Wiethe-Körprich and Bley *Empirical Res Voc Ed Train (2017) 9:8*
DOI 10.1186/s40461-017-0052-5

Empirical Research in Vocational Education and Training

RESEARCH Open Access

Prospective educators as consumers of empirical research: an authentic assessment approach to make their competencies visible

Michaela Wiethe-Körprich* and Sandra Bley

*Correspondence:
wiethe@bwl.lmu.de
Institute for Human Resource Education and Management, Munich School of Management, Ludwig-Maximilians-University in Munich, Ludwigstraße 28/RG, 80539 Munich, Germany

Abstract

Background: Educators today have to be able to make current empirical research results usable for everyday practice. Consequently, there are increasing endeavors to develop and assess competencies in consuming empirical research (CCER) on an academic level. However, problems with regard to recruiting and motivating test participants—rooted in the prevalence of low-stakes testing conditions—could limit confidence in the validity of the findings. The current study presents a structure and proficiency level modeling for CCER under high-stakes conditions.

Method: The sample comprises N = 155 bachelor students of Human Resource Education and Management. The assessment design of the 26 items complied with demanding standards for designing tests (such as Evidence-Centered Design and authenticity).

Results: The results are as follows: (1) We were able to confirm our expected structural model which consists of two dimensions ('conceptual competencies' and 'statistical competencies') instead of one overarching dimension. (2) The test items are of a high quality. (3) Three levels of CCER could be defined according to two task characteristics (cognitive processes and complexity) which explain nearly 100% of the prospective educators' CCER abilities.

Conclusion: The results of the study show that we succeeded in designing a reliable and valid test instrument for assessing (prospective) VET-educators' competencies in consuming empirical research.

Keywords: High-stakes testing, Item-response-theory, Research competencies

Background

Educators in vocational education and training (VET) should act as consumers of empirical research

Undertaking empirical research stimulates profitable innovation (Egeln et al. 2002). Natural scientists, for instance, have long taken it for granted that they should base their practical actions on current scientific research. In the healthcare sector, for example, hardly anybody wants to be treated by a doctor who refers to outdated research findings (Jahed et al. 2012). In many professions it is now common that practitioners are obliged to know about the latest relevant research results. Educators too have to be familiar with principles of empirical research in so far as they are able to reflect and to

critically question the findings of scientific research. Correspondingly, it has been suggested that evidence-based practice should prompt educational professionals to be aware of recent advances in their area of work (Darling-Hammond and Bransford 2005, pp. 15–16; Weber and Achtenhagen 2009). This enables them to monitor whether their educational activities are successful. But the sector of science also benefits if the latest research results are applied in economic practice. On the other hand, science can take up the research interests postulated by practice (Wuttke 2001, p. 40; Zurstrassen 2009, p. 41). Slavin provides a pithy summary of the situation: Educators need to be sophisticated consumers of research, regardless of whether they are also producers of research (2007, p. 2).

Referring to Slavin (2007) and further authors (e.g. Stark and Mandl 2001; Schweizer et al. 2011) research methodological tasks comprise two key challenges: (a) *reviewing* empirical academic literature, and (b) independently *performing* empirical research projects. Within this study, we consider educators as consumers—not as producers—of research, who must be able to avail themselves of scientific results—in terms of scientific studies—in their everyday practice. Therefore, we suggest, they need to have competencies in consuming empirical research (CCER). Slavin's (2007) as well as Darling-Hammond's and Bransford's (2005) requirements originally referred only to teachers. We claim that the active use of research findings is relevant for all people responsible for education, because teaching and learning take place in various settings, not only in schools. We chose VET-educators [Human Resource Education and Management students of the Ludwig-Maximilians-University in Munich (LMU)] as target group for our study, because they cover polyvalent professional areas. They are typically employed in various workplaces, for example (1) as teachers or trainers in schools and companies or in organizations for further education, (2) as organizers of vocational training, human resource management, and professional development within enterprises, (3) as administrative educators or politicians within chambers, associations, or ministries, or (4) as consultants or coaches in different educational environments. A teacher employed in a vocational school, for instance, needs CCER for aligning his/her instructional methods to the latest research results on efficient teaching. Another example constitutes employees operating within a company's apprenticeship department. They need CCER in order to design workplace learning processes according to current scientific findings within this field.

