Sangmeister *Empirical Res Voc Ed Train (2017) 9:3*
DOI 10.1186/s40461-017-0047-2

Empirical Research in Vocational Education and Training

RESEARCH Open Access

Commercial competence: comparing test results of paper-and-pencil versus computer-based assessments

Julia Sangmeister*

*Correspondence: sangmeister@die-bonn.de
German Institute for Adult Education-Leibniz Centre for Lifelong Learning, Bonn, Germany

Abstract

Background: Vocational education and training (VET) aims to enable young adults or trainees to participate in the workplace, and to promote their vocational capacities. In order to examine trainees' competencies at the end of VET, appropriate instruments are needed. This contribution aims: (1) to give an outline of such an instrument, one that has been designed to evaluate vocational competencies in the field of economics, and (2) to present the results of an empirical comparison of two possible test modes: computer-based assessment (CBA) versus paper-based assessment (PBA). The use of new technologies offers various opportunities for competence measurement: in particular, the computer as an assessment tool presents an authentic work tool drawn from professional life and promises novel ways of designing assessments. However, the current assessment practice in Germany is dominated by the use of traditional PBA, and there is less evidence about the possible effects of CBA. This study addresses the question of whether there are significant differences in the various ways of representing and measuring commercial competence with respect to specific content, item format, and, finally, motivational aspects.

Methods: A sample of 387 trainees from the German VET system was used to compare these two kinds of assessment. The analyses were realized using Item Response Theory and, particularly, Differential Item Functioning to detect differences between PBA and CBA at the item level. In addition to the performance data, motivational aspects, such as emotional state and test attractiveness, were also taken into account by a pre-/post-questionnaire.

Results: The study demonstrates that both test formats (CBA and PBA) can represent commercial competence in a valid and reliable way, but differences were found for certain items in the number of correct responses. The PBA shows a slight advantage in respect of overall item and model fit. Another key finding of our comparative study, at item level, is important from an instructive viewpoint: (domain) specific items are easier to solve in CBA than in PBA, whereas more general items are answered correctly more frequently in the latter. Contrary to expectations, we could not confirm the overall dominance of CBA against PBA on the basis of test takers' motivation, but values from CBA were more stable over time.

Conclusions: The study facilitated making the strengths and weaknesses of both test formats evident, and this implies the possibility of identifying opportunities for further development in assessment practice and in designing tests. Selected design criteria and aspects of test administration are discussed, with the aim of seeking to optimize

test development in order to create the best possible estimates for young adults' competence and capacity to participate in the world of work.

Keywords: Computer-based assessment, Differential item functioning (DIF), Mode-effect, Commercial competence, Validity

Background

The assessment of vocational competencies is a major challenge in the area of vocational education and training (VET), as appropriate instruments are still relatively scarce (Baethge and Arends 2009; Liedtke and Seeber 2015). In particular, there is a need for instruments to assess the content-specific competencies and skills needed for working life, in order to clarify whether vocational education prepares trainees adequately. This contribution aims: (1) to give an outline of such an instrument, one that has been designed to evaluate vocational competencies in the field of economics, and (2) to present the results of an empirical comparison of two possible test modes: computer-based assessment (CBA) versus paper-based assessment (PBA).

From an instructive perspective new technologies play an important role in VET, and in different ways. Firstly, they shift the educational content more closely toward the actual competencies needed in the workplace (Ludwig-Mayerhofer et al. 2011; Mayer and Solga 2008). For example, labelling the following themes "Ausbildung 4.0" (by which, in imitation of "Industry 4.0," we mean "VET 4.0"; Frey and Osborne 2013), Zinn (2015) picked out digitization, industrial automation, and interconnectedness as central themes for a future reorganization of learning processes in VET. The rationale behind this reorganization is that changes to the requirements of professional life must have a direct impact on the nature of the essential workplace competencies that have to be assessed. Given the background assumption that VET should primarily enable young adults or trainees to participate in the workplace and should promote their vocational capacities, these changes in the field of learning require us to review the test modes and formats that we have deemed adequate for assessment.

As observed by Pellegrino et al. (2003); see also Wilson (2005), the triadic elements of assessment, curriculum, and instruction all have to be integrally connected. New or changed learning content and learning forms must also be reflected in the assessment, and vice versa. Additionally, the results of recent studies on final examinations point to the need to redesign assessments: Herzog (2013), for example, speaks about the "gauging of education," and criticizes trends in schools whereby teachers align their lessons too strongly with external tests, notably with respect to content, task dimensions, and the methodological structure of the lesson. A test that is closely geared to real professional situations, in respect of content and type of work, could diminish the theory–practice gap: "teaching to the test" would, at the same time, be regarded as "teaching to the job."

Therefore, and secondly, new technologies offer new forms of, and mechanisms for, learning and competence measurement, given that it is necessary to "remove some of the constraints that have limited assessment practice in the past" (Pellegrino et al. 2003, p. 9). The computer presents—particularly in the commercial sector—an important and authentic work tool in professional life; thus, it seems helpful to administer computer-based tasks, in order to strengthen their authenticity by working with a relevant tool.

CBA offers the possibility of new (item) designs, as well as the integration of multimedia elements, such as video or audio sequences (Goldhammer et al. 2014) or interactive tools. These advantages have led to the assumption that tasks can be represented more authentically by CBA (Katz 2007; Huff and Sireci 2001) and that the capturing of strategic knowledge, as defined by more complex tasks, is made possible by this format (Quellmalz and Kozma 2003).

Thus, while CBA offers many opportunities and advantages for competence measurement, current assessment practice in Germany is dominated by the use of PBA (Gruttmann and Usener 2011), and there is a lack of evidence about the possible effects of putatively authentic assessment (Gulikers et al. 2005). Since the 1970s, several studies have made comparisons between CBA and PBA (Russell et al. 2003), but these have had very different designs, and the results are as multifarious as the studies themselves, notably with respect to the comparison criteria and their apparent impact on performance measurements (Bennett 2001; Pellegrino and Quellmalz 2010; Kingston 2009; Frahm 2012). Further, technological trends and computer-use behaviors are changing rapidly, so that the results of earlier studies can quickly become obsolete (Russell et al. 2003). An overview of comparability studies can be found in Wang and Shin (2009), and in a review of literature from the Texas Education Agency (2008).

A special characteristic of this study is that the CBA component of the assessment was designed, first and foremost, not to verify equivalent transferability (concurrent validity) from CBA to PBA, but to explicitly consider and utilize the opportunities and possibilities of a CBA by creating an authentic (workplace) environment. The PBA was adapted as closely as possible to the CBA by reprinting screenshots. The assessment was designed to measure competence in the world of work and training in an action-oriented and authentic way, which means: "requiring students to use the same competencies, 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). Authenticity does not inhere in relevant content, but must be derived from a generational process (Achtenhagen and Weber 2003). For this reason, several criteria have been developed for the design and evaluation of authentic assessments, including aspects such as complex and unstructured tasks, fidelity of the task to the conditions under which the performance would normally occur, as well as the production rather than reproduction of knowledge, and achieving validity and reliability with appropriate criteria for scoring varied products (e.g., Wiggins 1990; Herrington and Herrington 2006; Janesick 2006).

Thus, the aim of this study is to carry out an empirical comparison of CBA and PBA and to address the question as to whether there are significant differences in the ways in which commercial competence is presented in both test formats, with respect to specific content, item format, and, finally, motivational aspects.

Measuring VET competencies

Vocational competencies are defined in various ways, and are strongly linked to the professional domain to which they belong. In the following section, definitions of the concept of vocational competencies and the subordinate construct of commercial competence are given, before the assessment itself is introduced. Finally, the respective differences in test design for the two formats are discussed.

