

# Chapter 14

## Assessing Tomorrow's Potential: A Competence Measuring Approach in Vocational Education and Training

Viola Katharina Klotz and Esther Winther

**Abstract** Adequate measurement of action competence remains a central target of vocational education and training research; adequate measurement approaches in the vocational domain clearly are a prerequisite for accountable systems to authorize access to professional activities, as well as for future large-scale assessments. For the German Chamber of Commerce and Industry, competence assessments in the area of business and commerce rely mainly on final examinations that attempt to measure not just knowledge but also action competence. To evaluate and improve a test instrument, this chapter considers two questions: (1) how valid and reliable was the original test-format, and (2) how valid and reliable are the corresponding assessment results of a recently developed prototype? The study relies on statistical procedures (e.g., IRT scaling), applied empirically to a sample of 1768 final examinations of industrial managers in the original format, and to 479 industrial managers taking a prototype new format. The advanced prototype version appears as a more valid and accurate instrument to capture action competence. We conclude that several practical steps can be undertaken to improve current assessment practices in the area of business and commerce.

**Keywords** Vocational competence • Action competence • Assessment • Validity • Test reliability

---

V.K. Klotz (✉)  
University of Mannheim, Mannheim, Germany  
e-mail: [viola.klotz@bwl.uni-mannheim.de](mailto:viola.klotz@bwl.uni-mannheim.de)

E. Winther  
German Institute for Adult Education – Leibniz Centre for Lifelong Learning (DIE),  
Bonn, Germany  
e-mail: [winther@die-bonn.de](mailto:winther@die-bonn.de)

## 14.1 Background

### 14.1.1 *Prospects and Demand for Adequate Competence Assessments in Vocational Education*

Explicit or implicit measures of vocational competence are relevant to many facets of vocational education and training (VET), and thus constitute an ever-growing research field. They pertain to national educational factors, such as the relevant information and instruments for managing the quality of the vocational educational systems and developing adequate support programs, but increasingly, they also appear in international policy agendas (e.g., BMBF 2008). That is, international comparisons and the acknowledgement of qualifications, as well as the encouragement of lifelong, informal learning, require adequate measurement concepts and innovative evaluation methods. To meet these multiple expectations, two major conditions must be fulfilled a priori (Klotz and Winther 2012).

First, we require empirically confirmable competence models that encompass conceptual operationalizations of competencies but also reveal a well-postulated theoretical structure that captures their empirical structure. From a scientific perspective, researchers seek empirical results related to the “true” structure of professional competencies. From a political point of view, knowledge about the structure and comparability of competencies is required to achieve large-scale assessments of VET, such as across Europe. In this context, compulsory education likely refers to a common curriculum of basic competencies, such as literacy or numeracy, but the structure of competencies within VET is more varied in content and therefore tends to be more complex. Thus, VET content is heterogeneous not only between countries but also across different professions within nations (Baethge et al. 2009) and even in specific workplaces (Billett 2006). This abundant variation creates an ongoing dilemma in respect of the need to construct generally valid competence tests. Uncertainty about the structure of competencies also undermines international comparisons and the development of binding international agreements for consistent competence standards. Some (albeit scarce) empirical research into the appropriate structure or model of competence suggests a content-based classification, such that item content exerts a characteristic influence on the structure. Other studies assume dimensionality, based on different cognitive processing heuristics, which may determine response behaviors (Nickolaus 2011; Nickolaus et al. 2008; Rosendahl and Straka 2011; Seeber 2008; Winther and Achtenhagen 2009).

A second necessary condition pertains to the reliability of the test results—that is, the certainty with which we can classify students according to a chosen test instrument. Neglecting these conditions poses serious risks, because people can easily be misclassified on the basis of their test results, and such classification errors can have severe consequences for their future professional advancement—for example, in terms of admission requirements.

With this study, we seek to evaluate both necessary conditions with respect to the current testing efforts on the original final examinations—which were examined in

a former study (Klotz and Winther 2012; Winther and Klotz 2013)—but also on a newly developed assessment prototype within the research project “Competence-oriented assessments in VET and professional development”. Specifically, we first describe how the German VET system currently operationalizes and measures competencies in the economic domain. Empirical results obtained from a sample of 1768 final examinations of industrial managers<sup>1</sup> reveal the extent to which current German assessment instruments in the area of business and commerce are qualified, in terms of their validity and reliability, to measure and classify students’ economic action competence. We then describe our design criteria for a new prototype-version of final industrial examinations, and test this instrument on the empirical basis of 479 industrial managers. This study, in accordance with the SPP’s broader research program, therefore seeks to develop valid and reliable competence models and thereby to improve current assessment practices. The results offer guidelines for the design of the final examinations for industrial managers and possibly for assessment in the broader vocational sector of business and commerce.

### ***14.1.2 The Original Conceptualization of Final Examinations in the Area of Business and Commerce***

Action competence offers a constitutive element of the German vocational system, and has been a significant topic of scientific and political discourse since the early 1980s, particularly in relation to the didactic implications of action regulation theory (Hacker 1986; Kuhl 1994; Volpert 1983). In the mid-1990s, the Standing Conference of the Ministers of Education and Cultural Affairs (*Kultusministerkonferenz*) formally adopted the concept of action competence as a central target. Specifically, by law, students must be instructed in a way that enables them to *plan*, *execute*, and *monitor* an entire action process in a working environment. This concept appears largely heuristic, but still must form the foundation for any test construction (BBIG 2005; §5). In practice, these assessments come from the German Chamber of Commerce and Industry (GCCI) and comprise both oral and written components. The oral component consists of a presentation and then a related expert discussion; it accounts for 30 % of the assessment. The written component comprises practical tasks pertaining to economics and social studies (10 %), as well as commercial management and control (20 %). The last part of the examination contains situational tasks that take the form of case studies related to business processes. This last, business processes section, represents the most important assessment area, in terms of processing time (180 min) and weighting (40 % of the final grade). For this reason, this study focuses on this assessment component.

According to the GCCI (2009), the design of the business processes test component is intended to require test takers to model processes, undertake complex tasks,

---

<sup>1</sup>The data were acquired from six offices of the German Chamber of Commerce and Industry: Luneburg, Hanover, Frankfurt on Main, Munich, Saarland, and Nuremberg.

analyze business processes, and solve problems in an outcome- and customer-oriented way. To implement these goals, the test designers operationalized action competence as three mutually exclusive process dimensions: planning, executing, and monitoring (GCCCI 2009). Thus again, the business processes section seems particularly suitable for our empirical analysis of the structure of action competence.