Research-based training of VET-educators

The described evidence-based orientation has manifested itself within the many international and national professionalization standards for educators. Scientific studies on the effectiveness of teacher education and the corresponding professionalization efforts exist in the field of general (e.g. Baumert and Kunter 2006; Blömeke et al. 2008; Blömeke et al. 2011) and vocational (Bouley et al. 2015) education. Within the current scientific literature, the professionalization standards are often the starting point for measuring competencies. They reflect—among other competencies—the importance of educators' research methodological competencies: (1) All ten international 'Core Teaching Standards' modeled by the InTASC (Interstate Teacher Assessment and Support Consortium 2011) imply skills linked to the field of practices in consuming research; (2) concerning the twelve Swiss standards—formulated by Oser (1997)—nearly all these standards include practices in consuming

research implicitly; (3) within the national German KMK standards ('Kultusministerkonferenz'; Ministers' of Education and the Arts conference) (KMK 2004) competence number ten is assigned to the area of innovating and postulates that teachers should understand their profession as a lifelong learning task. In the light of the continuously decreasing half-life of knowledge, this is an essential claim that is only feasible if teachers master practices in consuming research. Accordingly, competence number ten stresses the aims and methods of educational research as well as the interpretation and application of its results as one central curricular focus. The corresponding standards emphasize inter alia that graduates of teachers' study programs must be able to receive and evaluate results from educational research and to use these results to optimize their educational activities (KMK 2004, pp. 5, 12). From the present authors' perspective, these requirements are transferable and necessary for all people responsible for education. Furthermore, it has to be guaranteed that the evidence-based orientation is also embedded within the corresponding instructional processes (e.g. Slavin 2008, pp. 5–14; Fichten 2010, p. 159).

Assessing prospective VET-educators' competencies in consuming research

The competencies defined within the curriculum and implemented within the instruction program have to be translated into an operationalized form in order to make them measurable. Older studies on research-methodological competencies by Stark and Mandl (2001), Schweizer et al. (2011) as well as Wagner and Maree (2007) focused more on the development, implementation, and evaluation of training programs for promoting these competencies. Comparable with our intention of modeling and measuring competencies in consuming empirical research, only the recently published AHELO project by the OECD (Tremblay et al. 2012) and the LeScEd (Learning the Science of Education) project (Groß Ophoff et al. 2014, 2015) exist. They aim at assessing competencies that are relevant in the field of working scientifically in higher education. Both initiatives address the application of research concepts and the adequate use of statistical tools. The LeScEd project deals with modeling and measuring the educational research literacy of students within the field of educational sciences and therefore adopts a similar approach to our work. But—in contrast to our study—within both existing initiatives (LeScEd and AHELO), testing is performed under low-stakes conditions.

As Wise and DeMars (2005, 2006) show, in some cases low-stakes testing conditions can lead to fundamentally biased test results. These effects could be intensified in the field of higher education, because of the lack of compulsory attendance (Wise and DeMars 2006; Wolf et al. 2015). Students who organize their studies independently—what is explicitly desired—could tend to neglect low-stakes tests. Due to competitive obligations during their studies, they will rather prioritize high-stakes tests that bear serious consequences for their academic progress. Consequently, low-stakes testing in higher education could lead to a higher probability of self-selection effects, as well as to a lower motivation for participating in the respective test. This can threaten the representativeness of the sample and raise the number of not-answered tasks (missing values). In addition, for some types of learners the application of statistical tools represents a serious obstacle, due to their anxiety over such formal methods. Such individuals are often unable to cope with corresponding tasks (Onwuegbuzie 2001). This phenomenon can reinforce the number of missing values and is likely to be intensified if low-stakes testing conditions prevail. This can entail

large losses of data points and bias the calculated estimators as well as the identified competence structure. Further, if we consider the missing values as representing participants who lack statistical competencies, this could result in a biased underestimation of these competencies. Therefore, we focus on modeling and measuring prospective VET-educators' competencies in consuming empirical research under high-stakes testing conditions.

