



Instructional sensitivity in vocational education



Viola Deutscher ^{a,*}, Esther Winther ^b

^a University of Mannheim, Business School, L4 1, 68161 Mannheim, Germany

^b German Institute for Adult Education (DIE), Heinemannstraße 12-14, 53175 Bonn, Germany

ARTICLE INFO

Article history:

Received 10 November 2016

Received in revised form

30 June 2017

Accepted 7 July 2017

Available online 8 September 2017

Keywords:

Instructional sensitivity

Item differential functioning (DIF)

Vocational educational training (VET)

Competence development

Competence-based assessment

ABSTRACT

Apprentices' performance after vocational educational training (VET) is commonly attributed to the effectiveness of the training. This implies the assumption that learners' development of vocational knowledge and ability is significantly affected by vocational instruction. However, the few analyses that have been made of instructional sensitivity within the general school-based educational system, have in most cases shown little or no effect of instruction (time in school) on performance in assessments. The question as to whether, and to what extent, VET in adult education is effective (in the sense that it fosters the development of vocational knowledge and ability), as well as the related question—whether we are able to track the resulting learning progress with adequate measures (i.e., assessments)—has hardly been investigated. In the present study, we propose modeling of instructional sensitivity via differential item functioning (DIF), and apply this method to a sample of $n = 534$ apprentices. We find that during vocational instruction, apprentices significantly improved their performance in an assessment of vocational knowledge and ability, and that we were able to track these changes in the quality of their abilities over the span of a three year initial VET program: that is, the first program of vocational study in which apprentices become qualified to work in a given trade. Moreover, with this proposed method, it is possible to identify items that are particularly sensitive to instruction and that appear therefore to be amenable to the future development of vocational assessments.

© 2017 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Premise

Schooling/training is commonly assumed to be responsible for learning (Burstein, 1989; Naumann, Hochweber, & Hartig, 2014). Somewhat surprising therefore are some empirical hints that performance on assessments in general education is often little or not at all sensitive to the effects of instruction. Diverse research (e.g., Chen, 2012; Court, 2013; Pham, 2009; Phillips & Mehrens, 1988; Popham, 2007; Popham & Ryan, 2012) suggests that many achievement tests fail to effectively reflect whether students successfully receive and absorb curricular content during instruction. This apparent paradox might result from one of two causes (or conceivably both): (1) That learners have indeed learned during instruction, but that the assessment applied was not able to capture the learning progress made. For example, Goe (2007) and Polikoff (2010) caution that the failure to detect instructional sensitivity does not necessarily imply that no learning progress has been

made. Rather, the weak relationship between curricular instruction and student performance could be due to the applied measurement tools not being sufficiently sensitive to capture the effect of instruction. These measures of learning outcomes possibly indicate what students know, but not necessarily what they learn during instruction (Popham, 2007).

The second possible cause (2) is expressed by Wiliam (2007, 12) who, providing an insightful analysis of the relevant research addressing instructional sensitivity, goes one step further, arguing for a more pessimistic second order explanation:

the fundamental issue is not that tests are insensitive to instruction; it is that achievement is insensitive to instruction. Put bluntly, most of what happens in classrooms doesn't change what students know very much, especially when we measure deep, as opposed to surface aspects of a subject.

This second explanation in turn might result from two causes: Either students' knowledge as a latent structure is generally insensitive to instruction, or instruction may not have been delivered (or not effectively).

* Corresponding author.

E-mail address: viola.klotz@bwl.uni-mannheim.de (V. Deutscher).

Even without any clear indication of which of the two explanations (or, conceivably, a combination) accounts for the empirical findings, both interpretations of the instructional *insensitivity* of diverse outcome measures pose a severe threat—especially to educational accountability. In some nations (e.g., the US), outcome measures have been used in recent times not only to evaluate the effectiveness of schools and teachers on the basis of their students' test proficiency, but also to allocate educational resources on the basis of test results (e.g., state tests used for the purposes of the No Child Left Behind Act). Without a doubt, an accountability test would—as one prerequisite, among other aspects of validity—at least have to be instructionally sensitive, in order to form an appropriate basis for making decisions with potentially far-reaching consequences. However, given unreliable and possibly inaccurate test-based evidence, achievement or learning progress, or even the lack thereof, instructional sensitivity cannot be accurately determined; this leaves the danger that teachers and schools will be misjudged, and even be unfairly denied resources.

Considering these potentially severe consequences, Polikoff (2010, 34), summarizing the overall state of instructional sensitivity research, comes to the conclusion that the lack of documentation of instructional sensitivity in accountability tests constitutes a “grievous oversight”. Even more strongly, Popham and Ryan (2012, 2) assail the current lack of empirical evidence regarding instructional sensitivity in most educational tests, describing it as an “intolerable state of affairs”. In view of the above, the internationally observable trend towards test-based accountability systems, and political reliance on outcome measures in making decisions affecting education, seems highly questionable. For this reason, some authors have demanded that the concept of instructional sensitivity become an explicit and integral part of a broadened conception of validity, for common standards in educational and psychological testing (e.g., AERA, APA, & NCME, 1999). They call for this to be applied at least for the outcome measures that are used to assess changes in learning and for those testing system effectiveness (e.g., teacher or school effectiveness; for example, Polikoff, 2010; Popham & Ryan, 2012).

Way (2014, 4) raises the concern that “despite these recent imperatives for explicitly making assessments instructionally sensitive, there is not agreement about how this is to be done (...)” Naumann et al. (2014) similarly believe that the question whether outcome measures are indeed sensitive to instruction is hardly empirically engaged, due to the lack of a commonly accepted definition and operationalization of the concept of instructional sensitivity. The methodological approaches to modeling instructional sensitivity are diverse, to say the least: this has led to mainly psychometric papers on the topic, and few practical applications combining the proposed methods with a didactical perspective (for one such application however, see the recent study by Naumann et al., 2014).

Although, as we have noted, instructional sensitivity is a crucial concept in instructional science, to our knowledge no studies have addressed the modeling of instructional sensitivity with respect to vocational education of adults. In Germany, about half of the population takes vocational educational training (VET) rather than academic training, after their school education. Most of this VET (60%) relates to commercial professions: for bankers, industrial management assistants, salesmen (National Educational Report, Hasselhorn et al., 2014). While development of measures of vocational knowledge and ability for this branch of education is very relevant, it is still in its infancy. In general, however, significant progress has been made in the last decade with respect to the measurement of learning outcomes in the vocational domains of auto mechanics (e.g., Nickolaus, Lazar, & Norwig, 2012) and apprenticeships in commercial professions: for example, industrial or

logistics apprentices (e.g., Klotz, Winther, & Festner, 2015; Rausch, Seifried, Wuttke, Kögler, & Brandt, 2016; Seeber, 2008; Weber et al., 2016; Winther & Achtenhagen, 2009). More recently, there has also been notable progress in the area of social health care (e.g., Seeber, 2015; Seeber, Ketschau, & Rüter, 2016). Therefore, the purpose of this study is to conceptualize and model instructional sensitivity in the area of vocational education, and to detect which item types are especially relevant to modeling the learning progress. More precisely, we focus on the occupation of industrial management assistant, and seek to explore whether instructional sensitivity is detectable in an assessment of vocational knowledge and ability.