Baethge and Arends (2009, p. 9) describe vocational competencies as "young adults' abilities to successfully apply their knowledge and experience to authentic occupational situations in selected vocational areas in the world of work." In this case, vocational competence denotes an overarching concept that includes economic or commercial competence as one possible thematic orientation in the professional world. While there are established comparative studies and assessments in the field of general education (e.g., PISA, TIMSS), an equivalent large-scale assessment in the VET system, which focuses on apprentices´ transition to the labor market, is a desideratum. VET, especially the company-based VET program (the so-called "dual system"), a long-established tradition in Germany, is of great importance, not least because many young people are trained in this system. As measured by the number of apprenticeship contracts, the commercial domain "trade and industry" is one of the major areas of VET (BMBF 2016, p. 25). Finally, there are good reasons to focus on this area, given the high level of participation and the high relevance of new technologies in the commercial domain.

Following Winther and Achtenhagen (2008, p. 100), based on a systemic understanding of operational sub-processes and their reconstruction from real company data in real professional situations, commercial competence is defined as the ability to make business decisions and to validate them. From a theoretical and instructive perspective, one can distinguish between domain-linked (dl) and domain-specific (ds) facets of commercial competence, as proposed by Gelman and Greeno (1989). The domain-linked category refers to the notion of key vocational abilities, such as knowledge that is general but also relevant for solving vocational or professional problems. Domain-linked content is related to the general skills (mostly linguistic and mathematical) that are relevant in (commercial) occupational practice (Klotz et al. 2015); it differs from the content of the general education sector in its emphasis on professional relevance to economic areas. By way of an example, we might cite the understanding or reading of simple mathematical algorithms (e.g., the rule of three to calculate currency conversions or discounts). Domain-specific content includes specific sets of rules and a practical knowledge of a professional community, exclusively in commercial professions (Klotz and Winther 2016; Oates 2004; Winther et al. 2016a).

Test environment

To represent and measure commercial competencies in the field of VET a test environment named ALUSIM was developed. It is a simulation of an industrial company that produces extruded aluminum products, such as beverage cans (see also, Winther and Achtenhagen 2009, first edition; Winther 2010). The assessment is designed in such a way that test-takers assume the role of an apprentice in the ALUSIM company and are introduced to different situations via video clip, and then have to act in the simulated work area with a real computer. For this purpose, various documents are made available, such as customer lists and product catalogs, as well as working tools like writing and tabulation programs, email programs, calendars, calculators, and notepads. The assessment is intended to be *authentic* in terms of occupational situations, and *appropriate* with respect to abilities acquired through vocational training. It can be described as a "simulated" performance test that assesses working products and processes (see Kubiszyn and Borich 2010, p. 186) and promotes and captures decision-making. The items are divided

into four main curricular areas of focus: work preparation, purchasing, corporate communications, and sales. The response format varies between short-answer items and extended-response items (see Kubiszyn and Borich 2010; Nitko and Brookhart 2014; Hanna and Dettmer 2004). In respect of the latter, for example, emails have to be formulated (e.g., to make formal requests); for short-answer items, brief answers, consisting of numbers or key words, have to be supplied within a given template. These item formats, in contrast to multiple-choice questions, are suitable for assessing competencies that are relevant in vocational situations, because they measure how well test-takers are able to generate answers, act, and express themselves (see Hanna and Dettmer 2004).

The term "computer-based" encompasses the representation *of*, and the working *on*, tasks, and (at least partially) their evaluation (for an overview of concepts of electronic assessment see Jurecka and Hartig 2007). In particular, the simulation, as a special form of CBA, is an "artificial representation of the real thing" (Hanna and Dettmer 2004) and can open up new possibilities, because it "ensures a measurement of authentic abilities without bringing the testees in such real-world situations" (Winther and Achtenhagen 2009, p. 98).

In a second step, the PBA was constructed on the basis of the CBA: all items were transferred to a paper-based version. In order to generate the highest possible (visual) similarity, screenshots from the simulated test environment were added into the PBA, video sequences were transcribed, and, as the name implies, the handling was based on paper and pencil. The two tests (CBA and PBA) both attempted to capture the same latent commercial trait.

To ensure that the selected test content could be interpreted as relevant, preliminary studies were realized by using document analysis (including curricula, teaching materials, and regulations), interviews with experts, and workplace observations (see Winther et al. 2016b). The selection of adequate test and item content is elementary but difficult to objectify, so it generally entails the involvement of experts (Hartig et al. 2012). In addition, technical reliability and usability considerations represent key conditions that affect the validity and general acceptability of tests (Ehlers et al. 2013). Therefore, in order to ensure the overall functionality of the CBA, a usability test was conducted using the think-aloud method (Boren and Ramey 2000; Ericsson and Simon 1984; McDonald et al. 2012; Yom et al. 2007). The results were used to adapt the user interface over several revision cycles, and to prevent any interference in the operation of the CBA in advance (Sangmeister et al. 2018, in preparation).

Differences in test design: mode effects

There are good reasons to assume that CBA is bringing changes to assessment practice that affect organizational and technical aspects for schools and teachers, as well as the test-design itself (Jerrim 2016). The test design constitutes the framework, setting up test-taker interactions and working steps within the assessment, and has the potential to influence performance and results. In the literature, effects that are due to differences in test design are called "mode effects," and studies describe a "difference between the latent competencies of a test-taker for two tests administered in different modes" (Kröhne and Martens 2011, p. 174). The medium of administration—in our case, computer versus paper-and-pencil—is therefore an obvious and central design feature from

which mode effects can arise; "however, in practice, the simple distinction between the two test delivery strategies is only a convenient way to communicate a conglomerate of differences between the actual test administrations that might trigger mode effects and result in nonequivalent test administrations" (Kröhne and Martens 2011, p. 171). So one challenge is to describe design criteria in a transparent way for each test format, in order to form a basis for interpreting or judging the applicability of the results. Some relevant criteria at the test level that are usually used in empirical analyses are: (a) information searches, (b) multimedia elements, (c) response modes, (d) input devices, and (e) navigation.

The information search criterion (a) is also closely related to the type of test and differs between PBA and CBA. While in CBA the necessary information to solve tasks must be partially filtered by the test-takers themselves, as part of the task, the relevant documents in the PBA are bundled beforehand into a document wallet. Disadvantages can arise by interacting with CBA, especially when the simulation is not intuitive to use and/or when the effort of searching for information triggers temporal limitations.

In respect of the multimedia element criterion (b), CBA offers a series of design options in addition to the written language format that is usually used in PBA. A key difference between the tests is that in CBA it is a video, rather than a continuous text, that introduces the situation; this automatically reduces the effort required for reading. Mangen et al. (2013) assume that students who read texts in print scored significantly better than students who read the texts digitally. The information in the PBA text is easier to read or mark up, and brings advantages in respect of time required to answer. So there are some indications that completing a PBA is faster than finishing a CBA (Johnson and Green 2006). On the other hand, the simulation could be perceived as a less formal and more practical testing format, compared to the traditional educational testing format of written instructions, and thus may have also a positive effect on processing motivation (Garris et al. 2002).

Aspects specifically related to handling include response mode (c) and input devices (d). Giving responses (c) in CBA entails touching or clicking a keyboard or mousepad, whereas in PBA, paper and pencil are used in handwriting. Input devices (d) can be responsible for differences in performance when test time is affected, because information and communication technology (ICT) expertise might be required in handling the computer. Navigation possibilities (e) must also be considered: this involves within-item navigation, which for CBA means scrolling and clicking. Several studies confirm that too much scrolling to see the entire item can result in mode effects (Kingston 2009). Thus, a computer simulation entails a certain degree of risk, because the information needed to solve the items is not always visible at a glance. In our test the within-test navigation, meaning the sequences in which items are answered, and the possibilities for revising and omitting items, were even more important. In the CBA, a predetermined task sequence has to be followed. Once completed, a task cannot be revised. It is possible, however, to finish an item without supplying an answer in the relevant location. The question of revision is often described in studies as the cause of mode effects (Lunz and Bergstrom 1994), but results and interpretations differ, and are influenced by test characteristics (speed or power). The possibility of revising may cause test-takers to deliberately seek out and work on items which they have a high probability of solving, and thus

achieve better test results. On the other hand, the revision of tasks and the possibility of "jumping back and forth" does adversely affect test time, and hence may lead to a lower number of correct answers.