We aimed at developing an appropriate performance measurement instrument including authentic test tasks which meets the ambitious standards for designing tests. The following approaches led our assessment development: Collegiate Learning Assessment (Shavelson 2008), Evidence-Centered Assessment Design (e.g. Mislevy and Haertel 2006), and the authentic assessment (Janesick 2006). In line with the described overarching goal, the primary objectives of our study are: (1) To develop a structural model for CCER and to prove this model empirically by using the Item-Response-Theory (IRT) (Hartig and Frey 2013). (2) To investigate if the 26 test items meet the central Rasch-modeling assumption of equal item discriminability and whether they allow for reliable and valid measurement. (3) To define a proficiency level model for the central CCER in order to make a statement about the prospective VET-educators' degree of competence.

Theoretical background and research questions

The underlying concept of competence

In accordance with the discussion of modeling and measuring professional competencies (Blömeke et al. 2015; Shavelson 2010) we use a holistic (complex) concept of competence, which integrates analytical as well as behavior-related aspects. Our understanding therefore corresponds with the conception proposed by Blömeke et al. (2015), who focus on "the latent cognitive and affective-motivational underpinning of domain-specific performance in varying situations" (p. 3), as well as Weinert (2001), the Curriculum-Instruction-Assessment Triade (Pellegrino et al. 2001), and the corresponding Evidence-Centered Design (Mislevy and Haertel 2006; Bley 2017).

Holistic competence models comprise a horizontal and a vertical layer. The horizontal competence structure (width of competence)—shaping the structural model (Hartig and Klieme 2006, p. 132)—represents theoretically assumed sub-dimensions of the particular construct, which are specified by internal cognitive and non-cognitive dispositions (National Research Council 2012, p. Sum-3). These internal dispositions required for performing situation-specific actions within a domain are not directly observable. They are only measurable through external observable behavior (=performance), which is evoked by test items reflecting workplace situations that depict the competence sub-dimensions. If a student is, for example, able to interpret the values presented by an SPSS-output of a correlation analysis correctly (=performance), the skill to judge outputs from relevant statistical software is attributed to this person. The test person's external response behavior results from combining internal cognitive and affect-motivational dispositions. Although we prefer a holistic concept of competence, we start from and therefore focus on cognitive dispositions of CCER (skills) within this study, because there are limited robust results in this field. Furthermore, non-cognitive affective-motivational dispositions—such as achievement motivation—are not explicated in our model, because they are not separately measurable (Shavelson 2012). However, they are implicitly covered by the actions that are required to solve our authentic test tasks (cf. chapter 3.2).

To facilitate making a differentiated statement regarding varying proficiency levels of students, the test items have to differ with respect to their level of difficulty. For the vertical competence structure (depth of competence)—shaping the proficiency level model (Hartig and Klieme 2006, p. 133)—various competence profiles are assumed. Through focusing on the particular degree of achievement, it provides information on the difficulty level of situational challenges and reflects the different proficiency levels of the particular construct. In line with internationally proven assessment standards, we assumed that competencies are supposed to be malleable and that the formation of competencies proceeds along a linear continuum (Blömeke et al. 2015, p. 7; Hartig 2007; Wilson 2005). For relating both the horizontal and the vertical modeling perspective to CCER, see sections "The domain of CCER: development of a structural model" and "Scaling CCER: development of a proficiency level model".