According to Polikoff (2010, 8–9), it is impossible to say

whether a finding of low or no sensitivity in any particular study is due to a poor-quality test that is actually insensitive to instruction or to poor quality instruction, so that the test results actually reflect the instruction received by students. In contrast, a finding of high sensitivity indicates both effective instruction and also a high-quality assessment sensitive to that instruction. Clearly, the goal is always to have instruction of maximum effectiveness, and to design a test to capture the effects of instruction.

So if we do not find instructional sensitivity, this does not necessarily mean that learners have not learned anything (e.g., due to poor instruction); it may possibly mean that our assessment failed to capture their learning (i.e., instructional *insensitivity* of the assessment). However, if we find instructionally sensitive items, this must mean that vocational knowledge and ability are being acquired during VET and that we are able to capture them. More precisely, in this study, the following research questions are addressed:

1. Is the developed assessment of vocational knowledge and ability sensitive to instruction (meaning that learning progress is made during VET and that we are able to capture that progress)?
2. Is the learning of specific (vocational) knowledge and ability equally sensitive to instruction as is the learning of generic knowledge and ability?

In order to explore this matter, the paper begins by reviewing different definitions of instructional sensitivity and different methodological approaches to its detection. Subsequently, the item and test design of an instrument to capture apprentices' knowledge and ability is introduced. We then apply the IRT-DIF approach to a vocational sample of $n = 877$ industrial apprentices, and outline and discuss the results.

2. Defining and detecting instructional sensitivity

In the theoretical research into instructional sensitivity, this term has often been used interchangeably with “instructional validity”, with both terms being treated as subfacets of other, common aspects of test validity, such as curricular validity and content validity (Polikoff, 2010). Li et al. (2012b, p. 2) note that the intended meaning of the term sometimes relates exclusively to the extent to which the curriculum content is taught successfully (e.g., Linn, 1983). Occasionally however, it also includes the nature of the teaching of the content (e.g., Burstein, Aschbacher, Chen, & Lin, 1990; Popham & Ryan, 2012; Yoon & Resnick, 1998). A definition that is open to both interpretations is the originally used, more technical definition of Haladyna and Roid (1981, p. 40), defining instructional sensitivity as “the tendency for an item to vary in difficulty as a function of instruction”. This relation is then specified

either by the duration of instruction only (Opportunity to Learn [OTL] as time for learning; see, e.g. Yu, Lei, & Suen, 2006) or by aspects referring to the quality, content and nature of instruction, such as is often implemented in broader OTL conceptions (e.g., Kao, 1990; Switzer, 1993; Yu et al., 2006). In this study we adopt the broader approach, as we focus on the effectiveness of VET and its assessment as a whole, and therefore define instructional sensitivity as the *tendency for a test or a single item to vary in difficulty as a function of the duration of vocational educational training*. According to this definition, if vocational instruction is reasonably effective, items should be easier for instructed students and more difficult when administered to uninstructed students. Conversely, if time on training does not change apprentices' performance on an assessment to any marked degree, then that assessment must be insensitive to instruction (Ruiz-Primo et al., 2012; Wiliam, 2007).

In the literature on detecting instructional sensitivity, a variety of approaches are distinguishable, but they can be subsumed under two basic headings: (1) Judgmental approaches are usually integrated into test design and development processes (Popham, 2007), but potentially can also be applied as ex-post evaluations of instructional sensitivity (e.g., Rovinelli & Hambleton, 1977). Judgmental approaches rely on trained experts in a domain rating the specified attributes of a test's instructional sensitivity. Ideally, in such methods, only instructionally sensitive items are selected for a final test instrument. However, the major drawback of these approaches is that it has not yet been demonstrated that experts can validly and reliably distinguish between tasks that are instructionally sensitive and those that are not (Chen, 2012; Polikoff, 2010; Way, 2014).

The second approach (2) is an empirical investigation of instructional sensitivity in learners' test outcomes, and includes a variety of empirical methods and respective designs.¹ One empirical method, an IRT-based Differential Item Functioning approach (DIF), has won recognition over recent years. In several studies (e.g., Naumann et al., 2014; Polikoff, 2010; Popham & Ryan, 2012) it has proved to be well suited to the purpose of detecting instructional sensitivity. This method goes back to the conceptual framework of Masters (1988). The major finding of Masters' framework is that, aside from the fact that high and low achieving students will usually score differently on a test, differential instructional sensitivity is reflected in some items being more highly discriminating than others. So the key technical element of DIF-based studies (e.g., Polikoff, 2010; Popham & Ryan, 2012) is that they compare the performance of groups on an assessment, controlling for the overall ability of the groups. In this respect, in either longitudinal or cross-sectional designs, DIF analyses can be run to compare instructed students to novice students, indicating whether the items are sensitive to the instruction experienced by the students (Polikoff, 2010, p. 17).

In this study, we seek to combine a judgmental with an empirical approach; Ruiz-Primo et al. (2012) have offered an example of such a triangulating approach.

3. Assessment design for VET

Especially in vocational education, where ability has to be demonstrated in the workplace on a regular basis, the concept of competence is more significant than the concept of mere knowledge, as a target construct of vocational assessment. We define competence in line with Mulder, Weigel, and Collins (2006, p. 79), as the "capability to perform by using knowledge, skills, and

attitudes that are integrated in the professional repertoire of the individual". A paper-pencil assessment is utilized to infer the apprentices' cognitive structures by assessing how well the test takers can do on authentic workplace-related tasks that they are expected to master at the end of their VET. So, in this contribution we specifically consider knowledge and ability as major cognitive prerequisites for the capability to perform in vocational situations, rather than attitude-related aspects of vocational competence in terms of attitudes and beliefs.²

Following the item classification system of Ruiz-Primo et al. (2012), our developed measure may be considered proximal to instruction: It is designed to take a snapshot of the relevant knowledge and skills in the curriculum. However, the exact content (e.g., a situation in the workplace) can be different to that studied during instruction. Our assessment was explicitly designed to align with the intended VET curriculum for industrial apprentices, and to be used as a paper-pencil test for the final examination of apprentices at the end of their VET (summative assessment). The design process was inspired by recent assessment theory (Pellegrino, Chudowsky, & Glaser, 2001; Mislevy & Haertel, 2006; Wilson, 2005; 2008).

In order to assure the validity of the assessment of instructional sensitivity, we undertook several assessment phases. After (1) defining our theoretical construct (as above), we (2) undertook a curricular analysis, in order to closely align our assessment of vocational knowledge and ability to the intended industrial VET curriculum. A particular feature of this phase was that in a VET assessment, we have to pay attention to the curricula of two learning sites: The German VET system is structured so as to equip apprentices with practical and theoretical knowledge by a dual system, of company-based training programs provided by the private sector (where the apprentices work about three days per week and are paid a wage by their employer), together with a school-based component (about two days per week, provided by the public sector).