At the level of the sample, contextual factors are often analyzed for comparison (e.g., age, socioeconomic background), as well as motivational aspects and computer experience. Studies that examine the age of the test-takers have found different results, but there seems to be a tendency for primary-school age to generate more effects (Pomplun et al. 2006). This, in turn, relates to the computer experience of children, but due to the progress in digitization this effect continues to be less of an issue (Wang and Shin 2009).

Research questions

In order to compare both test formats, we had to ensure that the construct of commercial competence had been tested in an appropriate form in both PBA and CBA. It is considered valid if the established parameters of item analysis and model fit comply with current conventions. If these psychometric conditions are achieved for both tests, then a positive interpretation of construct validity is allowable, and a comparison between PBA and CBA is reasonable. In IRT, internal validity is deemed to be approved through model fit (Rost 2006). Further, the focus on vocational competencies in the workplace is strongly linked to the increasing importance of the use of information and communication technologies. In this case, we assume that a computer-based and authentic assessment increases construct validity (Gielen et al. 2003) because of an improved representation of domains (Pellegrino and Quellmalz 2010). For the purposes of verification, reliabilities and item- and model-fit measures of CBA and PBA were compared. We assume that CBA maps the tasks in a way that is more authentic, so we hypothesize that:

(1) CBA displays a significantly better model fit than PBA

A detailed analysis was performed at item level to filter out items that are systematically easier (or more difficult) in either CBA or PBA, and thus have a DIF-Effect. Although PBA may well require declarative and procedural knowledge, its scope for mapping realistic operations in the workplace is limited (Goldhammer et al. 2014). Since authentic design (using job-specific documents and procedures) is important, it follows that job-related (specific) tasks can be represented and displayed more realistically in CBA (Huff and Sireci 2001; Katz 2007) and, thus, that items that are classified as domain-specific and more complex can be better reflected in CBA (Quellmalz and Kozma 2003). So we hypothesize that:

(2) DIF-Items for CBA are mainly domain-specific.

In addition to aspects based on the test level, issues at the level of sample should also be addressed. We examined to what extent motivational aspects make a difference between test formats, and whether there is a difference in respect of time (before and after testing). Motivation is operationalized using scales for emotional state and attractiveness. It was expected that emotional state would turn out to be higher before and after working with CBA because it seems to be more engaging than classical PBA and students prefer acting on CBA (see Johnson and Green 2006).

(3) The motivation of the test-taker, considering (a) emotional state and (b) test attractiveness, is significantly higher in CBA than in PBA before and after testing (ex ante/ex post).

Methods

The study amassed data from CBA and PBA performance tests, as well as from a motivation questionnaire. The survey was conducted in 2013 in twenty-six classes in three German federal states. The test-takers were in their second year of apprenticeship for industrial clerks and, on average, 20 years old, with a range of 17–37 years ($SD = 2.37$) (see Table 1).

Data collection was organized by a within-subject design, meaning that while each test-taker had to complete the whole test, one half was executed in PBA (first part) and the other half in CBA (second part). Therefore, a balanced rotation design with a total of eight booklet rotations was applied (each booklet with a sample of $n = 62–72$). Although both tests were designed to measure the same underlying construct there was no overlap of items. We received data sets from PBA ($n = 523$) and CBA ($n = 414$), with a total of thirty-four items being administered to measure commercial competence. The final sample consisted of 387 apprentices, because only those examinees who had worked on both tests (PBA and CBA) were included in the analysis. To scale the competence data, invalid responses or omitted items were rated as incorrect answers (0 = no credit). Items that were not administered due to the rotational design were excluded. The performance tests were framed by questionnaires asking for aspects of motivation. The test-takers were invited to perform a self-assessment ($n = 294$) before and after working on CBA and PBA, including the scales "emotional state" and "test attractiveness" (adapted from PISA studies; Kunter et al. 2002). Emotion and motivation actually represent two separate psychological dimensions but cannot be considered in isolation (Seifried and Sembill 2005; Sembill 1992; Schumacher 2002). The response scale ranges from 0 = no agreement to 3 = high agreement. Here also, only a reduced sample was realized, with missingness of $n = 93$. To analyze these data, a within-subject ANOVA was used.

The Item Response Theory (IRT) is an established method used to ascribe the competence values of trainees to contextual and situational requirements (Hartig and Frey 2013). Relevant models of probabilistic test theory (De Ayala 2009), therefore, were calculated with generalized Item Response Modelling Software (Acer´s ConQuest 2.0; Wu

Table 1 Sample of industrial clerks

Sample	N = 387
Gender	Female (67%)
	Male (33%)
Federal state	Baden-Württemberg (31%)
	Bavaria (31.5%)
	Hesse (37.5%)
Age	17–18 (33.3%)
	19–20 (37.5%)
	21–22 (21.2%)
	>23 (8%)

et al. 2007). The present work was based on a partial credit model (Masters 1982) as a form of the ordinal Rasch (1960) model that permits partial credit items, with 0 = no credit, 1 = partial credit, 2 = full credit, while the estimation of people and task parameters was determined by the maximum likelihood principle (multidimensional random coefficient multinomial logit model [MRCML]) (see Adams et al. 1997). The question of test validity is decisive for the acceptance of an instrument and for successful testing. Construct validity, in particular, is a key test criterion that can be seen as an overarching concept (Anastasi 1986; Hartig et al. 2012). Thus, validity is not a characteristic of the test per se, but refers to the interpretation of test scores (Borsboom et al. 2004).

In addition to these model tests, the detecting of DIF (Holland and Wainer 1993) was of importance, allowing for comparison of the two test formats at item level. DIF is most commonly used to assess fairness aspects in the process of instrument development and adaptation (Ackerman 1992; Roussos and Stout 1996) by checking whether individual items operate differently for different test subgroups, thus detecting violations of validity. Zumbo (2007, p. 228) divides changes in DIF-research into three generations, wherein the third generation is characterized by "the matter of wanting to know why DIF occurs." We refer to items as DIF-Items if the probability of a solution is not fully explained by the person's ability and/or item difficulty (Adams and Carstensen 2002). DIF-based studies permit statements about the performance of groups on an assessment, controlling for overall ability of the groups. According to Embretson and Reise (2000, p. 319), "a scale item displays DIF if examinees with the same latent-trait level have different probabilities of endorsing an item […]." In our case, we examined the extent to which group differences were caused by the test mode (CBA and PBA) and, therefore, not by personal characteristics such as gender or age. Group affiliation is characterized by the delivery strategy so Group 1 = PBA and Group 2 = CBA.

Results and discussion

Comparing facets of construct validity at test level (H1)

To check the overall fit in both formats, analyses of item, reliability, and model fit were used (see Table 2). The fit of several items to the model was evaluated using the weighted mean square (wMNSQ) approach. The wMNSQ describes the deviation of the observed probability for a correct response from the model-implied probability for a given ability level (Pohl and Carstensen 2012) and, therefore, a wMNSQ near 1 indicates a good fit with $0.75 \leq \text{wMNSQ} \leq 1.33$ (see Bond and Fox 2001). The respective t-values are inference statistical measures for the null hypothesis that the wMNSQ equals one with weighted fit t-values from $-2.0 > t < 2.0$ (see Wright and Masters 1982).