If these process dimensions actually characterize a test situation, their solutions should require different sets of cognitive abilities in the test taker. This possibility was tested within an analysis of the structural validity of the original final examinations (Klotz and Winther 2012; Winther and Klotz 2013) on the empirical basis of  $N = 1768$  industrial managers. As a result, the structure of the assessment did not follow this postulated process-oriented operationalization, but instead appeared to be organized according to the four content domains of the assessment: *marketing and distribution*, *acquisition*, *human resource management (HRM)*, and *goods and services*. Such an alternative content-related model of competence measurement appears in some other vocational assessments (Nickolaus 2011; Rosendahl and Straka 2011; Seeber 2008). However, in the case of the final examinations of industrial managers, this solution appears disputable. The items depicting one dimension are often in close neighborhood and/or characterized by a common initial situation. The empirical solution of a content-related structure (root mean square error of approximation, RMSEA: .041; comparative fit index, CFI: .957; Tucker-Lewis index, TLI: .965) therefore does not necessarily represent cognitive structures, but might also be the mere consequence of the previous curriculum of commercial schools—which was officially abolished in 1996, and replaced by cross-disciplinary learning fields that sought to foster greater action competence by introducing the idea of process-orientation—or possibly even a relict of test sequence.

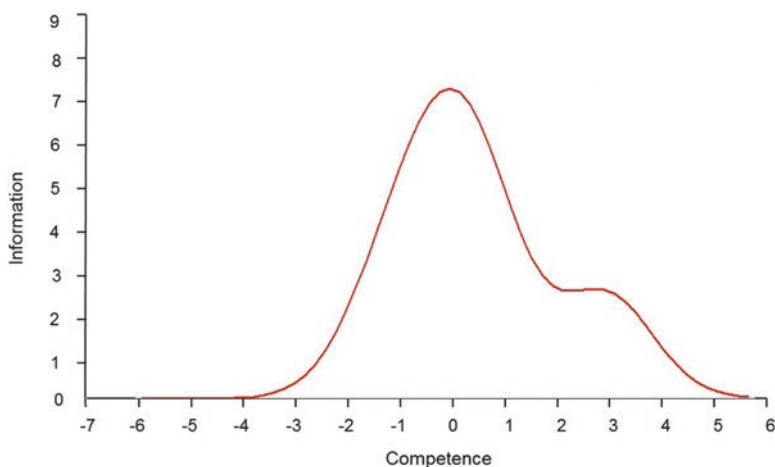
Besides the aspect of structural validity, we found infringements of the assessment's content validity in terms of content weighting (Winther 2011; Winther and Klotz 2013). As a final examination, the assessment should validly represent the commercial curriculum of industrial managers, which in turn should be based largely on real assignments in the workplace. With regard to content validity, a predominant part of the curriculum is dedicated to the *goods and services* domain (47 % of the curriculum and about one-third of practical training), and yet the proportion of content related to that topic in the original test was rather small (21 %). In particular, tasks related to modeling the processes of value creation and quantifiable production management are underrepresented, whereas the *marketing and distribution* content area appears overrepresented (38 % of the test), in relation both to percentage of the curriculum (26.7 %) and to practical relevance (25 %; see also Table 14.1).

In addition to these aspects of validity, the reliability of the original assessment was examined (Klotz and Winther 2012; see Fig. 14.1).

The information function for the test reaches its maximum for persons with an approximately average competence level. That is, near this area, it is possible to estimate, very precisely, test takers' true level of expertise (information = 7.4; reliability = .88). Further away from this maximum however, the test's estimation

**Table 14.1** Weighting of content

Content area	Prototype weighting (%)	Original test's weighting (%)	Curricular content weighting (%)	Practical learning (/25 months)
Marketing and distribution	25	38.00	26.67	5–7 months
Acquisition	18.33	20.00	13.33	5–7 months
Human resources	15	21.00	13.33	2–6 months
Goods and services	41.66	21.00	46.67	6–10 months



**Fig. 14.1** Information curve for the original test

precision decreases rapidly. Students of relatively high ability, who are located in the positive space, reveal a lower albeit still sufficient information value. In contrast, students with strongly below-average expertise are estimated with an information value tending to zero. The test provides many measurement items related to an average ability level, along with some items to measure high ability levels, but features few easy items, designed to measure low levels of expertise. Therefore, the GCCI final examination lacks power to effectively differentiate test takers of low versus very low ability. However, this fact does not necessarily cause problems. Some tests are constructed explicitly to differentiate students precisely at a specific, crucial point. That is, we need to consider the specific purpose of any particular test instrument to assess its reliability. The primary purpose of the final examinations is to regulate access to the industrial management profession, such that test takers are separated simply into those who pass the test, and thus receive certification to enter the professional community, and those who do not.

Annually, approximately 95 % of test takers pass the test, based on a norm-oriented test decision,<sup>2</sup> so the most important separation point must fall far below an average competence level. Yet the amount of test information available in this range tends toward zero. This lack of reliability in final examinations not only infringes on statistical test standards but also has severe implications for the professional development and life of a vast number of students. Considering that about 12,000 apprentices take this final examination<sup>3</sup> yearly, 600 test decisions, regulating access to the apprentice's targeted profession, are taken with low certainty and are therefore possibly false.

In summary, evaluation of the validity and reliability of the original assessment reveals that the test entails not the intended process-oriented structure but rather, a subject-specific content structure that reflects a previous, officially abolished teaching structure and curriculum. This makes it quite surprising that this conceptualization still dominates in the test. The empirical results pertaining to the structure of vocational competence are consistent with studies in other vocational areas that similarly suggest the high relevance of subject-related domains in the structuring of professional competence measures (e.g., Nickolaus et al. 2008; Seeber 2008). However, this approach seems unsatisfactory for measuring competence acquired in VET. In particular, on the basis of constructivist theory (Gijbels et al. 2006), a theory-based assessment design must capture students' skills in thinking and reasoning effectively, and in solving complex problems autonomously.