Consequently, not only did we analyze the official curriculum of vocational schools, but we also made a survey study in the industrial sector, investigating what content is commonly taught and considered necessary by the apprentices' training companies. The specific job analysis was guided by several questions: What content is processed in which departments? What materials are used? How does internal/external communication take place (infrastructure)? All results and data were incorporated into the development of a model of the typical business processes that occur within companies (Winther, 2010). The model, which followed the process perspective of the St. Galler Management Model (Rüegg-Stürm, 2004) includes three central processes in (industrial) companies: *value chain processes*, related to quantifiable goods and services and their marketing; *control processes*, including decision support for management; and *management processes* that comprise business management and organization concerns.

The phase of item construction (3) was implemented according to three guiding principles, the first of which was a) *authenticity* of the vocational assessment (e.g., Achtenhagen & Weber, 2003; Shavelson & Seminars, 1968). In order to secure maximum authenticity, we modeled a simulated company that produces ceramic products such as tableware, bath tubs or sinks (see Appendix). All assessment items developed were implemented within the simulated company framework, together with additional realistic material, and information with which respondents

¹ For a comparison of different approaches of an empirical investigation of instructional sensitivity see Li, Ruiz-Primo, and Wills (2012a).

² However, such aspects presumably also influence task solutions during the assessment and therefore are in all likelihood integrated in our measurement approach to some (unknown) extent.

were to solve the items (e.g., product lists or e-mails; see [Appendix](#)). With respect to the design of the single items, the assessment tasks were designed to measure economic knowledge and skills in the commercial sector by representing job-related skills in the industrial sector. For this purpose, the item format of all tasks was open-ended.³

In order to attend to b) the varying cognitive demands of vocational practice, the items were developed on three *cognitive levels*, according to the conceptual framework of [Greeno, Riley, and Gelman \(1984\)](#), which represents an action schema for performing vocational tasks. On the first level, conceptual competence implies an understanding of the principles in the domain. It corresponds to factual knowledge that can be translated into an action schema. At the second level, procedural competence takes the form of knowledge in action, such as dealing with facts, structures and knowledge nets. At the third level, interpretational competence refers to strategic decision making that reflects the cognitive process of grounded interpretation of the findings obtained, through conceptual and procedural knowledge. To assess these different types of cognitive process, we modeled six conceptual items, seven procedural and three interpretative items.

The third principle of item construction refers to c) the administration of tasks of varying *specificity*. In line with [Gelman and Greeno \(1989\)](#), we distinguish between domain-linked and domain-specific item content in the business domain. The former, decontextualized aspect is generally relevant to the business domain, while the latter is highly situational and reflects the specific aspects, guidelines, and action maxims of a particular occupation. More precisely, domain-linked aspects refer to basic knowledge and skills that are generic but are nonetheless relevant prerequisites for solving vocational problems ([Klotz, Winther, & Festner, 2015](#)). In business domains, concepts such as literacy and numeracy are examples of this type of general preknowledge ([OECD, 2003](#); [Winther & Achtenhagen, 2009](#)). Domain-linked knowledge and ability is needed, for example, to perform simple exchange rate calculations in the workplace. Such calculations do not require any specific vocational knowledge or ability, but can be dealt with simply by applying the general mathematical concept of the “rule of three”, with which learners were already familiar from their general school education. Domain-specific knowledge and ability, on the other hand, entails job- or enterprise-specific knowledge and skills ([Oates, 2004](#)). In a business domain, an example of this kind of knowledge and ability might be rules that are newly acquired during vocational educational training: for example, for preparing a balance sheet in accounting. Both aspects of vocational knowledge and ability—domain-linked and domain specific—are prerequisites for solving workplace-related tasks (for sample items see [Appendix](#)). For the study at hand, we modeled 10 domain-specific items and 6 domain-related items.

In the (4) test assembly phase (see e.g., [Mislevy & Haertel, 2006](#)), an important principle of assessment design for vocational education relates to the assembly of single tasks into one coherent business-process ([Klotz, 2015](#)). The test therefore starts as a simulated typical event in the company (e.g., an e-mail from a potential client) demanding certain responses from the test takers, which in turn lead to further events and tasks (see [Appendix](#)).

The final step of our assessment design process included (5) validating our test design. We asked 24 vocational experts (12 experts for each item) to rate all tasks in terms of authentic item design (relevance of content and realistic situational setting), as well as to rate the items as either domain-linked or domain-

specific. Items that received an average value below 3.5 on the five-point Likert scale, in respect of workplace relevance, were excluded from the instrument. Moreover, we used the expert judgments of the items as being either domain related or domain specific, as a basis for the empirical analysis. The experts mostly agreed, in relation to their categorization of each item; this is reflected in the high degree of inter-rater reliability (Intraclass-Correlation-Coefficient [ICC] = 0.940).

4. Theoretical assumptions about instructional sensitivity in VET

As [Ruiz-Primo et al. \(2012, 693\)](#) note, most research studies concerned with instructional sensitivity focus on evaluating assessment instruments already developed and used, but are silent on how to construct instructionally sensitive assessments. In our research, in contrast, we implemented theoretical design principles, *ex ante*, into the assessment, that could be manipulated systematically to model items of varying instructional sensitivity. With respect to Research Questions 1 and 2, we are interested not only in whether the assessment is instructionally sensitive, but also, if so, why. Often, item attributes causing difficulty in tasks, also reflect sources of instructional sensitivity. Detection of instructional sensitivity therefore requires strong familiarity with the vocational area and its theoretical difficulty, or gleaning the necessary attributes through interaction with vocational experts. Ideally, both circumstances would apply, to enable determining which vocational activities are complex, and for what reason, and how the capacity to achieve them might develop over time, with instruction.

In our assessment the item design characteristics potentially causing difficulty were the level of cognitive processing and the degree of specificity of the learning content. However, we believe that only the later attribute plays a predominant role in generating instructional sensitivity. In line with [Billett \(1994\)](#), we argue that most often, vocational novices do not lack cognitive ability. Rather, in most instances, apprentices lack the specific knowledge and experience within a vocational domain ([Glaser, 1990](#)) that would otherwise enable them to conceptualize and categorize workplace-related problems and to deploy their cognitive structures more effectively ([Billett, 1994, p. 4](#)). Similarly, [Dreyfus and Dreyfus \(1980\)](#) describe vocational learning as an expansion of novices' generic preknowledge, which develops with relevant knowledge about aspects, specific guidelines, and action schemes, such that it transforms into an increasingly organized form, as specific knowledge and ability. The newly acquired specific knowledge is then stored—in addition to general knowledge and ability (domain-linked)—to provide the learner with a broad knowledge base from which to act in similar vocational situations.

The existing theoretical and qualitative research offers support for the idea of vocational learning as acquirement of specific knowledge and ability (see research on the expert-novice paradigm in diverse vocational domains: e.g., [Dreyfus & Dreyfus, 1980](#); [Benner, 2004](#); [Worthy, 1996](#); [Ryan, Fook, & Hawkins, 1995](#); [Campbell, Brown, & DiBello, 1992](#); [Chmiel & Loui, 2004](#)). We therefore assume that instructional sensitivity in vocational domains is determined by the extent of content specificity of items in an assessment. More precisely, two hypotheses can be formulated in reference to the above-stated research questions:

1. Advanced vocational learners improve significantly, compared to novices, in respect of their performance in the assessment (Hypothesis 1).
2. Items that are domain specific are significantly more strongly instructionally sensitive than are items that relate to domain-related generic contents (Hypothesis 2).