As shown in Table 2, the data yielded good item statistics: for PBA, $0.83 \leq \text{wMNSQ} \leq 1.21$; $-2.24 \leq$ Item Thresholds (difficulty) $\leq +3.13$.; and for CBA, $0.78 \leq \text{wMNSQ} \leq 1.22$; $-1.77 \leq$ Item Thresholds (difficulty) $\leq +2.62$. There were found one item for PBA and nine items for CBA that provide values $t > |2|$. Nevertheless, these items were not excluded because of their relevance with regard to content. With regard to the differences in item thresholds, the PBA offered a higher range of item thresholds for the same item pool that can be identified for CBA. The reliabilities are described through WLE and EAP/PV coefficients that indicate higher values for PBA; however, the difference between the test formats is likely to be low (WLE difference of 0.019, EAP/PV

Table 2 Comparison of PBA and CBA by item statistics, reliability, and model fit

N = 387	Items = 34	PBA	CBA
Item statistics			
	wMNSQ	0.83 to 1.21 ($M = 0.99$; $SD = 0.11$)	0.78 to 1.22 ($M = 0.99$; $SD = 0.14$)
	t-value	−2.0 to 2.2 ($M = \|0.95\|$; $SD = 0.58$)	−2.9 to 2.5 ($M = \|1.41\|$; $SD = 0.81$)
	Item thresholds	−2.24 to 3.13 ($SD = 1.42$)	−1.77 to 2.62 ($SD = 1.31$)
Reliability			
	WLE	0.753	0.734
	EAP/PV	0.801	0.784
Model fit			
	Final deviance	6992.74	7442.67
	Estimated parameters	41	41
	BIC	7237.03	7686.97
	AIC	7074.74	7524.67
	cAIC	7084.72	7534.65

difference of 0.057). The psychometric requirements are thus complied with for both test formats, in respect of item fit and reliabilities.

There are no absolute measurements of model fit that can be used for comparison; therefore, different model quality indices based on deviance, such as the Akaike Information Criteria (AIC), the corrected Akaike Information Criteria (cAIC), and the Bayesian Information Criterion (BIC) were reported. Deviance is a measure of deviation between the model and the data: the greater the deviance, the greater the deviation from the model. Models with smaller AIC, cAIC, and BIC values are empirically preferred. In our case, the PBA data provided a higher model fit (BIC = 7237.03; AIC = 7074.74; cAIC = 7084.72).

Additionally, one can see the person's ability (WLE) and reached raw scores. The WLE describes the most likely competence score for each single person, given the item responses of that person (Pohl and Carstensen 2012). We found a mean value for WLE in PBA with −0.3 ($SD = 1.51$) and in CBA with −0.6 ($SD = 1.31$), so the person's ability in PBA generates a wider range (between −5.09 and 3.92) than CBA (between −4.56 and 2.48). On average, 8.93 ($SD = 5.81$) tasks were solved in PBA and 7.94 ($SD = 5.13$) tasks in CBA, one item less in CBA. Generally, it can be concluded that the items were rather too difficult for the sample, as indicated by a lower proportion of correct answers (see Fig. 1). Only 15 items for PBA and 12 items for CBA had a proportion of correct answers (full credit) higher than 50%.

Both tests provided basically permissible values, but PBA performed slightly better, considering item parameters, reliabilities, model fit, and personal ability.

Hypothesis (1), that CBA displays a significantly better model fit than PBA, is rejected.

Contrary to expectations formulated in Hypothesis 1, the data show a slight advantage for PBA in terms of the proportion of correct answers and model fit. However, the difference is very low (only one additional task was achieved in PBA, and a difference of deviance of Δ449.93). It was determined that test rotations had no significant effect on test

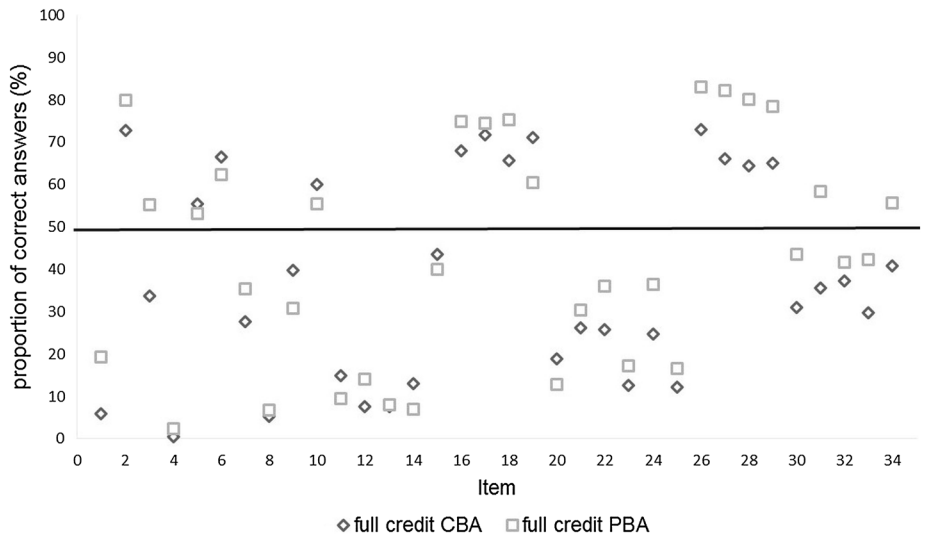

Fig. 1 Proportion of correct answers (%) for items in CBA and PBA

performance for PBA ($F < 1$, $p = 0.767$), but there is an effect for CBA [$F(7379) = 5.21$, $p = < 0.001$] that affects two combinations of rotation.

Mode effects at test level, as discussed above, may provide an explanation for these differences, as well as sample effects, like familiarity with the test format, because trainees at the vocational school usually use PBAs that are neither presented in the form of a simulation nor as CBAs. The novelty of CBA could conceivably lead to a situation in which test-takers initially try out functions and options simultaneously, so that the processing is less focused and accordingly more time consuming, whereas in PBA all the relevant information is present in a concentrated way (see "Measuring VET competencies" section: "information search"), and it is easier to get an overview of all the documents (see "Measuring VET competencies" section: "navigation"); this is less obvious in the CBA. Furthermore, the reprocessing of tasks (see "Measuring VET competencies" section: "revision") is not possible.

Detecting DIF to compare test delivery strategies (PBA vs. CBA) at item level (H2)

By using DIF analyses in a between-design, we identify items that are significantly easier or alternatively more difficult to solve, in each of the test formats (see Table 3). These items can be examined for similarities and differences related to configuration/design and type.

First, the DIF analysis confirmed the results above, and demonstrated for a group difference at test level in favor of the PBA. This means that test-takers underperformed in CBA, at 0.34 logits. At the item level, it becomes apparent that 21 items have a moderate (7) to large (14) DIF effect, of which 10 can be interpreted in favor of the PBA and 11 in favor of the CBA. This classification was done according to Paek (2002). Based on the underlying theoretical model of competence, all items were classified (ex ante) with regard to their domain-specific or domain-linked components (Winther et al. 2016b). Further investigation showed that of these CBA DIF items, a total of 8 items were classified as domain-specific (73%). This suggests that specific items can be displayed in the

Table 3 DIF items for PBA and CBA with domain-linked (dl) and domain-specific (ds) item classification

Item	Classification domain-linked (dl) domain-specific (ds)	PBA (logits)	CBA (logits)	Error	Chi square (df)	p value
Item 1	dl	−1.092		0.075	478.97 (33)	<0.001
Item 3	dl	−0.938		0.085		
Item 4	dl	−0.578		0.082		
Item 5	dl		0.740	0.083		
Item 6	ds		0.826	0.085		
Item 7	ds		0.766	0.071		
Item 8	ds		0.456	0.105		
Item 9	ds		1.118	0.085		
Item 10	dl		0.926	0.084		
Item 11	ds		0.762	0.073		
Item 12	ds		0.982	0.092		
Item 15	ds		0.432	0.084		
Item 19	dl		0.872	0.086		
Item 20	ds		0.788	0.094		
Item 24	ds	−0.438		0.087		
Item 26	dl	−0.446		0.092		
Item 27	dl	−0.796		0.09		
Item 28	dl	−0.710		0.089		
Item 29	dl	−0.550		0.089		
Item 31	dl	−0.918		0.084		
Item 34	ds	−0.486		0.496		
Total						
21 (of 34)	11 dl 10 ds	10	11			

CBA in a helpful way, so we have evidence, using the DIF analysis, that the specificity of the item had a positive effect on working with CBA.