The domain of CCER: development of a structural model

Standards for the design of assessments suggest that relevant and representative observable evidences for typical research-methodological reviewing activities have to be identified in order to develop tasks that are valid as regards the contents on which they focus. This was performed through a domain analysis (Mislevy and Haertel 2006). In order to understand (a) which substantial *content areas* CCER refers to in detail and which *challenges* (prospective) VET-educators are expected to master in the field of research methods, as well as (b) which *competence dimensions* are relevant to managing the respective challenges (Wiethe-Körprich and Trost 2013), a systematic literature review was conducted. Furthermore, within focus groups experts who run (lecture on) a course on research methods for Human Resource Education and Management students of the LMU were consulted and students who attended this course preceding our test participants were asked to state typical research-methodical challenges/tasks, abilities that are required to master these challenges, and tasks which were particularly difficult to solve. The subsequent "big ideas" (Pellegrino 2010, pp. 17–18) were derived from the domain analysis:

(a) Content areas and typical challenges

One crucial result is that research-driven learning is commonly structured alongside the typical scientific research process. This is pointed out in detail by several authors (e.g., Rost 2007; Bühner 2011). Hence, the *contents*—that students should possess in the field of research methods—identified as domain-typical can be classified by four central categories: (1) problem definition; (2) methodology used to investigate the research question(s) of interest; (3) analysis, depiction, and interpretation of the results; and (4) discussion, and conclusions derived from the research findings. Furthermore, experts stated that the main challenges could be divided into "working with research-methodological conceptual procedures" (such as capturing a study's statement from the abstract) and "working with statistical issues" (such as interpreting statistical representations; for more examples see Table 1). Based on an analysis of the module descriptions for courses on research methods of all German university study programs in the field of Human Resource Education and Management (N = 42), we identified that the curricular emphasis is on scientific literature that deals with research questions answered through

applying quantitative—by contrast to qualitative—research methods. As a consequence, in this study we focus on quantitative research methods for defining CCER.

(b) Definition and dimensions of CCER

Inspired by Schweizer et al. (2011), we derived the following definition for *competencies in consuming empirical research* (CCER):

> *CCER include competencies which enable an individual to reflect, interpret, and evaluate critically empirical quantitative studies—which are based on educational-psychological as well as sociological research questions—with regard to the quality of their theoretical foundation, their research questions and design, their methodical procedures, and their results including the practical relevance.*

Influential conceptual frameworks, which serve for analyzing educational research competencies are generally based on the following two concepts: (1) The SDDS model (Scientific Discovery as Dual Search model; Klahr and Dunbar 1988) defined that the process of gaining scientific knowledge requires three main components: searching for hypotheses, developing research designs, and evaluating empirical evidence (including the interpretation of statistical data analysis). (2) The EBR model (Evidence-Based Reasoning model), which is for example used by the LeScEd group, differentiates the three steps analyzing, interpreting, and applying (Groß Ophoff et al. 2014; Brown et al. 2010). The first step (analyzing) is particularly of interest with regard to instructional research on statistical literacy within the field of mathematics (Groth 2007). On the basis of first results it is questionable if the three different components of those approaches really address different latent sub-competencies of educational research competencies and therefore, if they are actually empirically distinguishable. The LeScEd group shows that a one-dimensional instead of a three-dimensional model fits the data better (Groß Ophoff et al. 2014). Summarizing, it is noticeable that all approaches define a specific "statistical" component while the other aspects are summarized in different variations. This separation of a statistics dimension also becomes evident in practical approaches in the instructional educational science's field of research methodology. They often distinguish between two main dimensions: *research methods* and *statistics* (e.g., Renkl 1994; Onwuegbuzie 2001; Dunn et al. 2007). As discussed by the authors, the use of statistical procedures to answer research methodological questions frequently constitutes a difficulty for prospective educators. The researchers point out a negative attitude towards statistical contents—manifesting itself in statistics anxieties, emotional hurdles, and mental stress—that prevents students from solving statistical tasks. This goes along with the assumption that the differentiation between statistical and further elements of research methods is referred to the underlying interests and talents of the target group (prospective educators), which are shaped more socially than analytically (Holland 1959). Based on these considerations, we expect two central content-related dimensions of *competencies in consuming empirical academic studies*: *Research-methodological conceptual competencies* (DIM1) is the ability to reflect, interpret, and critically evaluate empirical academic literature with regard to the research-methodological categories applied to the study's structure, theoretical foundation, and research questions, the selected research design, as well as the description and interpretation of the