³ The tasks were scored with 0 for “none” or a wrong solution, 1 for a partially correct answer and 2 for a fully correct answer.

Table 1
Sample description (n = 534).

Sample Characteristic	Group 1 (n ₁ = 136)	Group 2 (n ₂ = 398)	Statistical population ^a
Average years of initial VET	0.1 (= beginners)	2.3 (= advanced)	Ø 0.0 years (= beginners)
Age	19.2	21.3	19.1
Gender	Female: 56%	Female: 59%	Female: 58%
Educational career	Secondary school: 1% Intermediate school: 30% High school diploma: 69%	Secondary school: 1% Intermediate school: 33% High school diploma: 66%	Secondary school: 2% Intermediate school: 36% High school diploma: 62%
Migration background ^b	21%	19%	(not available)

^a Numbers for the Federal Republic of Germany, according to the Federal Institute for Vocational Education and Training (BIBB), of apprentices entering their initial vocational training.

^b Migrational background was assessed in the questionnaire by asking for the language spoken at the apprentice's parental home.

5. Data acquisition and method

A cross-sectional design was used for the acquisition of data. This design was sufficient for our purpose of detecting instructional sensitivity in an assessment, as we did not seek to estimate or explain individual differences within the cohort, but only to ascertain whether items were instructionally sensitive at the aggregate cohort level. Moreover, longitudinal data would have caused test repetition effect issues (e.g., Hoffman, Hofer, & Sliwinski, 2011; Salthouse & Tucker-Drob, 2008). The cross-sectional data were gathered in 2013 as a non-random sample from visits to vocational schools in locations spread widely across Germany (Munich, Hanover, Bielefeld, and Paderborn). For economic efficiency, schools with a large proportion of industrial apprentices were selected. Access was initiated by the German Chamber of Industry and Commerce (IHK). Within these schools all students enrolled in industrial apprentice programs were selected, and all agreed to participate. Table 1 presents the sample, subdivided into the two groups of vocational novices (n₁ = 136) and advanced vocational learners (n₂ = 398), and the basic characteristics of these groups. Even though the data were gathered as a non-random sample, the two groups were remarkably similar in regard to the distributional characteristics of all collected variables, and showed no differences with regard to gender (T = -0.748; p = 0.455), educational career paths (T = -0.169; p = 0.866) and migrational background (T = -1.011; p = 0.313). The two subsets (group 1 and group 2) only differed significantly with regard to the average time spent on vocational educational training (years spent in vocational training) and average age (T = -8.630; p = 0.000). Moreover, the distributions of the two collected subsamples are comparable to the general population of industrial apprentices in Germany (Table 1).

During test taking we observed, in regard to test motivation, that the students engaged very well with the instrument—most probably because it had been represented to them as a useful preparation for their final examination, and because we had assured them of individual feedback. This also likely explains the low rate of missing values (1.68%). The solutions to the items were corrected and coded according to a detailed scoring guide (Wilson, 2008). Two independent raters randomly corrected 16% of all 534 tests, in order to estimate the accuracy of the scoring process. The Intraclass-Correlation-Coefficient (ICC) proved a satisfactory degree of scoring objectivity (ICC = 0.914).

To analyze the open-ended items, we used a multidimensional-random-coefficient-multinomial logit-model (Adams, Wilson, & Wang, 1997) and analyzed the polytomous database of varying scaling with the program ConQuest (Wu, Adams, & Wilson, 1997). Then, thresholds for the two groups were estimated: for vocational novices (group 1) and advanced learners at the end of their training (group 2). A downward shift in the difficulty of items in a comparison of group 1 with group 2 would mean that the learners must

have progressed in their vocational knowledge and ability, as the items were relatively easy to solve for them, in comparison with vocational novices. In order to determine the difficulty of all items in both groups, we used a Differential Item Functioning (DIF) approach. DIF analyses explore whether the probabilities for the solving of items are different for different groups, after controlling for overall group performance (Holland & Wainer, 1993; Wilson, 2005, p. 165). For this purpose, the simple Rasch-Model was extended by a group term, in which an interaction term interacted with the single assessment items and therefore functioned as an empirical criterion for the existence of differential differences between the groups.

6. Results

The item statistics suggest infit for all items included in the model ($0.81 \leq \text{wMNSQ} \leq 1.12$)⁴ and satisfactory reliability values (EAP/PV reliability = 0.846).⁵ Applying the DIF approach to our database, we obtained the results given in Tables 2 and 3. As can be seen in Table 2, there was a significant difference in performance on our assessment from the beginning until the end of VET instruction. Given the large chi-square and p-value < 0.001, we reject H₀ (that there is no difference between novices and advanced learners). The estimated value of vocational knowledge and ability of group 2 for the assessment was, on average, 1.446 logits higher than for group 1; this, with a large effect size,⁶ is indicated by tasks being harder for beginners than for advanced learners. This means that learners acquired a significant additional degree of vocational knowledge and ability during training, and that we were able to capture this with our developed assessment instrument (Hypothesis 1).

Apart from this general change on the scale of vocational knowledge and ability, it was also possible to look at each item's difficulty for the total sample and compare it to the difficulty in each group. If the change in difficulty from group 1 to group 2 was larger than what could be expected from the general advancement given in Table 2 (1.446), the item must have been exposed to DIF. Table 3 shows the item difficulty for the total sample, and changes in the subsamples.

With respect to Table 3 it is important to note that the DIF approach consists of a strictly relative analysis. That is, every item

⁴ Adams and Khoo (1996) advocate infit for items with a weighted Mean Square (wMNSQ) value from 0.75 to 1.33.

⁵ The Expected A Posteriori/Plausible Value (EAP/PV) reliability indicates how much variance in a person's estimated ability is accounted for by the measurement model on average for all testees. As a scale reliability it can be compared to Cronbach's alpha and should be 0.80 or preferably higher for research designs based on correlative relations (Nunnally, 1978).

⁶ According to Paek (2002), absolute differences on the logit scale less than 0.426 are negligible. Differences up to 0.638 indicate medium-sized effects, and those higher than this level indicate a strong effect size for a learning progression.

Table 2
General advancement in performance on the assessment for vocational knowledge and ability.

Subgroups	Z	Ability Estimate	Standard Error	Chi-Square (df)	p-Wert
Group 1 (novices)	1	0.723	0.017	1797.35 (1)	0.000
Group 2 (advanced learners)	2	-0.723	0.017		

Table 3
Item difficulty for total sample and subsamples.