In terms of the test's reliability, it should be acknowledged, that the original test format yielded good reliability values for an average competence value. However, the items do not demonstrate reliability in their ability to show up rather low competence values. The low reliability in this crucial area limits accurate identification of failures. Therefore, some examinees may—possibly wrongly—be denied certain positions within the professional community and within society as a whole. The reliability of the GCCI test instrument thus could be improved in this crucial competence area.

### 14.1.3 *Assessment Model for Commercial Vocations*

In order to improve the current examination we designed a new foundational conceptualization of the assessment, following the subsequent construct, design standards and concrete implementation steps (Winther and Klotz 2013):

1. *Construct Definition: A Domain Model:* The design of an evidence-based assessment is always initiated by a theoretical model of a given construct (Mislevy and Haertel 2006; Wilson 2005). We adopted the modeling approach of Gelman and Greeno (1989), who suggest that “failure due to the absence of knowledge of a

---

<sup>2</sup>Acquired from GCCI statistics for Munich and Upper Bavaria.

<sup>3</sup>Acquired from GCCI statistics for Chemnitz.

principle should be distinguished from failure due to the lack of the domain-relevant knowledge” (p. 141). We believe that such a competence model, comprising both general competencies in the economic domain (domain-related) and specific competence components (domain-specific) at a first stage, better depicts the development and nature of commercial competence. We modeled items for both competence dimensions, focusing on work requirements in specific occupations, but with a varying degree of generalizability (Winther and Achtenhagen 2008; Winther and Achtenhagen 2009; Winther 2010). We further assumed, in line with findings in general education, the existence of a verbal and a numerical component of domain-related competence (e.g., National Educational Psychological Service–NEPS). From a didactic as well as from an empirical point of view, verbal and numerical domain-related components (numeracy, literacy) influence the formation of domain-specific vocational competence (e.g., Nickolaus and Norwig 2009; Lehmann and Seeber 2007). Such a separation might also prevail for domain-specific competence, as the commercial curricula entail both verbal and numerical abilities. Therefore, the two dimensions of domain-linked and domain-specific competence might, at a second stage, subdivide into a verbal and a numerical component respectively. These two considerations generate a four-dimensional structure consisting of domain-related economic literacy, domain-related economic numeracy, domain-specific verbal competence and domain-specific numerical competence, such as is depicted in Fig. 14.2.

From a developmental perspective, however, we assume that at the end of the vocational training the domain-specific and the domain-related dimensions could

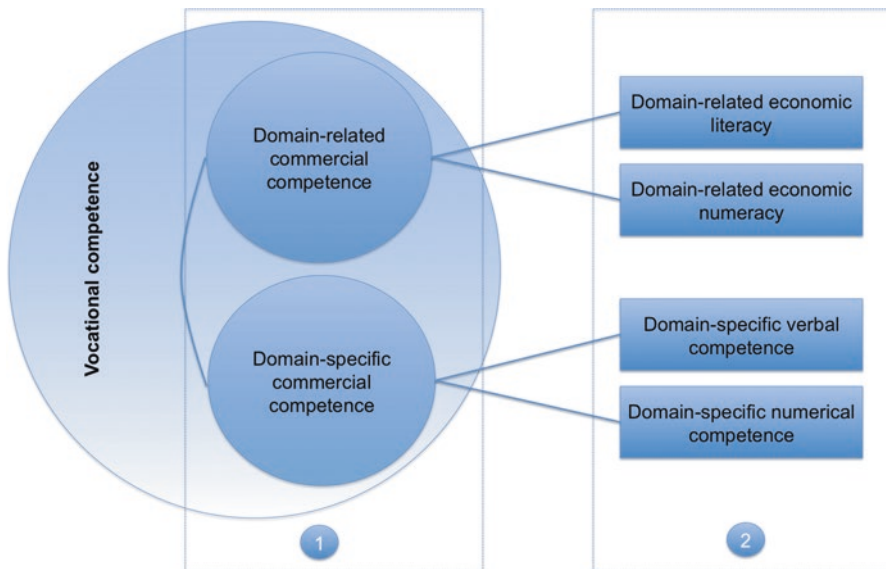


Fig. 14.2 Domain model of commercial competence

integrate into one dimension as the result of knowledge integration (Piaget 1971; Bransford et al. 1999). Knowledge integration should occur as the process of incorporating new information (domain-specific knowledge) into a body of existing knowledge (domain-related knowledge).

2. *Increasing Curricular Content Validity*: Within this general model of competence in the commercial domain, a crucial step within the item construction process was filling this model with concrete curricular-valid contents, focusing on work requirements in the specific occupational context of industrial managers. Here we designed, in accordance with our curricula analysis, more items pertaining to the *acquisition* and *goods and services* content areas, in order to better depict the vocational curriculum in terms of content weighting.
3. *Offering Sufficiently Complex Test Situations*: Recent commentary (e.g., Schmidt 2000; Winther 2010) suggests that the current test practices fail to give students sufficient room or potential to apply their knowledge to solve complex problems in a working context. We therefore, referring to the theoretical framework of Greeno et al. (1984), modeled items on three cognitive levels within our item design process:
  - Conceptual competence corresponds to factual knowledge as knowledge of facts and structures, which can be transmitted into action schemata;
  - Procedural competence subsumes the application of knowledge: that is, how to operate with facts, structures, knowledge nets and their corresponding elements;
  - Interpretative competence focuses on an interpretation of results and on decision processes.

Forming a vertical competence structure based on a cognitive construct map (Wilson 2005) to test different competence qualities was also intended to increase the interpretability of the IRT (item response theory) test scores (i.e., criterion-based assessment).

4. *Securing Adequate Vocational Authenticity*: Test tasks for vocational education are authentic if they model real-life situations (Shavelson and Semnara 1968; Achtenhagen and Weber 2003). We therefore designed a model company as a test setting on the basis of a real company, and within this company modeled realistic work situations and work tasks.
5. *Implementing the Concept of Process-Oriented Within Test-Design*: We implemented the concept of process-orientation (Hacker 1986) by stimulating company operations across departments and their specific economic interrelations. In our design, learners had to analyze certain problems across departments and to integrate preceding information on the operating work process. That is, they could not—with regard to information management—exclude the informational context given for the other items, within the operating process. For example, with regard to our sample sequence of an operating process, given in Appendix, learners had to anticipate that the cheapest sub-contractor for the acquisition department would not meet the goods and services department's production deadline. Also, the apprentices had to deduce information from foregoing



client-relation events. For example, if an offer had already expired at its acceptance date, no binding contract would be in place.