Hypothesis (2), that DIF-Items for CBA are mainly domain-specific can be confirmed.

Conversely, we found that items that are easier to solve in PBA were predominantly domain-linked. Thus, the results are independent of the response format (essay items and short-answer items).

Evaluation of motivational aspects at sample level (H3)

The descriptive results of motivational aspects from the self-assessment are represented in Table 4 for PBA and Table 5 for CBA.

For both tests there is a higher agreement before testing (ex ante) compared to after testing (ex post) for emotional state and attractiveness of test. Overall, the motivational rating for both formats is rather sobering. The lowest rating was found for "Attractiveness of test".

For each scale four things are measured: approval ratings for the two test forms (PBA and CBA) at two times (ex ante and ex post).

Figure 2 shows mean emotional state as a function of the two test forms (PBA vs. CBA) and time (ex ante vs. ex post): there was no significant effect on test forms, with $F(1293) = 1.4$, MSE $= 0.215$, $p = 0.238$ (n.s.), $\eta_p^2 = 0.005$, but there was an effect for time,

Table 4 Mean values of motivation scales for PBA (ex ante/ex post)

Scale (items included) $n = 294$	M	SD	Cronbach's α
Emotional state (5)			
Ex ante	1.85	0.64	0.84
Ex post	1.54	0.64	0.80
Attractiveness of test (3)			
Ex ante	1.14	0.63	0.84
Ex post	1.02	0.65	0.79

3-point Likert-scale from 0 = no agreement to 3 = high agreement

Table 5 Mean values of motivation scales for CBA (ex ante/ex post)

Scale (items included) $n = 294$	M	SD	Cronbach's α
Emotional state (5)			
Ex ante	1.73	0.63	0.82
Ex post	1.60	0.67	0.82
Attractiveness of test (3)			
Ex ante	1.11	0.65	0.89
Ex post	1.09	0.66	0.80

3-point Likert-scale from 0 = no agreement to 3 = high agreement

with $F(1293) = 72.86$, MSE $= 0.188$, $p < 0.001$, $\eta_p^2 = 0.199$, and an interaction effect for format and time: $F(1293) = 15.01$, MSE $= 0.156$, $p < 0.001$, $\eta_p^2 = 0.049$. The post hoc t test illustrates that there were significant differences in emotional state before testing (ex ante) between PBA and CBA, $t(293) = 3.74$ with $p < 0.001$, but not after testing (ex post), $t(293) = -1.5$, $p = 0.135$ (n.s.). These results also show that over time emotional state was more stable in CBA than in PBA.

Figure 3 depicts the attractiveness of the test and offers a significant main effect for time with $F(1293) = 10.715$, MSE $= 0.147$, $p = 0.001$, $\eta_p^2 = 0.035$, but not for format $F(1293) = 0.439$, MSE $= 0.187$, $p = 0.508$ (n.s.), $\eta_p^2 = 0.001$. The interaction term was significant with $F(1293) = 4.552$, MSE $= 0.141$, p $= 0.034$, $\eta_p^2 = 0.015$. The post hoc t-Test provided an effect on time (ex ante/ex post) only for PBA, with $F(293) = 3.883$, $p < 0.001$.

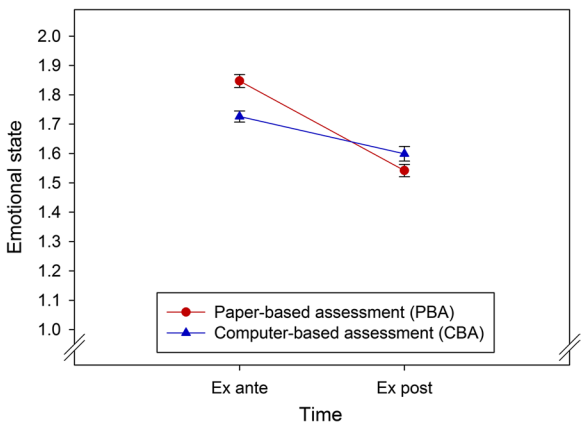

Fig. 2 Emotional state as a function of test format (PBA vs. CBA) and time (ex ante vs. ex post). *Error bars* show the within-subjects standard errors of the means, using the algorithm suggested by Cousineau (2005)

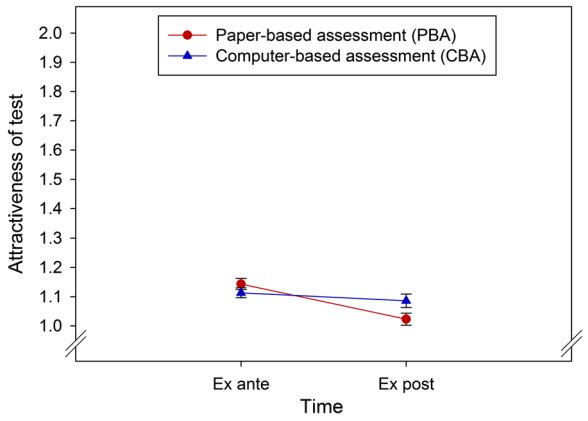

Fig. 3 Attractiveness of test as a function of test format (PBA vs. CBA) and time (ex ante vs. ex post). *Error bars* show the within-subjects standard errors of the means, using the algorithm suggested by Cousineau (2005)

In summary, the mean values before testing (ex ante) differ significantly for emotional state between the test formats, but not after testing (ex post). In particular, we expected higher ratings for CBA due to the aforementioned restructuring of the tests. Following the observation that the use of computers is no longer "anything special" for the sample of so-called "digital natives," and that no disadvantages were expected in using a computer, we can simultaneously assume that there were no motivational effects, simply because a computer is now part of everyday life (see also Frey et al. 2009). However, an interesting finding here is that emotional state (ex ante) in the CBA was lower than in the PBA, but CBA-values were more stable over time (ex post) and decreased less strictly. The latter finding was also observed for attractiveness of test.

Hypothesis (3), that the motivation of test-takers, considering (a) emotional state and (b) test attractiveness, is significantly higher in CBA than in PBA before and after testing (ex ante/ex post), cannot be confirmed. However, motivation of test-takers was more stable over time for CBA.

Conclusions

In the context of VET competence measurement, the crucial question is not just *what* test-apprentices know, but, in particular, what they are *able* to do and *how* they act. Therefore, instruments are necessary in order to determine and externalize these competence aspects effectively for evaluation of performance (status quo), the quality and improvement of training, and the prospects of success in the labor market. This article presents an advanced instrument from the field of electronic assessment for measuring commercial competence, which claims to optimize competence measurement through the use of authentic content and presentation. To achieve a better understanding of effects and mechanisms, a comparison was carried out between CBA and PBA, taking mode effects into account. Because mode effects vary depending on the test content and, thus, are not necessarily transferable and generalizable (see Kröhne and Martens 2011; Ito and Sykes 2004), it is necessary to have a look at the specific instrument. The aim was to illuminate the characteristics of CBA and to use this information for further development and optimization of those new assessments.