Item	Absolute item difficulty for the total sample	DIF (Item*Instruction) for group 2 compared to group 1	Error	Chi-Square (df)	p-value
1 (dl)	-0.026	0.528	0.038	698.48 (15)	0.000
2 (dl)	0.220	0.066	0.039		
3 (ds)	-0.793	-0.986	0.046		
4 (dl)	1.126	0.786	0.035		
5 (ds)	0.483	-0.202	0.044		
6 (ds)	1.846	0.218	0.045		
7 (dl)	-0.603	0.058	0.038		
8 (dl)	-0.587	0.480	0.037		
9 (ds)	-0.910	-1.160	0.046		
10 (ds)	-0.926	-1.068	0.046		
11 (dl)	1.201	0.160	0.044		
12 (ds)	0.107	-0.370	0.045		
13 (dl)	-0.139	0.378	0.039		
14 (dl)	-0.210	0.360	0.039		
15 (ds)	-0.343	-0.102	0.037		
16 (dl)	-0.446	0.852	0.161		

on the assessment shows a positive gain from beginner to advanced, in absolute terms. However, looking at the DIF from group 1 to group 2, it becomes obvious that all items that were disproportionately easier for subsample 2 compared to subsample 1 (indicated by a negative sign) were domain-specific tasks. These assessment items were highly sensitive to instruction.

Domain-linked items (dl) on the other hand, underestimated the total improvement of learners during VET. This does not mean that learners did not improve in respect of their general abilities during VET (the group effect that adds to the DIF analysis was 1.446 logits, and thus was always larger than the disadvantage for advanced learners of an item being domain-linked), but that their improvement with respect to those items was less than what we would expect, given the total learning progress on the scale of vocational knowledge and ability. In our assessment, 9 items demonstrated negligible DIF, two demonstrated medium DIF (two domain-linked items) and five items demonstrated large DIF (three domain-specific items and two domain-linked items).

In order to further quantify this difference, it is possible to calculate the average DIF for domain-linked and domain-specific items. Domain-linked items demonstrated an average DIF of 0.389 as a disadvantage for advanced learners. Domain-specific items contrarily demonstrated an average DIF of 0.648 as an advantage for advanced learners, indicating that items that were domain specific were significantly more instructionally sensitive than items that related to domain-related contents (Hypothesis 2). Taking on an absolute view, learners progress by 1.057 when administered domain-linked items and by 2.094 when administered domain-specific items. So the increase in specific knowledge and ability is on average roughly double the increase in domain-linked ability.

Fig. 1 summarizes all results graphically. The IRT-Wright-Map on the left hand side orders the items of the assessment for the total sample from least to most difficult. Items 6 or 11 are most demanding with respect to the required quality of vocational knowledge and ability. Items 12 or 1 were of about average difficulty, and items 9 and 10 were the easiest items of the assessment.

The instructional function given in the middle of the graph illustrates that for advanced vocational learners (group 2), the average difficulty of the whole assessment dropped from 0.723 to -0.723. This in turn means that vocational knowledge and ability must have improved by 1.446 logits, given the higher probability of solving items. The last column shows the interaction of the single items with the duration of instruction. E.g. item 9 had a DIF effect of -1.160 logits for group 2 (9.2) compared to group 1 (9.1).

If we add the DIF effect of (for example) item 9 to the total group effect (1.446), we can calculate the absolute difference in difficulty for this item in both groups. This value (2.606), which can be obtained in Fig. 2, indicates the absolute distance of an item on the logit scale for both groups (from 1.9 to 2.9). Here, the first number of an item indicates the group to which it belongs, the second, the item name; the third number refers to the different thresholds a polytomous item can possibly have. Looking at the graph, it again becomes obvious that all items were easier for advanced learners (shaded items) than they were for vocational novices.

Although it has been shown that domain-specific items were more instructionally sensitive and therefore, advantaged advanced learners in an assessment, in our data there was one exception to this rule. The one item not fitting into this scheme was item 6 (see Appendix), which disadvantaged advanced learners. A possible explanation for this phenomenon is that as learners gain in knowledge, this leads to more intra-individual cognitive conflicts (see Foster, 2011; Naumann et al., 2014; Vosniadou, 2007). As learners gain additional knowledge, newly acquired knowledge structures might conflict with existing knowledge, generating greater uncertainty in the answering process. This explanation becomes probable when one looks at learners' answers. Item 6 asked students whether a binding purchase contract was in place, in respect of the business process described in the realistic scenario. Learners new to the domain mostly followed their intuition, arguing that no binding purchase contract was in place, as the purchase offer had already expired by the acceptance date (correct). Advanced learners, on the other hand, who could give a legal definition of a purchase contract, were often less sure if a purchase