6. *Raising the Test's Reliability*: To appropriately assign learners into grade-categories, and to achieve greater accuracy at the most crucial separation point of the test, at a low competence level, we designed some rather difficult items and also some items targeting a low competence level.

By incorporating these guiding principles into the final examination, we aimed to render the assessment instrument more valid and reliable and to move it beyond the current focus on component skills and discrete bits of knowledge, to encompass theoretically sound aspects of student achievement (Pellegrino et al. 2001) and of vocational competence as a coherent and transgressional concept. Furthermore, such a test structure might offer more information about the level of competence students actually acquire, and concrete starting points for developing support measures to improve their learning process, as well as a more detailed view of the development of the apprentices' competence.

## 14.2 Method

### 14.2.1 Sample

We implemented the above guiding principles within 30 tasks for a new prototype final examination for industrial managers. The test took 125 min, including reading test instructions (10 min) and completing a context survey (10 min). Sample tasks of our instrument can be found in Appendix. We determined the tests' validity and reliability on an empirical basis of  $N = 479$  industrial managers who were assessed in March, April, October and November 2013 at four German vocational schools.<sup>4</sup> The sample consisted of 55 % women and 45 % men. The test takers were on average 21 years old.

### 14.2.2 Examination of Validity

Our evaluation of the validity criterion comprises two facets. First, it describes the operationalization of a theoretical concept, together with its potential subdimensions and observable indicators, to determine whether the focal approach offers a good notion of measurement in relation to the latent trait. It therefore entails the translation of the latent trait into contents, and then the contents into reasonable measurement items, and in this sense, it refers to *content validity* (Mislevy 2007). But even if an abstract concept is carefully operationalized, including all theoretical

---

<sup>4</sup>Munich, Hanover, Bielefeld and Paderborn.

aspects and a reasonable item design, it remains possible that the theoretical concept simply does not exist in the real world—or at least not in the way assumed by the researcher. Second, to address the potential gap between theory and observed reality, validity assessments entail testing *construct validity* to determine if the postulated theoretical structures arise from empirical test results (Embretson 1983; Mislevy 2007).

In order to ensure this first aspect of content validity, we first operationalized the vocational curriculum into content areas, then further into individual learning contents and, on the basis of this operationalization, developed test tasks. We then gave our developed test tasks to  $N = 24$  vocational experts (10 industrial teachers and 14 industrial staff managers) in order to ensure that our situated item setting, as well as the content of the developed items, modeled real-life, authentic situations (Achtenhagen and Weber 2003; Shavelson 2008). The experts had to rate on a five-point Likert scale whether the test tasks referred to realistic work assignments carried out in the occupational practice, and on what level of cognitive complexity they resided. These expert ratings formed an integral part of our test design: If the external criterion of authenticity in terms of workplace relevance was evaluated as low for an item, such items were withdrawn from the assessment.

Because competence, as measured by final examinations, seemingly constitutes a multidimensional concept, the confirmation of its structure requires a multidimensional modeling approach. To analyze construct validity, we used multidimensional item response theory (MIRT). We implemented this approach in Mplus (Muthén and Muthén 2010) and used 1PL Rasch modeling.

### 14.2.3 Examination of Reliability

The term “reliability” describes the replicability and thus the accuracy with which each item measures its intended trait (Kiplinger 2008). According to Fischer (1974), item precision can be depicted by item information curves (or functions), which indicate the range over the measurement construct in which the item discriminates best among individuals. The inverse of the squared standard measurement error is equivalent to item information with respect to the latent trait (in our case, vocational competence). If the information is expansive, it is possible to identify a test taker whose true ability is at that level, with reasonable precision. For this analysis, we again applied an IRT standard. An important characteristic of IRT models is that they describe reliability in terms of measurement precision as a continuous function that is conditional on the values of the measured construct. It is therefore possible to model the test’s reliability for each individual value of competence for every test taker (Hambleton and Russell 1993).

## 14.3 Results

### 14.3.1 Results for the Test's Validity

Our final weighting of test content was determined by relating the developed items back to the content domains of the vocational curriculum. We show the content weighting of each content area relative to all items of the test instrument, in Table 14.1.

Regarding the test's authenticity, in terms of the workplace relevance of the developed tasks, the items of the instrument achieved an average expert rating of workplace relevance four from five, indicating a "rather high" level of workplace authenticity. In terms of complexity, 38 % of the tasks were rated on a conceptual competence level, another 38 % on a procedural level and 24 % on an interpretative competence level.

After an analysis of the instrument's items, 2 tasks from 30 had to be removed, due to low separation ability, so that the instrument then comprised 28 tasks (7 for domain-related literacy, 11 for verbal domain-specific tasks, 7 for numeric domain-specific tasks and 3 for domain-related numeracy). In order to examine the construct validity of the prototype-version, we implemented, besides our theoretically assumed model (Model 6, depicted in Fig. 14.2, and here the second model stage), all alternative models (five possible combinations of lower dimensionality) and calculated the respective relative and absolute fit indices (see Table 14.2).

As the result of a relative consideration (Chi-Square difference testing), the theoretically-assumed four-dimensional structure fitted the data significantly better than the lower dimensional models. In terms of absolute fit, this model assumed a domain-related economic literacy component, a domain-related economic numeracy component, a domain-specific verbal competence component and a domain-specific numerical competence component, in which strong global model fit (RMSEA: .041; CFI: .931; TLI: .954) inhered. The four resulting dimensions correlate moderately to highly (Table 14.3).

