: Emerald.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York, London: Norton.

Bundesministerium für Bildung und Forschung (BMBF) [German Federal Ministry of Education and Research]. (2012). *Vocational skills and competences made visible: The ASCOT research initiative*. Bonn: German Federal Ministry of Education and Research.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. London: Lawrence Erlbaum Associates.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*, 31–43. doi:10.1037/a0026975

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement*, *34*(2), 141–161. doi:10.1111/jedm.1997.34.issue-2

Clauser, B. E., Subhiyah, R. G., Nungester, R. J., Ripkey, D. R., Clyman, S. G., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement*, *32*, 397–415. doi:10.1111/jedm.1995.32.issue-4

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220. doi:10.1037/h0026256

Cook, D. A., Byrdges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*, 560–575. doi:10.1111/medu.12678

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Dörner, D. (1987). Denken und Wollen: Ein systemtheoretischer Ansatz [Cognition and volition: A system-theoretical approach]. In H. Heckhausen, P. M. Gollwitzer, & F. E. Weinert (Eds.), *Jenseits des Rubikon* [Beyond the rubicon] (pp. 238–250). Berlin: Springer.

Dörner, D. (1997). *The logic of failure. Recognizing and avoiding error in complex situations*. New York: Merloyd Lawrence.

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology*, *8*, 1-11.doi:10.3389/fpsyg.2017.01153

Duncker, K. (1945). *On problem solving*. Washington: The American Psychological Association.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Evers, A. T., & van der Heijden, B. I. J. M. (2017). Competence and professional expertise. In M. Mulder (Ed.), *Competence-based vocational and professional education bridging the world of work and education* (pp. 83–101). Cham: Springer International.

Fischer, A., Greiff, S., & Funke, J. (2017). The history of complex problem solving. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: Using research to inspire 21st century learning* (pp. 107–121). Paris: OECD Publishing.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*(3), 193–202. doi:10.1037/0003-066X.39.3.193

Frederiksen, N., Saunders, D. R., & Wand, B. (1957). The in-basket test. *Psychological Monographs*, *71*(9), 1–28. doi:10.1037/h0093706

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem solving. *Applied Psychological Measurement*, *2*, 1–24. doi:10.1177/014662167800200101

Frey, B., Schmitt, V. L., & Allen, J. P. (2012). Defining authentic classroom assessment. *Practical Assessment, Research & Evaluation*, *17*, 1–18.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. doi:10.1016/j.techfore.2016.08.019

Funke, J. (2003). *Problemlösendes Denken*. [Problem solving thinking]. Stuttgart: Kohlhammer.

Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Frontiers of Psychology*, *5*, 739. doi:10.3389/fpsyg.2014.00739

Funke, J., Fischer, A., & Holt, D. V. (2018). Competences for complexity: Problem solving in the twenty-first century. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills: Research and applications* (pp. 41–53). Dordrecht: Springer.

Gulikers, J., Bastiaens, T., & Kirschner, P. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–85. doi:10.1007/BF02504676

Herl, H. E., O'Neil, H. F., Chung, G. K. W. K., Bianchi, C., Wang, S., Mayer, R., … Tu, A. (1999). *Final report for validation of problem solving measures*. Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*(1), 72–98. doi:10.1037/0033-2909.96.1.72

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, *48*, 63–85. doi:10.1007/BF02300500

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*, 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.2013.50.issue-1

Kaufman, D., & Ireland, A. (2016). Enhancing teacher education with simulations. *TechTrends*, *60*(3), 260–267. doi:10.1007/s11528-016-0049-0

Kiefer, T., Robitzsch, A., & Wu, M. (2015). *TAM: Test analysis modules* (Version 1.3) [R]. Retrieved from http://cran.r-project.org/package=TAM

Lane, S., & Stone, C. A. (2006). Performance assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: Praeger Publishers Inc.

Langfield-Smith, K., Smith, D. A., Andon, P., Hilton, R., & Thorne, H. (2018). *Management accounting: Information for creating and managing value* (8th ed.). Sydney: McGraw-Hill Education.

Leutner, D., Funke, J., Klieme, E., & Wirth, J. (2005). Problemlösefähigkeit als fächerübergreifende Kompetenz [Problem solving competence as cross-curricular competence]. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern* [Students' problem solving