results including their practical relevance. *Research-methodological statistical competencies* (DIM2) is the ability to reflect, interpret, and critically evaluate the choice and the application of central statistical procedures which are used to answer research questions or to test hypotheses deduced from scientific problems of empirical research (Stark and Mandl 2001, pp. 5–6).

We suggest that prospective VET-educators should be able to review a holistic study. To master the relevant challenges that occur within the different research-methodological content areas, various situation-specific skills are required. In line with the idea of the Evidence-Centered Assessment Design approach (Mislevy and Haertel 2006), in Table 1 the two dimensions of CCER are further operationalized and we present a selection of the relevant evidence a student has to adduce in order to demonstrate that he or she has accomplished the respective research-methodological skill (Mislevy and Haertel 2006, p. 7).

Both the LeScEd project and our study on CCER include dimensions of conceptualization and statistics in order to operationalize latent sub-dimensions of the competence model. While LeScEd differentiates three dimensions—'information literacy', 'statistical literacy', and 'critical thinking' (Groß Ophoff et al. 2014, p. 254)—statistical and conceptual competencies are differentiated in our interpretation of the domain analysis. Conceptual competencies therefore cover aspects of 'information literacy' and 'critical thinking'.

Scaling CCER: development of a proficiency level model

For scaling a competence scale in proficiency levels, a continuous competence dimension—as it is used for CCER—is divided into discrete, ordinal categories (Fleischer et al. 2013, p. 8). Only if the items are distinguishable by a varying degree of difficulty can

Table 1 Situation-specific skills and according evidences of dimension 1 and dimension 2

Situation-specific skills of	Exemplary evidences
DIM1: research-methodological conceptual competencies:	
1.1 The student can understand and differentiate basic research methodological terms, concepts, and models	The common quality criteria for tests—objectivity, reliability, and validity—are explained correctly (including different kinds of these criteria)
1.2 The student can scrutinize studies' structure, rigor, and relevance	The research questions/hypotheses of a study are identified appropriately
1.3 The student can assess the appropriateness of research designs critically	The justification provided for the data collection method selected by the author(s) of a study is convincing related to the research question
1.4 The student can make and work with interpretations, causal explanations, and predictions	A scientific paper's results and conclusions are analyzed critically regarding their practical and scientific significance
DIM2: research-methodological statistical competencies	
2.1 The student can justify the selection of statistical routines	The author's decision to apply a correlation and a regression analysis respectively related to the scientific question is assessed correctly
2.2 The student can express the relevance of central quality criteria for procedures of statistical testing	Quality criteria for factor analyses—eigenvalue, explained proportion of total variance, and specificity—are identified correctly
2.3 The student can judge outputs from relevant statistical software	The important impact factors based on a presented output of a regression analysis are identified and interpreted correctly (based on significant β-values)

a differentiation between diverse proficiency levels be effected (Embretson 2002). The selection of the task features used to scale CCER was made in the light of the characteristics which had turned out to be significant determinants of the item's difficulty in previous studies on measuring situation-specific skills by using stage-oriented models, drawing notably on Blum et al. (2003), Kauertz and Fischer (2008), Winther and Achtenhagen (2009). According to these authors, the following three criterions are assumed to have a relevant impact on the difficulty of solving research-methodological tasks: (I) Kind of cognitive process according to the Cognitive System of the 'New Taxonomy of Educational Objectives' by Marzano and Kendall (2007, 2008); (II) Complexity concerning the number of content-related elements; (III) Degree of familiarity. A detailed description of these criterions—including their operationalization and examples—is presented in Appendix (Table 6).