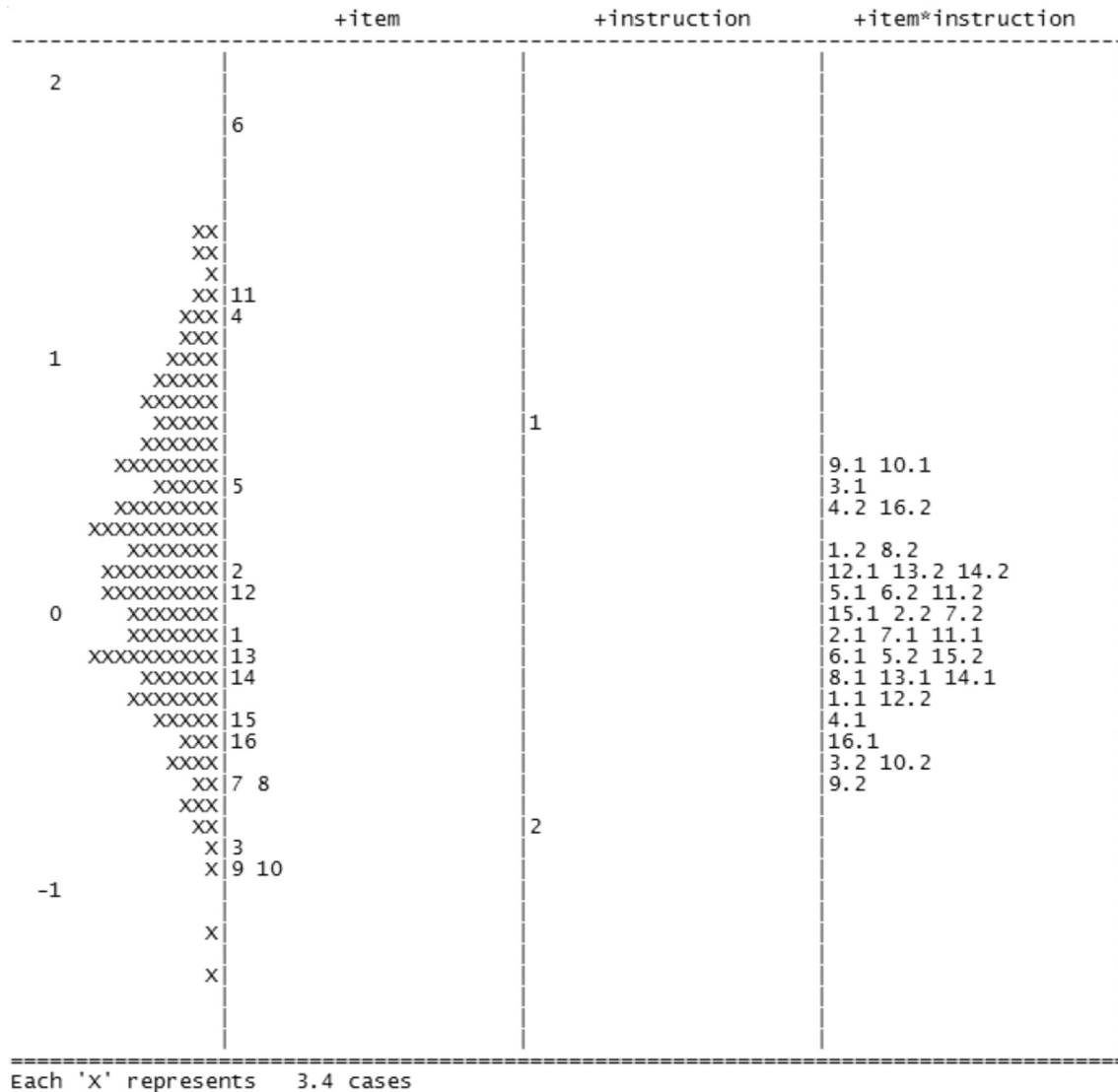


Fig. 1. DIF-analysis for the assessment.

contract was in place, as their theoretical definition did not quite fit the situational setting of the business scenario.

7. Conclusion and discussion

The definition and detection of instructional sensitivity is not an isolated endeavor but rather is a matter of what is supposed to be taught in the classroom (curriculum), what is actually taught in the classroom (instruction), and how well tests and items align with what is taught (assessment). Instructional sensitivity should therefore be evaluated according to the notion of the curriculum-instruction-assessment triad (Pellegrino, 2012). With respect to Research Question 1, the results suggest that during vocational instruction, apprentices significantly improve their performance (a large effect) and that it is possible to track these changes in the quality of vocational knowledge and ability over the span of initial VET via an instructionally sensitive assessment for vocational knowledge and ability that aligns with the vocational curriculum.

The results strengthen the proposition that dual vocational learning is a powerful system for skill acquisition (Bonnal, Mendes, & Sofer, 2002; Griffin, 2016; OECD, 2008; OECD, 2010) positioned at the boundary between learning and working (Harteis, Rausch and

Seifried, 2014). The high exposure of almost all items to instructional sensitivity over a relatively short period of time (about two years) points to dual VET being an effective education system for conveying workplace-related knowledge and ability to adults successfully.

With respect to Research Question 2, we were interested not only in whether the assessment was instructionally sensitive, but why it proved to be so. As we have demonstrated, theoretically and empirically, the question of instructional sensitivity is also a theoretical question in respect of the target construct assessed. The more generic the assessed knowledge and ability, the less sensitive this construct is to instruction, whereas the more specific the assessed knowledge and ability, the more sensitive this construct is to instruction. Hence, the results empirically support Dreyfus and Dreyfus's (1980) profound conception of vocational learning as an expansion of novices' generic abilities through specific knowledge and ability, together allowing for the solving of situational problems. In the past this theory has been supported by qualitative research in diverse vocational domains (e.g., Campbell, Benner 2004; Campbell et al., 1992; Chmiel & Loui, 2004). The results also point to the possibility that for adults, the acquirement of specific knowledge and ability is less laborious and amenable than is the acquirement of general abilities.

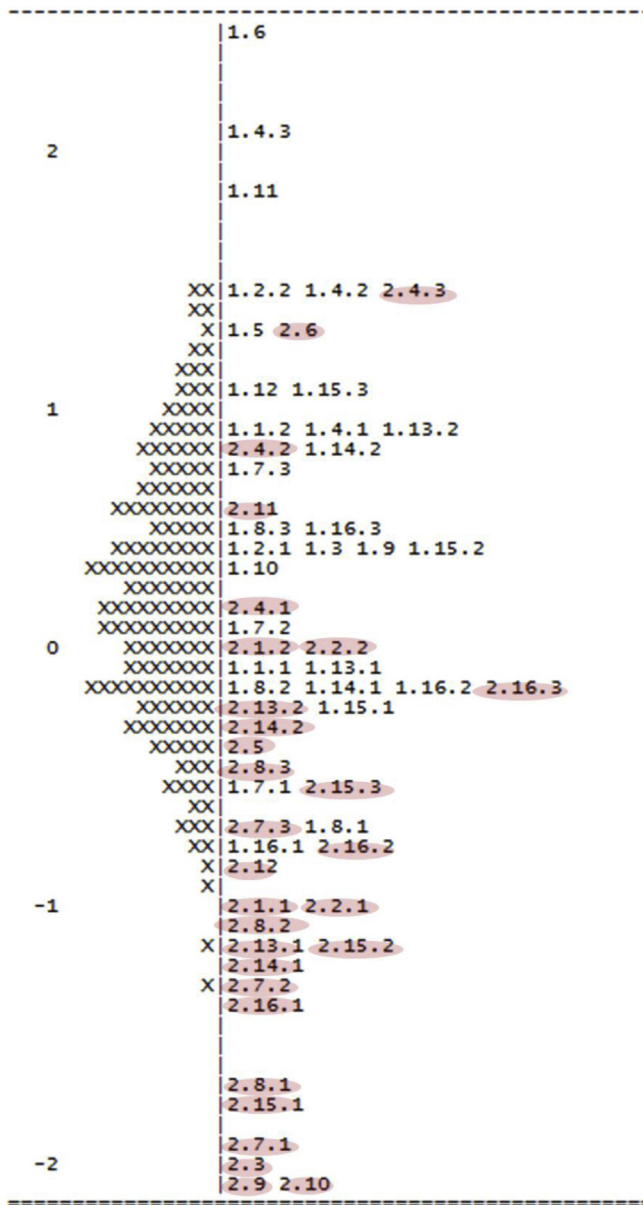


Fig. 2. Absolute item difficulty in each sample (group 1 and group 2).

However, more surprising is the finding that during VET, adults also significantly increase their general abilities, such as numeracy and literacy—although to a lesser extent. These abilities should have been learnt already at school, but were only successfully acquired during VET. This challenges the notion that vocational learning consists solely of the transition from generic to specific knowledge. Rather, it appears that vocational learning settings also incidentally stimulate the acquisition of general abilities, presumably through experience and/or the didactical approaches of situated or problem-based learning (as suggested e.g. by Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991; Gruber, Harteis, & Rehr, 2008). This suggestion is confirmed explicitly by research in the domain of management learning (Kolb & Kolb, 2009; see also the empirical research of; Klotz, 2015).

On a practical level, the finding of different categories of vocational items with varying degrees of instructional sensitivity, allows for future assessment development to model item characteristics that can be systematically manipulated to develop items and

assessments that prove to be instructionally sensitive. Via an ex ante classification of item specificity, by experts, items that are sensitive to instruction and therefore especially valuable with respect to the information they can impart on learning success, could be identified.