The study showed that both test formats (CBA and PBA) can represent commercial competence in a valid and reliable way. Contrary to expectations, PBA manifested a slight advantage in respect of item and model fit. The dominance of CBA against PBA, based on the quality of the construct representation, or the test motivation of the test-takers, could not be confirmed. Overall, there was only little endorsement for both formats in absolute terms, but the motivation of test-takers was more stable over time for CBA and decreases less strictly. Furthermore, the specificity of tasks is of importance: CBA represents (domain) specific items in an easier way than PBA, while PBA fits better for domain-linked items (more general tasks). Depending on the purpose of a test, a combination of both forms may be appropriate. From a scientific perspective, we have gained information about assessment design criteria that can improve test performance, and that should be enhanced in vocational learning settings, in order to promote competence development affecting the socialization of young adults into the world of work. This also supports educators, at a practical level, by designing assessments.

Several limitations of the study suggest the need for future study improvements. For example, with respect to design, on an individual level, there were no identical tasks in both CBA and PBA for a pairwise comparison. The advantages of this type of design are that a smaller sample is sufficient, and it obviates the need to test twice, which could possibly have negative effects on motivation, such as fatigue or boredom (Texas Education Agency 2008), and, moreover, apprentices are generally better able to carry out the task on a second attempt. This approach is also justified by the fact that, from a pragmatic, test-economic perspective, the rotating design reduces test time, even though all items could be included, and, on the other hand, carry-over effects could be avoided.

Moreover, additional surveys would be able to provide a deeper analysis of the mode effects. We described differences between CBA and PBA that may affect test scores, and the empirical analysis showed evidence for this, but it has not been possible to identify the specific driving force (see also Jerrim 2016). For example, while we generated log data for CBA that can supply information on processing procedures and times, such information is not available for the PBA, and so a comparison is not possible. The assumption that differences occurring are based on processing times (for example, tapping vs. writing, or introductory text vs. video introduction) cannot be adequately tested.

To assess the "suitability" of e-assessment, the whole testing process needs to be considered, from preparation to implementation to post-processing (Ehlers et al. 2013). Questions such as avoiding interference, providing infrastructure, and data protection are essential for the use of CBA, but were not singled out here as central themes. To make a final judgment, these criteria would have to be discussed in more detail.

With regard to content, we may discuss to what extent the use of computers (in terms of ICT Literacy) must be regarded as another dimension of competence, and to what extent it is part of commercial expertise. Assessments used to measure computer skills—for example, the ISkills (Information and Communication Technology Literacy Test) of the ETS (2002), or the BCS (Basic Computer Skills) of Goldhammer et al. (2014), could help to identify the level of ICT competence and monitor the impact on the CBA test results for commercial apprentices.

Abbreviations
ASCOT: Technology-based Assessment of Skills and Competencies in VET; PBA: paper-based assessment; CBA: computer-based assessment; CoBALIT: Competencies in the Field of Business and Administration, Learning, Instruction and Transition (Research project funded by the German Federal Ministry of Education and Research [BMBF]); ICT: information and communication technology; PISA: Programme for International Student Assessment; ManKobE: Mathematics and Science Competencies in Vocational Education and Training (Research project funded by the Leibniz Society); TIMSS: Trends in International Mathematics and Science Study; VET: vocational education and training.

Acknowledgements
This research study is part of the ManKobE (Mathematics and Science Competencies in Vocational Education and Training) project, funded by the Leibniz Society and under the direction of the IPN—Leibniz Institute for Science and Mathematics Education at Kiel University (http://www.ipn.uni-kiel.de/en/research/projects/mankobe?set_language=en).

Preliminary studies based on the CoBALIT project (Competencies in the Field of Business and Administration, Learning, Instruction, and Transition) are embedded in the research initiative ASCOT (Technology-based Assessment of Skills and Competencies in VET; http://www.ascot-vet.net/) and sponsored by the Federal Ministry of Education and Research (BMBF): Reference Nos. 01DB1115 and 01DB1118.

Competing interests
The author declares that she has no competing interests.

Availability of data and materials
All data are available from the author.

Consent for publication
The author hereby consent to publication.

Ethics approval and consent to participate
The study was examined and approved by the ministries of education of the collaborating federal states of Germany (Hesse, Bavaria, Baden-Württemberg). Participants were informed about the aim of the study and that their participation was anonymous and voluntary.

Funding information
Leibniz-Association.
Award Number: SAW-2012-IPN-2 | Recipient: not applicable.
Federal Ministry of Education and Research.
Award Number: 01DB1115 | Recipient: not applicable.

Received: 2 August 2016 Accepted: 17 February 2017
Published online: 27 February 2017

References
Achtenhagen F, Weber S (2003) "Authentizität" in der Gestaltung beruflicher Lernumgebungen ["Authenticity" in designing professional learning environments]. In: Bredow A, Dobischat R, Rottmann J (eds) Berufs- und Wirtschaftspädagogik von A-Z. Grundlagen, Kernfragen und Perspektiven. [Vocational, economic, and business education from A-Z. Basics, core issues and perspectives]. Schneider, Baltmannsweiler, pp 185–199

Ackerman TA (1992) A Didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. J Educ Meas 29:67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

Adams R, Carstensen C (2002) Scaling outcomes. In: Adams R, Wu M (eds) PISA 2000 Technical Report, OECD, Paris, pp 149–162. http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33688233.pdf. Accessed 20 Dec 2016

Adams RJ, Wilson MR, Wang WC (1997) The multidimensional random coefficients multinomial logit. Appl Psychol Meas 21:1–24. doi:10.1177/0146621697211001

Anastasi A (1986) Evolving concepts of test validation. Annu Rev Psychol 37:1–15. doi:10.1146/annurev.ps.37.020186.000245

Baethge M, Arends L (eds) (2009) Feasibility study VET-LSA. A comparative analysis of occupational profiles and VET programmes in 8 european countries. International report. Soziologisches Forschungsinstitut, Göttingen

Bennett RE (2001) How the internet will help large-scale assessment reinvent itself. Educ Policy Anal Arch 9(5):1–23. doi:10.14507/epaa.v9n5.2001

Bond TG, Fox CM (2001) Applying the Rasch-model: fundamental measurement in the human sciences. Lawrence Erlbaum, Mahwah

Boren MT, Ramey J (2000) Thinking aloud: reconciling theory and practice. IEEE Pro Comm Discipl 43(3):261–278. doi:10.1109/47.867942

Borsboom D, Mellenbergh GJ, van Heerden J (2004) The concept of validity. Psychol Rev 111:1061–1071. doi:10.1037/0033-295X.111.4.1061

Bundesministerium für Bildung und Forschung (BMBF) (2016) Berufsbildungsbericht 2016. [VET report 2016] Bonn, Berlin 2016. https://www.bmbf.de/pub/Berufsbildungsbericht_2016.pdf. Accessed 1 Nov 2016