competence] (pp. 11–19). Wiesbaden: VS Verlag für Sozialwissenschaften.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437–448. doi:10.3102/0013189X07311286
Lohmann, G., Pratt, M. A., Benckendorf, P., Strickland, P., Reynolds, P., & Whitelaw, P. A. (2019). Online business simulations: Authentic teamwork, learning outcomes, and satisfaction. *Higher Education*, 77, 455-472. doi:10.1007/s10734-018-0282-x.
Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*, 121–138. doi:10.1080/10627197.2018.1427570
Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring for complex tasks in computer-based testing* (pp. 123–167). Hillsdale, NJ: Lawrence Erlbaum Associates.
Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.
Mayer, R. E. (1994). Problem solving. In M. W. Eysenck (Ed.), *The Blackwell dictionary of cognitive psychology* (pp. 284–288). Oxford: Blackwell.
Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). New York: Routledge.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, *23*(2), 13–23. doi:10.3102/0013189X023002013
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741
Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.
Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463–469. doi:10.3102/0013189X07311660
Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions new directions and applications* (pp. 83–108). Charlotte, NC: Information Age.
Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment* (CRESST Report 800). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centred design (RR-03-16)*. Princeton, NJ: Educational Testing Service.
Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161. doi:10.1111/jedm.1992.29.issue-2
Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph Series–Issues and Methodologies in Large Scale Assessments*, *4*, 131–158.
Mulder, M. (2017). Competence and the alignment of education and work. In M. Mulder (Ed.), *Competence-based vocational and professional education bridging the world of work and education* (pp. 229–251). Cham: Springer International.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
Nokes, T. J., Schunn, C. D., & Chi, M. T. H. (2011). Problem solving and human expertise. In V. Grøver Aukrust (Ed.), *Learning and cognition in education* (pp. 104–111). Oxford: Elsevier.
OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: Author.
OECD (2014). *PISA 2012 technical report*. Paris: Author.
Padilla, J., Diallo, S. Y., & Armstrong, R. K. (2018). Toward live virtual constructive simulations in healthcare learning. *Simulation in Healthcare: the Journal of the Society for Simulation in Healthcare*, *13*, 35–40. doi:10.1097/SIH.0000000000000317
Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59–81. doi:10.1080/00461520.2016.1145550
Peng, J., & Abdullah, I. (2018). Building a market simulation to teach business process analysis: Effects of realism on engaged learning. *Accounting Education*, *27*, 208–222. doi:10.1080/09639284.2017.1407248
Ramm, D., Thomson, A., & Jackson, A. (2015). Learning clinical skills in the simulation suite: The lived experiences of student nurses involved in peer teaching and peer assessment. *Nurse Education Today*, *35*, 823–827. doi:10.1016/j.nedt.2015.01.023
Rausch, A., Seifried, J., Wuttke, E., Kögler, K., & Brandt, S. (2016). *Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain*. Empirical Research in Vocational Education and Training. Retrieved from https://link.springer.com/article/10.1186/s40461-016-0035-y
Rausch, A., & Wuttke, E. (2016). Development of a multi-faceted model of domain-specific problem-solving competence and its acceptance by different stakeholders in the business domain. *Unterrichtswissenschaft*, *44*, 169–184.
Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. M. (2017). Investigating neural architectures for short answer scoring. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 159–168). Copenhagen, Denkmark: The Association for Computational Linguistics.
Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, *58*, 1009–1037. doi:10.1111/peps.2005.58.issue-4
Rychen, D. S., & Salganik, L. H. (2003). A holistic model of competence. In D. S. Rychen & L. H. Salganik (Eds.), *Key competences for a successful life and well-functioning society* (pp. 41–62). Seattle: Hogrefe & Huber.
Schoppek, W., & Fischer, A. (2015). Complex problem solving—single ability or complex phenomenon?. *Frontiers in Psychology*, *6*, 1-4. doi:10.3389/fpsyg.2015.01669

Sembill, D., Rausch, A., & Kögler, K. (2013). Non-cognitive facets of competence. Theoretical foundations and implications of measurement. In K. Beck & O. Zlatkin-Troitschanskaia (Eds.), *From diagnostics to learning success. Proceedings in vocational education and training* (pp. 199–212). Rotterdam: Sense.

Sheridan, T. T. (1995). Management accounting in global european corporations: Anglophone and continental viewpoints. *Management Accounting Research*, *6*, 287–294. doi:10.1006/mare.1995.1020

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247. doi:10.1111/jedm.1991.28.issue-3

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation based assessment of clinical competence. *Advances in Health Sciences Education*., *23*, 323-338. doi:10.1007/s10459-017-9800-3

Trigwell, K., & Prosser, M. (2014). Qualitative variation in constructive alignment in curriculum design. *Higher Education*, *67*(2), 141–154. doi:10.1007/s10734-013-9701-1

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181–195. doi:10.1037/1082-989X.6.2.181

van Gog, T. (2012). Expertise. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 1238–1240). New York: Springer.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*, 9–36.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education*, *8*, 157–186. doi:10.1207/s15324818ame0802_4

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

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149. doi:10.1177/0146621604271053

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/BF02294627

Weber, S., & Achtenhagen, F. (2017). Competence domains and vocational-professional education in Germany. In M. Mulder (Ed.), *Competence-based vocational and professional education bridging the world of work and education* (pp. 337–359). Cham: Springer International.

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

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1–14). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2004). Assessment, accountability and the classroom: A community of judgment. *Yearbook of the National Society for the Study of Education*, *103*(2), 1–19. doi:10.1111/j.1744-7984.2004.tb00046.x

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. doi:10.1111/jedm.1993.30.issue-3

Young, M. N., Markley, R., Leo, T., Coffin, S., Davidson, M. A., Salloum, J., ... Damp, J. B. (2018). Effects of advanced cardiac procedure simulator training on learning and performance in cardiovascular medicine fellows. *Journal of Medical Education and Curricular Development*, *5*, 1–5. doi:10.1177/2382120518803118

***Cogent Education*** **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**