Research questions

We addressed the following three research questions (RQs), which were derived from the presented theoretical explanations:

RQ 1 (Structural model): Are the two theoretically modeled dimensions of CCER empirically distinguishable?
RQ 2 (Quality of the test instrument): Do the empirical quality measures concerning the performance test instrument indicate:
RQ 2a) … that the central Rasch-modeling assumption of equal discriminability regarding all test items is met?
RQ 2b) … that the test items allow a reliable and valid measurement of CCER?
RQ 3 (Level model): Which levels of CCER can be defined by task characteristics that significantly determine the item difficulty?

Methods

Target group, course structure, and sample

As target group, undergraduates of the Human Resource Education and Management (HRE&M; in German: "Wirtschaftspädagogik") study program at the LMU were chosen. Their polyvalent educational profile prepares them for the various workplace settings where VET-educators are typically employed.

The students are offered a small group course on empirical research methods which integrates essential research-methodological content. Through focusing on empirical research methods, the course follows the trend towards an empirical research orientation which prevails in the field of educational sciences (Gesellschaft für Empirische Bildungsforschung 2012). It aims at two superordinate learning objectives that are compliant with the two key challenges of research methodology: (a) *reviewing* empirical academic literature and (b) independently *performing* an empirical research project. In order to develop these competencies, an innovative instructional design consisting of different course elements is provided. The course is offered every semester. With reference to the learning objective (b), an independent research project has to be performed and the results have to be presented in a short research paper. The test designed for our study addressed whether learning objective (a) is being achieved. The total grade for the

course is composed of both performance measures. Participation is compulsory for undergraduates who intend to write their bachelor theses in the field of human resource education. Alternatively, they attend a course and address their theses to the field of business administration. The test on CCER (for answering RQ1 and RQ2) took place at the end of the respective semester of the study program HRE&M at the LMU. Within our cross-sectional research study, test data are available for a total of 155 students. They were derived from the full surveys of four consecutive semesters ($n_1 = 54$, $n_2 = 30$, $n_3 = 23$, $n_4 = 48$)[1] starting in the winter term 2011/12. The students are on average in their sixth semester (SD = 1.34) and two-thirds of them are female.[2]

Intended information for the high-stakes assessment and test design

Our test of CCER was designed for a real 60-min exam under high-stakes conditions. Therefore, the test result has considerable consequences for the respective student: It decides if the respective test person has passed or failed the course, and enters the students' final grade for the study program. Compared with a voluntary survey without important consequences, high-stakes testing situations lead to higher motivation and a significantly lower probability of guessing and skipping test tasks. So, the number of missing responses will be minimized. Furthermore, there is no sample self-selection effect. Students commonly dedicate little effort to low-stakes testing assessments as an act of prioritization and to save their energy for meaningful academic tasks. These points reduce the score validity of low-stakes testing approaches in the most basic sense (Wise and DeMars 2005; 2006; Wolf et al. 2015). Missing responses for omitted items are usually not random. This may lead to biased estimates of item and person parameters (Mislevy and Wu 1996). However, at least for low-stakes testing assessments, several authors propose ignoring missing responses instead of scoring them as incorrect (de Ayala et al. 2001). But if this results in an unequal distribution of omitted items concerning different competence dimensions (e.g. relatively more statistical questions are skipped), the consequence may be a misjudgment of the competence structure. In a high-stakes power test—as we intended to design—it can be expected that omission occurs when participants do not know the answer and therefore missing at random is less plausible (Mislevy and Wu 1996). Instead, there is a significant correlation of ability and the number of missing responses (Pohl et al. 2014). Despite all its benefits, high-stakes testing conditions have—with regard to assessment development—several restrictions concerning the number and administration opportunities of the test tasks. That means: (a) With regard to local conditions we had to use a paper-and-pencil test instead of a more realistic technology-based approach. (b) The number of test tasks was limited by the test time, because each participant had to get exactly the same items. And (c) a substantial number of easy tasks had to be implemented, because a student had to reach 50% of the maximum score to pass the test.