Cousineau D (2005) Confidence intervals in within-subject designs: a simpler solution to Loftus and Massons's method. Tutor Quant Methods Psychol 1(1):42–45. doi:10.20982/tqmp.01.1.p042
De Ayala RJ (2009) Theory and practice of item response theory. Guilford Press, New York
Educational Testing Service (ETS) (2002) Digital transformation: a framework for ICT literacy. ETS, Princeton. http://www.ets.org/Media/Tests/Information_and_Communication_Technology_Literacy/ictreport.pdf. Accessed 1 Aug 2016
Ehlers JP, Guetl C, Höntzsch S, Usener CA, Gruttmann S (2013) Prüfen mit Computer und Internet. Didaktik, Methodik und Organisation von E-Assessment. [Computer and internet-based testing. Didactics, methodology and organization of e-assessment] In: Ebner M, Schön S (eds) Lehrbuch für Lernen und Lehren mit Technologien (L3T) [Textbook for learning and teaching with technologies], vol 2E. http://www.pedocs.de/volltexte/2013/8348/pdf/L3T_2013_Ehlers_et_al_Pruefen_mit_Computer.pdf. Accessed 1 Aug 2016
Embretson SE, Reise SP (2000) Item response theory for psychologists. Lawrence Erlbaum, Mahwah
Ericsson KA, Simon HA (1984) Protocol analysis: Verbal reports as data. Bradford books/MIT Press, Cambridge
Frahm S (2012) Computerbasierte Testung der Rechtschreibleistung in Klasse fünf—eine empirische Studie zu Mode-Effekten im Kontext des Nationalen Bildungspanels [Computer-based testing of literacy in 5th grade—an empirical study of mode effects in the context of the national educational panel]. Logos-Verlag, Berlin
Frey C, Osborne MA (2013) The future of employment: how susceptible are jobs to computerization? University of Oxford. http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf. Accessed 1 Aug 2016
Frey A, Hartig J, Moosbrugger H (2009) Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Test. [The effects of adaptive testing on test-taking motivation using the example of the Frankfurt Adaptive Concentration Test]. Diagnostica 55(1):22–28. doi:10.1026/0012-1924.55.1.20
Garris R, Ahlers R, Driskell JE (2002) Games, motivation, and learning: a research and practice Model. Simul Gaming 33(4): 441–467. doi:10.1177/1046878102238607. http://brainadvantage.com/PDF/Games,%20Motivation,%20and%20Learning.pdf. Accessed 1 Aug 2016
Gelman R, Greeno JG (1989) On the nature of competence: principles for understanding in a domain. In: Resnick LB (ed) Knowing, learning and instruction: essays in honor of Robert Glaser. Lawrence Erlbaum, Hillsdale, pp 125–186
Gielen S, Dochy F, Dierick S (2003) Evaluating the consequential validity of new modes of assessment: The influence of assessment on learning, including pre-, post-, and true assessment effects. In: Segers M, Dochy F, Cascallar E (eds) Optimising new modes of assessment: in search of quality and standards. Kluwer Academic Publishers, Dordrecht, pp 37–54. doi:10.1007/0-306-48125-1
Goldhammer F, Kröhne U, Keßel Y, Senkbeil M, Ihme JM (2014) Diagnostik von ICT-Literacy: multiple-choice- vs simulationsbasierte Aufgaben [Assessment of ICT literacy: multiple-choice vs. simulation-based tasks]. Diagnostica 60:10–21. doi:10.1026/0012-1924/a000113
Gruttmann S, Usener C (2011) Prüfen mit Computer und Internet. Didaktik, Methodik und Organisation von E-Assessment [Computer and internet-based testing. Didactics, methodology and organization of e-assessment]. In: Ebner M, Schön S (eds) Lehrbuch für Lernen und Lehren mit Technologien (L3T) [Textbook for learning and teaching with technologies (L3T)]. BIMS, Bad Reichenhall. http://l3t.tugraz.at/index.php/LehrbuchEbner10/article/view/37. Accessed 1 Aug 2016
Gulikers JTM, Bastiaens TJ, Kirschner PA (2004) A five-dimensional framework for authentic assessment. Educ Technol Res Dev 52(3):67–86. doi:10.1007/BF02504676
Gulikers JTM, Bastiaens ThJ, Martens R (2005) The surplus value of an authentic learning environment. Comput Hum Behav 21:509–521. doi:10.1016/j.chb.2004.10.028
Hanna GS, Dettmer PA (2004) Assessment for effective teaching: using context-adaptive planning. Pearson A&B, Boston
Hartig J, Frey A (2013) Sind Modelle der Item-Response-Theorie (IRT) das "Mittel der Wahl"für die Modellierung von Kompetenzen? [Benefits and limitations of modeling competencies by means of Item Response Theory (IRT)]. Zeitschrift für Erziehungswissenschaft 16(1):47–51. doi:10.1007/s11618-013-0386-0
Hartig J, Frey A, Jude N (2012) Validität [Validity]. In: Moosbrugger H, Kelava A (eds) Test- und Fragebogenkonstruktion [Test construction]. Springer, Berlin, pp 143–171
Herrington JA, Herrington AJ (2006) Authentic conditions for authentic assessment: aligning task and assessment. In: Bunker A, Vardi I (eds) Proceedings of the 2006 annual international conference of the Higher Education Research and Development Society of Australasia Inc (HERDSA): critical visions: thinking, learning and researching in higher education: Research and Development in Higher Education, 29, pp 141–151
Herzog W (2013) Bildungsstandards. Eine kritische Einführung [Standards of education. A critical introduction]. Kohlhammer Verlag, Stuttgart
Holland PW, Wainer H (1993) Differential item functioning. Lawrence Erlbaum, Hillsdale
Huff KL, Sireci SG (2001) Validity issues in computer-based testing. Educ Meas 20(3):16–25. doi:10.1111/j.1745-3992.2001.tb00066.x
Ito K, Sykes RC (2004) Comparability of scores from norm-reference paper-and-pencil and web-based linear tests for grades 4–12. Paper presented at the annual meeting of the American Educational Research Association, San Diego. http://www.ctb.com/img/pdfs/raPaperVsWebLinearTests.pdf. Accessed 1 Aug 2016
Janesick VK (2006) Authentic assessment. Peter Lang, New York
Jerrim JP (2016) PISA 2012: how do results for the paper and computer tests compare? Assess Educ Princ Policy Pract 23(4):495–518. doi:10.1080/0969594X.2016.1147420
Johnson M, Green S (2006) On-line mathematics assessment: the impact of mode on performance and question answering strategies. J Technol Learn Assess 4(5):4–35
Jurecka A, Hartig J (2007) Anwendungsszenarien computer- und netzwerkbasierten Assessments [Application scenarios for computer- and network-based assessments]. In: Hartig J, Klieme E (eds) Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung. [Possibilities and requirements of technology-based competence diagnostics]. BMBF, Bonn pp 69-79. https://www.bmbf.de/pub/Bildungsforschung_Band_20.pdf. Accessed 1 Aug 2016