**Table 14.2** Relative and absolute fit indices of the postulated and alternative models

Model	Parameter	df	Relative fit indices			Absolute fit indices		
			AIC	BIC	$\chi^2$	RMSEA	CFI	TLI
1	45	–	17,831	17,861	549,042	.075	.779	.848
2	47	2	17,805	17,836	493,386	.069	.810	.869
3	47	2	17,771	17,802	436,166	.063	.842	.892
4	50	3	17,761	17,795	452,960	.065	.832	.885
5	50	3	17,637	17,671	319,141	.048	.908	.938
6	54	4	17,591	17,627	277,329	.041	.931	.954

df degrees of freedom, AIC Akaike information criterion, BIC Bayes information criterion,  $\chi^2$  Chi-Square, RMSEA root mean square error of approximation, CFI comparative fit index, TLI Tucker-Lewis index

**Table 14.3** Correlations, variance ( $\sigma^2$ ) and reliability (based on EAP/PVs and WLEs) for model 6

Model 6	1.	2.	3.	4.	$\sigma^2$	EAP/PV	WLE
1. Domain-related literacy	1				0.92	.74	.47
2. Domain-specific verbal	.78***	1			1.01	.78	.67
3. Domain-specific numerical	.76***	.71***	1		0.96	.71	.50
4. Domain-related numeracy	.34***	.37***	.50***	1	1.29	.71	.45

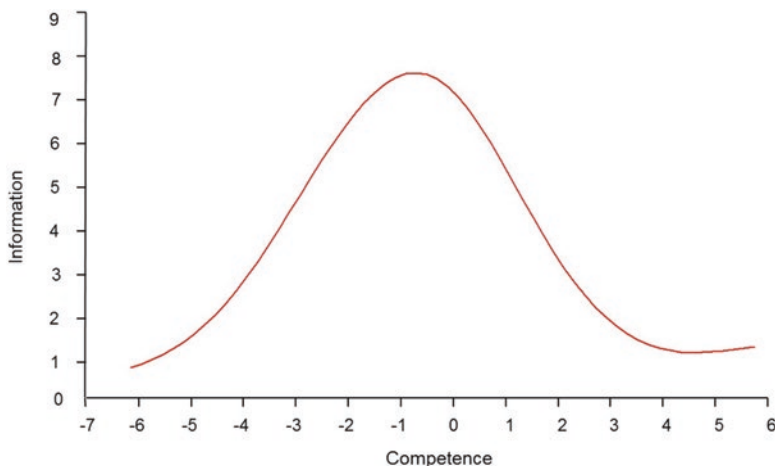
EAP/PV reliability based on expected a posteriori scores, WLE reliability based on weighted likelihood estimates, \*\*\* $p < .001$

Taking a closer look, the dimensions correlated strongly among the degree of specificity and verbal versus numerical access. It is further noteworthy that the domain-specific components correlate more strongly than do the domain-related components.

### 14.3.2 Results for the Test's Reliability

The test's overall WLE (weighted likelihood estimates) reliability was .826. However, due to only 28 items being used in the final instrument, given the restricted test time for the final examination, the values for the four-dimensional model were not sufficient to accurately depict each test taker on all of the four scales, as can be seen in Table 14.3. Only the EAP/PV (expected a posteriori scores) scale reliability, as a value of internal scale coherence, yielded sufficient values for all four scales. Using the IRT standard, we then computed the amount of information for each ability level for the developed prototype test, in order to compare it with that on the original GCCI test instrument. Here we used the one-dimensional model—not only because we had to, in order to make a comparison between the original and the prototype version—but also because we found it appropriate to enable us to make statements about how precisely students can be distinguished from one another through the use of this instrument, in respect of their final grading. This final test decision has to be made for the instrument as a whole. The resulting information function was generally increased in height and spread, and was characterized by an overall flatter gradient compared to the original test's function (see Fig. 14.3).

The information function for the prototype test reaches its maximum for persons with an approximately  $-.05$  competence value on the logit scale. That is, at this point, test takers' true level of expertise is estimated with high precision (reliability = .89). However, for this competence area the original instrument seemed just as good (reliability = .88). For students with relatively high ability (2 on the logit scale) the test still revealed a good informative value (reliability = .78). Even for the best student, with a value of 5.8 on the logit scale, the reliability still amounted to .50, compared to zero for the original test. For students with a rather low competence level ( $-2$  on the logit scale) their competence value was estimated with a reliability of .86 (compared to a reliability of .69 for the original instrument). And even for a



**Fig. 14.3** Information curve for the prototype test

very low competence value, constituting the crucial separation area of passing or failing the test, at about a logit of  $-4$ , a reasonable reliability value of  $.69$  was obtained, compared to a value of zero in the original test format.

## 14.4 Discussion

Regarding validity, we examined two concepts: (1) the translation of the latent trait into contents, as well as the resulting contents into reasonable measurement items (*content validity*), and (2) the potential gap between the assumed theoretical content structure and observed reality (*construct validity*). Regarding content validity, we adapted the test assembly in such a way that the final weighting of the test content now related more adequately to the amount of content weight within the vocational curriculum, compared to the original version. Further, a more salient distribution of the test items over the taxonomy of cognitive complexity was implemented within the test design and then confirmed by expert ratings. Finally, the expert ratings also functioned as a critical counterpoint within the assessment design. Within the test assembly process (“assembly model”; Mislevy and Riconscente 2005) the rating of workplace authenticity formed a crucial criterion for the final item selection; that is, that only items with an above average rating were taken into the final test. The final degree of authenticity of the instrument in terms of workplace relevance is satisfactory, but possibly could be further improved by a second round of item modeling and expert selection.

Regarding construct validity, the comparison of relative model fit, as well as Chi-Square difference testing, suggests a four-dimensional structure, comprising a domain-related economic literacy component, a domain-related economic numeracy

component, a domain-specific verbal competence component and a domain-specific numerical competence component. The correlations between the resulting dimensions suggest that the structures are sufficiently divergent in terms of discriminant validity (see Table 14.3). However, in terms of an absolute model fit, a model suggesting a three-dimensional structure consisting of a domain-specific component, a numeracy and a literacy component, such as that suggested by Winther (2010) already attains a sufficient global model fit (RMSEA: .048; CFI: .908; TLI: .938). This is due to the higher correlation of numerical and verbal aspects for domain-specific competence ( $r = .71$ ;  $p < .001$ ) than for domain-related competence ( $r = .34$ ;  $p < .001$ ). It seems that, with an increasing degree of vocational specificity, the importance of the distinction of numerical versus verbal access decreases appreciably, supporting the idea of the integration of numerical and verbal knowledge aspects in specific vocational abilities.

However, the integration of domain-related and domain-specific competence at the end of the vocational training, in the sense of a total integration into one dimension, like that suggested by Winther (2010), cannot be found within our data. It is imaginable that this integration takes place at a later developmental stage, with an increasing degree of vocational experience and routine. Or we may have to acknowledge that there is no absolute integration of domain-related and domain-specific competence dimensions, and that the two competence dimensions are indeed related (correlation of domain-specific and domain-related competence within a two-dimensional model:  $r = .77$ ;  $p < .001$ ) but remain separate dimensions in terms of dimensionality over the vocational trajectory.