With regard to the high-stakes testing conditions and the intended assessment information, as well as on the basis of our competence model and the identified evidences, we developed 26 paper-and-pencil test items. They were designed along the lines of the

[1] The tasks were all the time not accessible for students, so that participants of earlier semesters did not have systematic disadvantages (ANOVA regarding the total scores for the four groups: F value = 1.087, p = .357 > .05).

[2] Test data were analyzed anonymously. Biographical data derived from course registration information.

typical research process set out by Rost (2007, p. 26). Each content area of the research process was covered by different situations in the form of items which depict the two theoretically expected dimensions of CCER (DIM1: 11 items; DIM2: 15 items). In order to cover the whole spectrum of proficiencies, and taking into account that undergraduates need 50% of the maximum score to pass the test, we constructed tasks of all degrees of difficulty. Development requirements originating from (i) *content-related instructional science* (in German: 'Fachdidaktik'), (ii) *cognitive psychology*, and (iii) *psychometrics* were considered for constructing the items.

From the perspective of (i), *content-related instructional design*, the standards for designing authentic assessments had to be fulfilled, such as realistic illustrations, orientation towards real professional circumstances/environments, permission of judgements and reflections, focusing actions and the comprehension of these actions, replicating or simulating tasks which originate from the occupational routine, and inspirations for further learning (Janesick 2006, p. 4; Mislevy and Haertel 2006; Weber et al. 2014).

For this reason, all test items are based on one empirical study (published in the *Journal of Pedagogical Psychology*) which bears the title "Personal responsibility for academic achievement: Dimensions and correlatives" by Koch (2006). This research study deals with the effect of the latent construct 'personal responsibility' for academic achievement. The implementation of different authentic situations requiring CCER within a superordinate context—the real study by Koch (2006)—offers various advantages: Koch's article was selected due to the probable attractiveness that the topic would hold for the students as well as the students' involvement triggered by the topic. It can be expected that the study's context equally constitutes relevant issues for all test persons and that all students should obtain a comparable interest, previous knowledge, and experiences concerning the addressed content area. Consequently, situation-specific affect-motivational effects are largely negligible. Furthermore, this paper was chosen since the statistical sophistication is consistent with the abilities which can be expected as concerns consumer behavior, which the participants have acquired during the course. The study had not been utilized by the instructors during the course on empirical research methods or during other courses. Therefore, the tasks outline new situations. All tasks were presented using real data, text excerpts, and figures. They follow the judgment of a complete research process. Correspondingly, the test person—considered qua consumer of research—had to grasp central contents and information about the paper (cf. sample item 1; see Appendix (Table 5) fro the whole item pool); evaluate sources and methods used for the survey as well as methods for analyzing data (e.g. correlation and factor analyses) performed in the study; and analyze, interpret, and assess statistical diagrams [cf. sample item 20; see Appendix (Table 5)]. In addition, transfer cases based on so-called "what-if tasks" had to be solved. These tasks go beyond the research methodological situations covered by Koch's (2006) study. In one of the "what if tasks" the test persons had to create an experimental design for evaluating the effectiveness of a training on strengthening personal responsibility for academic achievement what is not part of Koch's (2006) study.

Apart from few matching tasks, the test items are designed primarily using open-ended formats in terms of performance tasks and analytic writing tasks (cf. the Collegiate

Learning Assessment approach by Shavelson 2008). Subsequently, two examples of test items are illustrated:

Item 1 is assigned to the content area of 'problem definition' and refers to skill 1.2 concerning the conceptual dimension of CCER (cf. Table 1). The abstract of Koch's (2006) article illustrated in Fig. 1 is presented. Based on this, the student is prompted to identify the paper's two central research questions.