However, this study is subject to several limitations that need to be considered and that may inspire future research endeavors: First, the results reported here are limited by the fact that a convenience sample of participants was used, and consequently the obtained results cannot be considered strongly generalizable. However, the two groups were remarkably similar in regard to the distributional characteristics of all collected variables, and to the general population of industrial apprentices in Germany. Second, the cross-sectional nature of the data did not allow for controlling the baseline achievement of the two subsamples. Third, as noted above, according to Polikoff (2010, 9) a finding of high sensitivity indicates both good instruction and a high quality test that is sensitive to that instruction. However, this is true only for the effectiveness of training, and it is possible that the test questions, or even the learning goals set in the curriculum, were relatively unchallenging. That is, while, on the basis of our results, the training appears to have achieved the desired outcomes adequately, actually more could have been achieved (see e.g., Popham & Ryan, 2012). Therefore, on the basis of this study, we are not able to make a statement about the efficiency of the VET.

Another limitation can be seen in the way we assessed the authenticity of the assessment. For the purposes of determining and improving the authenticity of the assessment, we only gathered expert data. Authenticity however is in the eye of the beholder (Gulikers, Bastiaens, Kirschner, & Kester, 2008), so future research in the vocational domain should also gather data on apprentices' perceptions of the authenticity of an assessment, as the perspectives of experts and testees may yield different outcomes (Khaled, Gulikers, Biemans, & Mulder, 2015).

Finally, while the data showed significant progress in domain-linked and domain-specific knowledge and ability, it did not show why. So, while we know that the assessment was instructionally sensitive, we do not know which aspect or aspects of instruction yielded the educational outcomes. For instance, we are yet unable to say which part of the dual education—the learning at a vocational school or the working and learning at a training company—contributed most to this finding, given that “instruction”, for our sample, refers to the dual VET treatment as a whole. The same applies for the respective didactical methods used by the teachers in vocational schools and at the workplace.

Therefore, the exact causalities for the learning processes observed here, on the boundary between learning and working, remain hidden, and future research might adopt a broader understanding of the topic of instructional sensitivity, including measures of the pedagogic quality of instruction, in order to probe the issue of instructional sensitivity more deeply and, with respect to vocational education, to understand more fully the qualities and potentials of vocational training as an environment in which not merely domain-specific, but also broader educational goals, can be addressed.

Conflict of interest

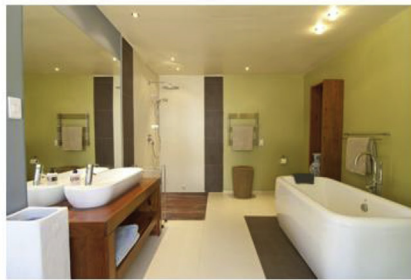
The authors declare that they have no conflict of interest.

Acknowledgments

This research was supported by the German Research Foundation, within the projects “Competence development through enculturation” (KL 3076/2-1) and “Competence-oriented assessments in VET and professional development” (Wi 3597/1-2).

Appendix

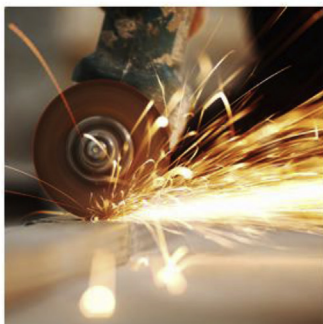
Ceraforma Keramik AG



Since its foundation in 1982, the 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 – used to operate too independently from 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 Friebel 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.

- 5 After repeated negotiations the company Ceraforma accepts the order from the DIY Bauhannes at the price stipulated by Mr Schwiener. 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 contacted various suppliers by mail and you received 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 Bestellungseingang
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