Katz IR (2007) Testing information literacy in digital environments: the ETS iSkillsTM assessment. Inform Technol Libr 26:3–12
Kingston NM (2009) Comparability of computer- and paper-administered multiple choice tests for K-12 populations: a synthesis. Appl Meas Educ 22(1):22–37. doi:10.1080/08957340802558326
Klotz VK, Winther E (2016) Zur Entwicklung domänenverbundener und domänenspezifischer Kompetenz im Ausbildungsverlauf: Eine Analyse für die kaufmännische Domäne [Developing domain-related and domain-specific competence in VET: an analysis for the commercial domain]. Zeitschrift für Erziehungswissenschaft 19(4):765–782. doi:10.1007/s11618-016-0687-1
Klotz VK, Winther E, Festner D (2015) Modeling the development of vocational competence: a psychometric model for economic domains. Vocat Learn 8(3):247–268. doi:10.1007/s12186-015-9139-y
Kröhne U, Martens T (2011) Computer-based competence tests in the National Educational Panel Study: the challenge of mode effects. Zeitschrift für Erziehungswissenschaften 14(2):169–186. doi:10.1007/s11618-011-0185-4
Kubiszyn T, Borich G (2010) Educational testing and measurement: classroom application and practice, 9th edn. Wiley, New York
Kunter M, Schümer G, Artelt C, Baumert J, Klieme E, Neubrand M, Prenzel M, Schiefele U, Schneider W, Stanat P, Tillmann KJ, Weiß M (2002) PISA 2000: Dokumentation der Erhebungsinstrumente [PISA 2000: documentation of survey questionnaires]. MPI für Bildungsforschung, Berlin
Liedtke M, Seeber S (2015) Modellgeltungstests und Einflussfaktoren auf differentielle Itemfunktionen in einem computergestützten Assessment für kaufmännische Berufe [Factors on differential item functioning for a computer-assisted assessment for commercial professions]. Zeitschrift für Berufs- und Wirtschaftspädagogik 111(2):242–267
Ludwig-Mayerhofer W, Solga H, Leuze K, Dombrowski R, Künster R, Ebralidze E, Fehring G, Kühn S (2011) Vocational education and training and transitions into the labor market. Zeitschrift für Erziehungswissenschaft 14(2):251–266. doi:10.1007/s11618-011-0189-0
Lunz ME, Bergstrom BA (1994) An empirical study of computerized adaptive test administration conditions. J Educ Meas 31(3):251–263. doi:10.1111/j.1745-3984.1994.tb00446.x
Mangen A, Walgermo B, Bronnick K (2013) Reading linear texts on paper versus computer screen: effects on reading comprehension. Int J Educ Res 58:61–68
Masters GN (1982) A rasch model for partial credit scoring. Psychometrika 47(2):149–174. doi:10.1007/BF02296272
Mayer KU, Solga H (eds) (2008) Skill formation. Interdisciplinary and cross-national perspectives. Cambridge University Press, Cambridge
McDonald S, Edwards HM, Zhao T (2012) Exploring think-alouds in usability testing: the findings of an international survey. IEEE Pro Comm 55(1):2–19. doi:10.1109/TPC.2011.2182569
Nitko A, Brookhart S (2014) Educational assessment of students, 6th edn. Pearson, Boston
Oates T (2004) The role of outcome-based national qualifications in the development of an effective vocational education and training system: the case of England and Wales. Policy Futures Educ 2(1):53–71
Paek I (2002) Investigations of differential item functioning: comparisons among approaches, and extension to a multidimensional context. University of California, Berkeley
Pellegrino JW, Quellmalz ES (2010) Perspectives on the integration of technology and assessment. J Res Technol Educ 43(2):119–134. doi:10.1080/15391523.2010.10782565
Pellegrino JW, N Chudowsky, Glaser R (eds) (2003) Knowing what students know: the science and design of educational assessment. National Academy Press, Washington
Pohl S, Carstensen CH (2012) NEPS technical report—Scaling the data of the competence tests (NEPS Working Paper No. 14). Otto-Friedrich-Universität, Bamberg. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf. Accessed 1 Aug 2016
Pomplun M, Ritchie T, Custer M (2006) Factors in paper-and-pencil and computer reading score differences at the primary grades. Educ Assess 11(2):127–143
Quellmalz ES, Kozma R (2003) Designing assessments of learning with technology. Assess Educ 10(3):389–407
Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen
Rost J (2006) Item-Response-Theorie [Item-response-theory]. In: Petermann F, Eid M (eds) Handbuch der Psychologischen Diagnostik [Handbook of psychological diagnostics]. Hogrefe, Göttingen, pp 261–274
Roussos L, Stout W (1996) A multidimensionality-based DIF analysis paradigm. Appl Psychol Meas 20:355–371
Russell M, Goldberg A, O'Connor K (2003) Computer-based testing and validity: a look back into the future. Assess Educ Princ Policy Pract 10(3):279. doi:10.1080/0969594032000148145
Sangmeister J, Klotz VK, Winther E (2018) Technology-based competence assessment—designing high fidelity workplace simulations. In: Ifenthaler D (ed) Digital workplace learning. Bridging formal and informal learning with digital technologies. Springer, Berlin **(in preparation)**
Schumacher L (2002) Emotionale Befindlichkeit und Motive in Lerngruppen [Emotional state and motivation in learning groups]. Kovač, Hamburg
Seifried J, Sembill D (2005) Emotionale Befindlichkeit in Lehr-Lern-Prozessen in der beruflichen Bildung [Emotional state in learning processes in VET]. Zeitschrift für Pädagogik 51(5):656–672
Sembill D (1992) Problemlösefähigkeit, Handlungskompetenz und Emotionale Befindlichkeit. Zielgrößen Forschenden Lernens. [Problem-solving skills, empowerment and emotional state]. Hogrefe, Göttingen
Texas Education Agency (2008) A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests. Texas Education Agency (TEA) Technical Report Series
Wang H, Shin CD (2009) Computer-based and paper–pencil test comparability of studies. Meas Res Serv Bull 13(9):1–7
Wiggins G (1990) The case for authentic assessment. American Institutes for Research, Washington. http://files.eric.ed.gov/fulltext/ED328611.pdf. Accessed 1 Aug 2016
Wilson M (2005) Constructing measures: an item response theory approach. Lawrence Erlbaum, Mahwah. doi:10.1016/j.evalprogplan.2005.07.008

Winther E (2010) Kompetenzmessung in der beruflichen Bildung. [Measuring competence in VET]. Bertelsmann, Bielefeld. doi:10.3278/6004148w

Winther E, Achtenhagen F (2008) Konzeptuale Kompetenz und Selbstregulation als Grundlagen einer berufsbezogenen Kompetenzforschung [Conceptual competence and self-regulation as the basis of professional competency research]. In: Münk D, Gonon P, Breuer K, Deißinger T (eds) Modernisierung der Berufsbildung. Neue Forschungserträge und Perspektiven der Berufs- und Wirtschaftspädagogik [Modernization of VET. New research achievements and perspectives in vocational and economic education]. Budrich, Opladen, pp 100–110

Winther E, Achtenhagen F (2009) Measurement of vocational competencies—a contribution to an international large-scale assessment on vocational education and training. Empir Res Voc Educ Train 1:88–106

Winther E, Festner D, Sangmeister J, Klotz VK (2016a) Facing commercial competence: modeling domain-linked and domain-specific competence as key elements of vocational development. In: Wuttke E, Schumann S, Seifried J (eds) Economic competence and financial literacy of young adults: status and challenges. Budrich, Opladen, pp 149–164

Winther E, Seeber S, Festner D, Sangmeister J, Liedtke M (2016b) Large scale assessments in der kaufmännischen Berufsbildung—Das Unternehmensassessment ALUSIM (CoBALIT) [Large-scale assessments in commercial VET—the company assessment ALUSIM (CoBALIT)]. In: Beck K, Landenberger M, Oser F (eds) Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT [Technology-based measurement of competencies in VET—findings from the BMBF research initiative ASCOT]. Bertelsmann, Bielefeld, pp 55–74

Wright BD, Masters GN (1982) Rating scale analysis. MESA Press, Chicago

Wu ML, Adams RJ, Wilson MR, Haldane SA (2007) ACER ConQuest version 2.0. Generalised item response software. ACER Press, Camberwell

Yom M, Wilhelm T, Gauert S (2007) Protokolle Lauten Denkens und Site Covering [Thinkaloud protocols and site covering]. In: Buber R, Holzmüller HH (eds) Qualitative Marktforschung Konzepte-Methoden-Analysen [Qualitative market research concepts-methods-analyses]. Hogrefe, Göttingen, pp 637–652

Zinn B (2015) Conditional variables of Ausbildung 4.0—vocational education for the future. J Tech Educ 3(2):1–9

Zumbo BD (2007) Three generations of differential item functioning (DIF) analyses: considering where it has been, where it is now, and where it is going. Lang Assess Q 4:223–233. http://faculty.educ.ubc.ca/zumbo/papers/Zumbo_LAQ_reprint.pdf. Accessed 1 Aug 2016