Regarding the test's reliability, we designed the assessment instrument explicitly in regard to the specific purpose of the final examinations. That is, first of all, to differentiate students precisely at the most crucial point of separation, of passing or failing the test and therefore being granted or denied access to the vocational community as a full member. Second, to allow for a signaling function for future employers in the vocational final assessment, in terms of a dependable grading (Weiß 2011). We therefore designed more items targeting a low competence level and also some more difficult items, in order to also differentiate precisely for a progressed level of competence.

The obtained information curve suggests that the prototype examination is capable of a precise measurement of an around average ability—similarly to the original instrument. It also effectively differentiates test takers of low versus very low ability and test takers with high versus very high ability, and therefore measures precisely along the logit scale and visibly adds significant value, compared to the original instrument. However, this only applies to the test instrument as a whole. Accuracy at an individual diagnostic level was not reached for each of the four dimensions separately. We conclude therefore that the desired function of the new prototype

examination of classifying students, can be administered with an adequate degree of certainty only for a one-dimensional model and not for the postulated four-dimensional structure, within restricted test times.

## 14.5 Conclusions

Our research endeavor focusing explicitly on the improvement of current assessment practice, illustrates that the identification of theoretically sound and empirically confirmable structures and reliability is not intended as a question of statistical test esthetics, but is a necessary prerequisite of school policy and assessment as they move towards an evidence-based practice (Slavin 2002), in turn optimizing educational processes and educational decisions (Koeppen et al. 2008).

First, evaluation of the validity of the original final examinations, provided by a former study within our research project, reveals that the criteria of content validity were not completely adhered to. Furthermore, the analyzed GCCI assessment entailed not the intended, process-oriented structure but rather a content-related structure. Finally, with regard to the accuracy with which the final examination distinguishes and classifies students, the test did not provide enough items to measure below-average competence levels accurately.

In order to improve the final examination of commercial competence for industrial clerks, we designed a new foundational conceptualization of the assessment, following the idea of an evidence-based assessment design, including a careful construct operationalization, a reviewed item design process and an extensive empirical checkup on the obtained data, in order to draw inferences about students' knowledge and skills (Mislevy and Riconscente 2005). Our results suggest that the developed prototype version of the final examinations can capture students' skills in thinking and reasoning effectively and in solving complex problems (Pellegrino et al. 2001) in a more valid and also precise way: The items are adequate in terms of their content (curricular validity, complexity, authenticity) and in terms of their intended structural validity (construct validity). The instrument furthermore demonstrates reliability in its ability to differentiate adequately among students and to assign them to classes with a sufficient degree of certainty, as a prerequisite for fair opportunities to attain certain positions within the professional community and within society as a whole.

**Acknowledgments** The preparation of this chapter was supported by grant "Competence-oriented assessments in VET and professional development" (Winther, 2009-2014; Wi 3597/1-1; 1-2) from the German Research Foundation (DFG) in the Priority Program "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" (SPP 1293).

## Appendix

### *Ceraforma Keramik AG*



Since its foundation in 1982, Ceraforma Keramik AG has developed into an expanding and globally active industrial enterprise, having their head office in Aachen, Germany. The company is involved in the production of ceramic goods, such as china and porcelain for tableware and vases or sanitary ware.

In the past, the management of Ceraforma Keramik realized that the four divisions: procurement logistics, production, human resource management as well as marketing and sales, were operating too independently of each other, which caused disturbances in the performance process and led to customer complaints. In response to these problems, so-called *horizontal teams* were established, consisting of work members from different company divisions.

You have been employed with Ceraforma Keramik in such a horizontal team since the beginning of this year. Here, the allocated customer orders are being handled in all business processes, ranging from the receipt of orders to the settlement of accounts. Ms. Kenk, the team leader, Mr. Friebe and Ms. Hoffmann, the new trainee, are your colleagues in the horizontal team.





## Business Process 1

### Situation

Your team just received a new customer enquiry. Your colleague, Mr. Friebel, shows you the following e-mail, which arrived on 30 March 20... at 10:17.

An...	info@ceraforma.de;
Cc...	
Bcc...	
Betreff:	Waschbecken der Reihe "Swing"

Dear Sir/Madam,

Whilst seeking manufacturers of ceramic goods, we have come across your company, which has attracted our attention.

We are DIY retailers and our head office is in Hannover. We are looking for a supplier for sanitary ceramics and are especially interested in the washbasin in your design series „Swing“.

We would appreciate receiving your corresponding quotation for 2,400 pieces, your soonest delivery date and terms of delivery at your earliest convenience.

Sincerely yours,

Karl Schwiene´  
Head of Procurement

---

Bauhannes GmbH  
Junkersstraße 8  
30179 Hannover  
eMail: [karl.schwiener@bauhannes.de](mailto:karl.schwiener@bauhannes.de)  
Tel.: +49 (0)511-123321  
Fax: +49 (0)511-456654  
Mobil: +49 (0)176-123654

---

Amtsgericht Hannover, HRB 1234, Geschäftsführer:  
Dr. Konrad Kluge, Aufsichtsratsvorsitz: Emanuel Windig  
Umsatzst.-Id: DE123456789

1.1 Since there have not yet been any business relations with the Bauhannes Ltd. company, you are requested by Mr. FriebeI to gather detailed information on the financial standing of the potential customer.

Which two kinds of information would you gather to assess the risk and which two outside sources would you contact?

---

---

---

---


---

---

---

---

1.4 After repeated negotiations the company Ceraforma accepts the order from the DIY Bauhannes at the price stipulated by Mr. Schwienert. Receipt of confirmation of the order by email is on Friday, 6 April 20... You have been informed that there is no sufficient quantity of quartz crystal on stock to execute the order. You are therefore required to order 25 tons of new quartz crystals. You then contact various suppliers by mail and you receive the emails below from Mineral Seifert AG from Aachen, and Tam-Quarz Ltd. from South Africa:

	An:	horizontalteam3@ceraforma.de
	Kopie:	
	Blindkopie:	
	Betreff:	Unser Angebot für Sie

Lieber Herr Friebel,

wir freuen uns, dass wir Sie wieder einmal von unseren Produkten und Leistungen überzeugen konnten.