References

- Achtenhagen, F., & Weber, S. (2003). "Authentizität" in der Gestaltung beruflicher Lernumgebungen. 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.
- Adams, R. J., & Khoo, S. T. (1996). *Quest-Interactive test analysis system*. Victoria: Australian Council for Educational Research.
- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement*, 21, 1–23.
- AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Benner, P. (2004). Using the Dreyfus Model of Skill Acquisition to describe and interpret skill acquisition and clinical judgement in nursing practice and education. *Bulletin of Science, Technology and Society*, 24(1), 188–199.
- Billett, S. (1994, December). Situated cognition: Reconciling culture and cognition. In *Paper presented at reforming post compulsory education and training: Reconciliation and reconstruction* (Brisbane, Australia).
- Bonnal, L., Mendes, S., & Sofer, C. (2002). School-to-work transition: Apprenticeship versus vocational schools in France. *International Journal of Manpower*, 23(5), 426–442.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–34.
- Burstein, L. (1989, March). Conceptual considerations in instructionally sensitive assessment. In *Paper presented at the annual meeting of the American Educational Research Association*. San Francisco, CA.
- Burstein, L., Aschbacher, P., Chen, Z., & Lin, L. (1990). *Establishing the content validity of tests designed to serve multiple purposes: Bridging secondary-postsecondary mathematics*. CSE Technical Report 313. Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA: University of California, Los Angeles.
- Campbell, R. L., Brown, N. R., & DiBello, L. A. (1992). The programmer's burden: Developing expertise in programming. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 269–294). New York: Springer.
- Chen, J. (2012). *Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods*. Lawrence, KS: University of Kansas.
- Chmiel, R., & Loui, M. C. (2004). Debugging: From novice to expert. In *Proceedings of the 35th SIGCSE technical symposium on computer science education*, 17–21 (New York, USA).
- Court, S. (2013, November). DIF and SPGS: Implementing the popham-ryan design. In *Presentation at the first international conference on instructional sensitivity* (Lawrence, Kansas).
- Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Berkley: University of California.
- Foster, C. (2011). A slippery slope: Resolving cognitive conflict in mechanics. *Teaching Mathematics and Its Applications*, 30(4), 216–221.
- 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, NJ: Erlbaum Associates.
- Glaser, R. (1990). Re-emergence of learning theory within instructional research. *American Psychologist*, 45(1), 29–39.
- Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.
- Greeno, J. G., Riley, M. S., & Gelman, R. (1984). Conceptual competence and children's counting. *Cognitive Psychology*, 16(1), 94–143.
- Griffin, T. (2016). *Costs and benefits of education and training for the economy, business and individuals*. Adelaide: NCVET.
- Gruber, H., Harteis, C., & Rehr, M. (2008). Professional learning: Skill formation between formal and situated learning. In K. U. Mayer, & H. Solga (Eds.), *Skill formation. Interdisciplinary and cross-national perspectives* (S. 207–229). Cambridge: Cambridge University Press.
- Gulikers, J. T. M., Bastiaens, T. J., Kirschner, P. A., & Kester, L. (2008). Authenticity is in the Eye of the beholder: Student and teacher perceptions of assessment authenticity. *Journal of Vocational Education and Training*, 60(4), 401–412.
- Haladyna, T., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement*, 18(1), 39–53.
- Harteis, C., Rausch, A., & Seifried, J. (Eds.). (2014). *Discourses on professional learning: On the boundary between learning and working*. Dordrecht: Springer.
- Hasselhorn, M., Baethge, M., Füssel, H.-P., Hetmeier, H.-W., Maaz, K., Rauschenbach, T., et al. (2014). *National Educational Report, Education in Germany 2014—an indicator-based report including an analysis of the situation of people with special educational needs and disabilities*. Federal Ministry of Education and Research: Bertelsmann.
- Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychology and Aging*, 26(4), 778–791.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Kao, C.-F. (1990). *An investigation of instructional sensitivity in mathematics achievement test items for U.S. eighth grade students*. Los Angeles: University of California.
- Khaled, A., Gulikers, J., Biemans, H., & Mulder, M. (2015). How authenticity and self-directedness and student perceptions thereof predict competence development in hands-on simulations. *British Educational Research Journal*, 41(2), 265–286.
- Klotz, V. K. (2015). *Diagnostik beruflicher Kompetenzentwicklung: Eine wirtschaftsdidaktische Modellierung für die kaufmännische Domäne*. Berlin: Springer.
- Klotz, V. K., Winther, E., & Festner, D. (2015). Modeling the development of vocational competence: A psychometric model for business domains. *Vocations and Learning*, 8(3), 247–268.
- Kolb, A. Y., & Kolb, D. A. (2009). Experiential learning theory: A dynamic, holistic approach to management education and development. In S. J. Armstrong, & C. V. Fukami (Eds.), *The SAGE handbook of management learning, education and development* (pp. 42–68). London: Sage.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
- Linn, R. L. (1983). Curricular validity: Convincing the courts that it was taught without precluding the possibility of measuring it. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 115–132). Boston, MA: Kluwer-Nijhoff Publishing.
- Li, M., Ruiz-Primo, M. A., Giamellaro, M., Wills, K., Mason, H., & Lan, M.-C. (2012b). Instructional sensitivity and transfer of learning at different distances: Close, proximal and distal assessment items. In *Paper presented at the AERA annual meeting* (Vancouver, Canada).
- Li, M., Ruiz-Primo, M. A., & Wills, K. (2012a). *Comparing methods to estimate the instructional sensitivity of items*. DEISA. Paper 4.
- Masters, J. R. (1988). A study of the differences between what is taught and what is tested in Pennsylvania. In *Paper presented at the annual meeting of the national council on measurement in education*. New Orleans, LA.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(1), 6–20.
- Mulder, M., Weigel, T., & Collins, K. (2006). The concept of competence in the development of vocational education and training in selected EU member states: A critical analysis. *Journal of Vocational Education and Training*, 59(1), 65–85.
- Naumann, A., Hochweber, J., & Hartig, J. (2014). Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach. *Journal of Educational Measurement*, 51(4), 381–399.
- Nickolaus, R., Lazar, A., & Norwig, K. (2012). Assessing professional competences and their development in vocational education in Germany: State of research and perspectives. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: Learning outcomes in science education* (pp. 141–161). Münster: Waxmann.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- 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 in Education*, 2(1), 53–71.
- OECD. (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.
- OECD. (2008). *Costs and benefits in vocational education and training*. Paris: OECD.
- OECD. (2010). *Learning for Jobs: Synthesis report of the OECD reviews of vocational education and training*. Paris: OECD.
- Paek, I. (2002). *Investigation of differential item functioning: Comparisons among approaches, and extension to a multidimensional context*. Berkley, CA: University of California.
- Pellegrino, J. W. (2012). The design of an assessment system focused on student achievement: A learning sciences perspective on issues of competence, growth and measurement. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible—learning outcomes in science education* (pp. 79–107). Münster: Waxmann.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know—the science and design of educational assessment*. Washington, DC: National Academy Press.
- Pham, V. H. (2009). *Computer modeling of the instructionally insensitive nature of the Texas assessment of knowledge and skills (TAKS) Exam*. Austin: The University of Texas at Austin.
- Phillips, S. E., & Mehrens, W. A. (1988). Effects of curricular differences on achievement test data at item and objective levels. *Applied Measurement in Education*, 1(1), 33–51.
- Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, 29(4), 3–14.
- Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 89(2), 146–155.
- Popham, W. J., & Ryan, J. M. (2012, April). Determining a high-stakes test's instructional sensitivity. In *Paper presented at the annual meeting of the national council on educational measurement*. Vancouver, B.C., Canada.
- Rausch, A., Seifried, J., Wuttke, E., Kögler, K., & Brandt, S. (2016). Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. *Empirical Research in Vocational Education and Training*, 8(9). <http://dx.doi.org/10.1186/s40461-016-0053-y>.
- Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research*, 2(1), 49–60.
- Rüegg-Stürm, J. (2004). Das neue St. Galler management-modell. In R. Dubs, D. Euler, J. Rüegg-Stürm, & C. Wyss (Eds.), *Einführung in die Managementlehre* (pp. 65–134). Bern: Haupt.
- Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M.-C., Mason, H., et al. (2012). Developing and evaluating instructionally sensitive assessments in science.

- Journal of Research in Science Teaching*, 49(6), 691–712.
- Ryan, M., Fook, J., & Hawkins, L. (1995). From beginner to graduate social worker: Preliminary findings of an Australian longitudinal study. *British Journal of Social Work*, 25(1), 17–35.
- Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, 22(1), 800–811.
- Seeber, S. (2008). Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 104(1), 74–97.
- Seeber, S. (2015, April). The impact of institutional training settings on competence development in the area of social and health care. In *Presentation at the AERA annual meeting* (Chicago).
- Seeber, S., Ketschau, T., & Rüter, T. (2016). Structure and level of vocational competence of Medical assistance. *Unterrichtswissenschaft*, 44(2), 185–203.
- Shavelson, R. J., & Seminars, J. (1968). Effect of lunar gravity on man's performance of basic maintenance tasks. *Journal of Applied Psychology*, 52(3), 177–183.
- Switzer, D. M. (1993). Differential item functioning and opportunity to learn: Adjusting the Mantel-Hansel chi-square procedure. *Practical Assessment, Research & Evaluation*, 13(7), 1–16.
- Vosniadou, S. (2007). The cognitive-situative divide and the problem of conceptual change. *Educational Psychologist*, 42(1), 55–66.
- Way, W. (2014). *Memorandum on instructional sensitivity considerations for the PARCC assessments*. Washington DC: Pearson.
- Weber, S., Wiethe-Körprich, M., Bley, S., Weiß, C., Draxler, C., & Güter, C. (2016). Modellierung und Validierung eines Intrapreneurship-Kompetenz-Modells bei Industriekaufleuten. *Unterrichtswissenschaft*, 44(2), 149–168.
- Wiliam, D. (2007). Sensitivity to instruction: The missing ingredient in large-scale assessment systems?. In *Paper presented at the annual meeting of the international association for educational assessment* (Baku, Azerbaijan).
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wilson, M. (2008). Cognitive diagnosis using item response models. *Journal of Psychology*, 216(2), 74–88.
- Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: W. Bertelsmann Verlag.
- 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(1), 85–102.
- Worthy, C. (1996). Clinical ladders: Can we afford them? *Nursing Management*, 27(1), 33–34.
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Multi-aspect test software*. Camberwell: Australian Council for Educational Research.
- Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn and equity: New standards examinations for California Mathematics Renaissance*. CSE Technical Report 484. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California.
- Yu, L., Lei, P.-W., & Suen, H. K. (2006). Using a differential item functioning (DIF) procedure to detect differences in opportunity to learn (OTL). In *Paper presented at the american educational research* (San Francisco, California).