Aufgrund unser langjährigen Geschäftsbeziehungen, können wir Ihnen zu Ihrer Anfrage folgende Konditionen anbieten:

Produkt: reiner Quarz, Bergkristall, weiß  
 Preis/Menge: 500,00 EUR/t inkl. MwSt  
 Zahlungsbedingungen: 10 Tage 3 % Skonto; 60 Tage netto Kasse  
 Bezugskosten: 100,00 EUR pauschal/Lieferung  
 Lieferzeit: 3 Werktage ab Bestelleingang  
 Angebot ist gültig: bis zum 15.04.20..

Wir würde uns freuen, wenn Sie sich erneut für unsere Produkte und unseren Service entscheiden würden.

Einen schönen Tag noch und freundliche Grüße

Jörg Schewe  
 Vertrieb  
 Mineral Geifert AG Aachen

	An:	horizontalteam3@ceraforma.de
	Kopie:	
	Blindkopie:	
	Betreff:	RE: Angebotsanfrage

Dear Sir/ Madam,

In reply to your enquiry dated xy.20.. we are pleased to make the following offer:

- white mountain quartz crystal: € 450.00/ ton
- minimum order quantity: 10 tons
- shipping charges: € 13.00/ 100kg

Since this is your first order, we allow a quantity discount of 3% per ton for orders exceeding 30 tons.

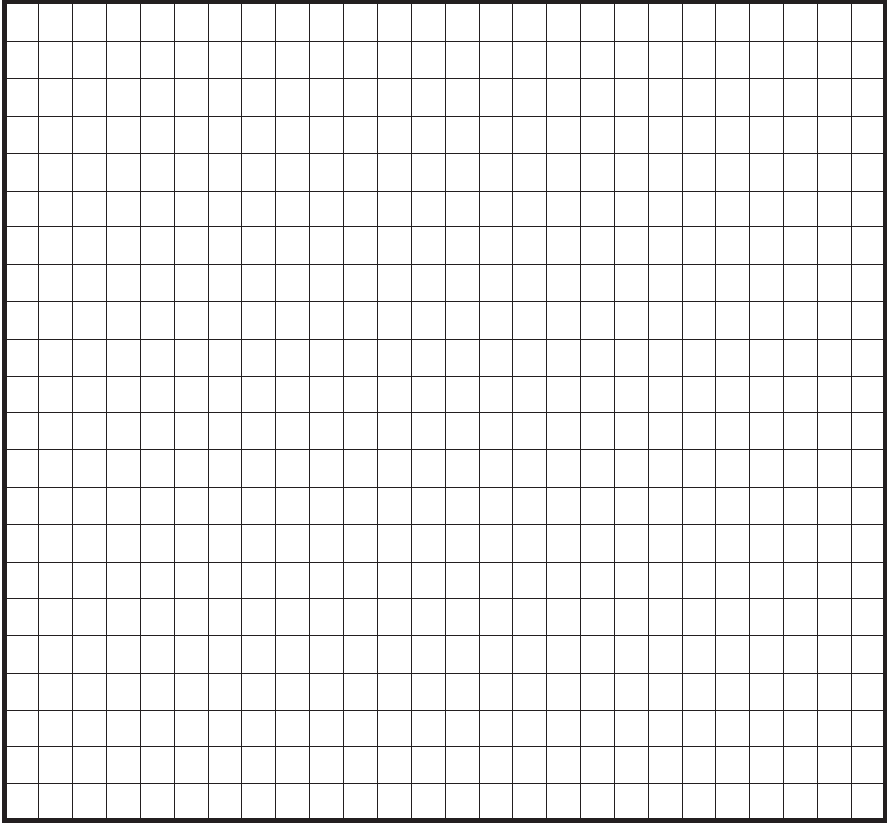
This offer is valid until April 15, 20.. Please consider that transportation by ship might takes up to a month.

We look forward to hearing from you soon!

Sincerely, (alternative: Yours sincerely/ Kind regards)

J. Stones  
 Tam-Quarz, South Africa  
 Mail: stoness@tam.za

Please compare both offers and give reasons for which offer you would decide. When making your decision you should consider also possible risks and social and ecological issues, besides financial aspects. Also bear in mind that Ceraforma have sufficient liquid funds and that discounts granted can be fully exploited.



---

---

---

---

---

## References

- Achtenhagen, F., & Weber, S. (2003). "Authentizität" in der Gestaltung beruflicher Lernumgebung ["Authenticity" in the design of vocational learning environments]. In A. Bredow, R. Dobischat, & J. Rottmann (Eds.), *Berufs- und Wirtschaftspädagogik von A-Z. Grundlagen, Kernfragen und Perspektiven: Festschrift für Günter Kutscha* (pp. 185–199). Baltmannsweiler: Schneider.
- Baethge, M., Arends, L., & Winther, E. (2009). International large-scale assessment on vocational and occupational education and training. In F. Oser, U. Renold, E. G. John, E. Winther, & S. Weber (Eds.), *VET boost: Towards a theory of professional competences: Essays in honor of Frank Achtenhagen* (pp. 3–24). Rotterdam: Sense.
- BBIG (Berufsbildungsgesetz) 1. (2005, April). Retrieved from [http://www.gesetze-im-internet.de/bundesrecht/bbig\\_2005/gesamt.pdf](http://www.gesetze-im-internet.de/bundesrecht/bbig_2005/gesamt.pdf).
- Billett, S. (2006). *Work, change and workers*. Dordrecht: Springer.
- BMBF (Bundesministerium für Bildung und Forschung). (2008). *Framework programme for the promotion of empirical educational research*. Bonn: Author.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. New York: National Academy of Sciences.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197. doi:10.1037/0033-2909.93.1.179.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [An introduction to the theory of psychological tests]. Bern: Huber.
- GCCI (German Chamber of Commerce and Industry), & Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen (AKA). (Eds.). (2009). *Prüfungskatalog für die IHK-Abschlussprüfungen* [Test catalog for the GCCI's final examinations]. Nürnberg: AKA.
- Gelman, R., & Greeno, J. G. (1989). On the nature of competence: Principles for understanding in a domain. In L. B. Resnick (Ed.), *Knowing and learning: Essays in honor of Robert Glaser* (pp. 125–186). Hillsdale: Erlbaum.
- Gijbels, D., Van De Watering, G., Dochy, F., & Van Den Bossche, P. (2006). New learning environments and constructivism: The students' perspective. *Instructional Science*, *34*, 213–226. doi:10.1007/s11251-005-3347-z.
- Greeno, J. G., Riley, M. S., & Gelman, R. (1984). Conceptual competence and children's counting. *Cognitive Psychology*, *16*, 94–143. doi:10.1016/0010-0285(84)90005-7.
- Hacker, W. (1986). *Arbeitspsychologie. Psychische Regulation von Arbeitstätigkeiten* [Action-regulation-theory. Psychological regulation of occupational actions]. Bern: Huber.
- Hambleton, R. K., & Russell, W. J. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement*, *12*(3), 38–47. doi:10.1111/j.1745-3992.1993.tb00543.x.
- Kiplinger, L. (2008). Reliability of large scale assessment and accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 93–113). New York: Routledge.
- Klotz, V. K., & Winther, E. (2012). Kompetenzmessung in der kaufmännischen Berufsausbildung: Zwischen Prozessorientierung und Fachbezug [Competence measurement in commercial vocational training: Between processual and content-related perspectives]. *Bwp@ Berufs- und Wirtschaftspädagogik —Online*, *22*.
- Koepfen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Journal of Psychology*, *216*(2), 61–73.
- Kuhl, J. (1994). A theory of action and state orientation. In J. Kuhl & J. Beckmann (Eds.), *Volition and personality: Action vs. state orientation* (pp. 97–129). Seattle: Hogrefe.
- Lehmann, R., & Seeber, S. (Eds.). (2007). *ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen* [ULME III. Examination of performance, motivation and attitudes of students at the end of their vocational training]. Hamburg: HIBB.

- Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, 36, 463–469. doi:10.3102/0013189X07311660.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement*, 25(4), 6–20. doi:10.1111/j.1745-3992.2006.00075.x.
- Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical report 9). Menlo Park: SRI International.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles: Author.
- Nickolaus, R. (2011). Die Erfassung fachlicher Kompetenz und ihrer Entwicklungen in der beruflichen Bildung: Forschungsstand und Perspektiven [Assessing professional expertise and its development: Current state of research and future perspectives]. In O. Zlatkin-Troitschanskaia (Ed.), *Stationen empirischer Bildungsforschung: Traditionslinien und Perspektiven* (pp. 331–351). Wiesbaden: Springer.
- Nickolaus, R., & Norwig, K. (2009). Mathematische Kompetenzen von Auszubildenden und ihre Relevanz für die Entwicklung der Fachkompetenz: Ein Überblick zum Forschungsstand [Mathematical competences of apprentices and their relevance for the development of vocational expertise: An overview regarding the current state of research]. In A. Heinze & M. Grüßing (Eds.), *Mathematiklernen vom Kindergarten bis zum Studium. Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht* (pp. 204–216). Münster: Waxmann.
- Nickolaus, R., Gschwendter, T., & Geißel, B. (2008). Modellierung und Entwicklung beruflicher Fachkompetenz in der gewerblich-technischen Erstausbildung [Assessing vocational expertise and its development over commercial-technical initial trainings]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 104, 48–73.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Piaget, J. (1971). *Biology and knowledge; an essay on the relations between organic regulations and cognitive processes*. Chicago: University of Chicago Press.
- Rosendahl, J., & Straka, G. A. (2011). Kompetenzmodellierungen zur wirtschaftlichen Fachkompetenz angehender Bankkaufleute [Modeling commercial expertise of beginning bankers]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 107, 190–217.
- Schmidt, J. U. (2000). Prüfungen auf dem Prüfstand: Betriebe beurteilen die Aussagekraft von Prüfungen [Examining final examinations: Firms evaluate the explanatory power of final examinations]. *Berufsbildung in Wissenschaft und Praxis*, 29(5), 27–31.
- Seeber, S. (2008). Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen [Approaches for the modeling of vocational expertise in commercial vocations]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 104, 74–97.
- Shavelson, R. J. (2008). Reflections on quantitative reasoning: An assessment perspective. In B. L. Madison & L. A. Steen (Eds.), *Calculation vs. context: Quantitative literacy and its implications for teacher education* (pp. 27–47). Washington, DC: MAA.
- Shavelson, R. J., & Semnara, J. L. (1968). Effect of lunar gravity on man's performance of basic maintenance tasks. *Journal of Applied Psychology*, 52, 177–183.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15–21. doi:10.3102/0013189X031007015.
- Volpert, W. (1983). *Handlungsstrukturanalyse als Beitrag zur Qualifikationsforschung* [Analysis of the structure of actions as a contribution to qualification research]. Köln: Pahl-Rugenstein.
- Weiß, R. (2011). Prüfungen in der beruflichen Bildung: Ein vernachlässigter Forschungsgegenstand [Examinations in vocational education: A neglected field of research]. In E. Severing & R. Weiß (Eds.), *Prüfungen und Zertifizierung in der beruflichen Bildung: Anforderungen–Instrumente–Forschungsbedarf* (pp. 37–52). Bielefeld: Bertelsmann.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah: Erlbaum.
- Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung* [Competence measurement in vocational education]. Bielefeld: Bertelsmann.

- Winther, E. (2011). Kompetenzorientierte Assessments in der beruflichen Bildung: Am Beispiel der Ausbildung von Industriekaufleuten [Competence-oriented assessments in vocational education]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, *107*, 33–54. doi:[10.1186/1877-6345-5-2](https://doi.org/10.1186/1877-6345-5-2).
- Winther, E., & Achtenhagen, F. (2008). Kompetenzstrukturmodell für die kaufmännische Bildung. Adaptierbare Forschungslinien und theoretische Ausgestaltung [A structural competence model for commercial education]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, *104*, 511–538.
- Winther, E., & Achtenhagen, F. (2009). Measurement of vocational competencies—A contribution to an international large-scale assessment on vocational education and training. *Empirical Research in Vocational Education and Training*, *1*, 88–106.
- Winther, E., & Klotz, V. K. (2013). Measurement of vocational competences: An analysis of the structure and reliability of current assessment practices in economic domains. *Empirical Research in Vocational Education & Training*, *5*(2), 1–12. doi:[10.1186/1877-6345-5-2](https://doi.org/10.1186/1877-6345-5-2).