Item 20—allocated to the content area of 'results' and depicting skill 2.3 of the statistical dimension of CCER (cf. Table 1)—presents the subsequent output for a correlation analysis performed in Koch's (2006) study. To decrease extraneous cognitive load, some side notes and highlighting elements are integrated in the output (cf. Fig. 2).

The item quotes the following statement which a researcher had framed based on the output: "Final university examination grade and commitment to the studies are two independent criteria for success". The test persons are requested to mark the value that resulted in the given statement within the presented output (e.g. by circling) and to explain why the argument is derived correctly.

As all test items refer to the described study, the test persons do not have to become acquainted with a new context in every task [*cognitive psychological perspective*, (ii)]. Additionally, appropriate linguistic complexity of the tasks' instructions, reasonable signaling, and the avoidance of redundancies were considered when constructing the test tasks [in accordance to Bley et al. (2015)]. Therefore, the extraneous cognitive load as well as the time for introducing tasks can be reduced (van Merriënboer and Kirschner 2013, p. 22).

Despite of the advantages of embedding test-items in one real study (e.g. authenticity), we are aware that this approach results in the fact that the test instrument is subsumed under a single anchor. As a consequence, the assumption of local stochastic independence could be violated [*psychometric perspective* (iii): Koller et al. 2012]. Therefore, we made a great effort in providing all the necessary information (e.g. text excerpts or statistical outputs from the study) relevant to solve each new situation. The result of the

Personal Responsibility for Academic Achievement: Dimensions and Correlatives

Abstract. This article adopts the concept of «personal responsibility», developed in the field of social psychology, for the setting of higher education. The main thesis is that personal responsibility consists of three dimensions: Personal relevance, knowledge about strategies of action, and action control. It is supposed that students who feel personal responsibility for academic achievement will show higher effort to achieve and better grades in examinations. In Study 1 (subjects: 133 undergraduate students) these three dimensions were tested using factor analysis; results show that they are independent from each other. In Study 2 (subjects: 223 postgraduate students), the influence of these three dimensions on study effort and grades in examinations was investigated. Regarding undergraduate studies, the dimensions relevance and knowledge about strategies had a significant influence on study effort and grades in examinations. Implications of the findings concerning the interaction of the three dimensions in order to explain academic achievement are discussed.

Keywords: responsibility, academic achievement (college), higher education, intrinsic motivation, social psychology

Fig. 1 Abstract of the study by Koch (2006)

Tabelle 2
Korrelationsmatrix der Variablen für das Universitätsstudium (N = *223*)

		M	*SD*	1.	2.	3.	4.	5.
personal responsibility	1. Bedeutsamkeit	4.37	1.05	(.81)	.31**	.12	.35**	.33**
	2. Ziel-/Prozessklarheit	4.13	1.05		(.76)	.50**	.25**	.39**
	3. Kontrolle	4.36	1.14			(.85)	.01	.24**
academic achievement	4. Studienengagement	3.21	1.06				(.53)	.12
	5. Staatsexamensnote	7.82	1.99					

Anmerkungen. Angaben für Cronbachs α in der Diagonalen; $^{**}p < .01$; $^{*}p < .05$.

Fig. 2 Correlation matrix presented in the study by Koch (2006)

non-parametric T11 test (p value = .824) (Ponocny 2001) shows that this procedure was successful.

By discussing all items with seven experts, who are instructors in empirical research methods for students of HRE&M at the LMU, content as well as substantive validity were ensured. The experts were asked to evaluate the items with respect to the relevance of content-related aspects, the appropriateness of the tasks' scoring, as well as students' cognitive solution processes that are intended to be activated by the test tasks [cf. Appendix (Table 6)]. Slight revisions of our tasks were performed corresponding to the experts' assessment.

Handling of missing responses and coding

High objectivity in implementing the test can be assumed, because of legally defined examination rules. No data set had to be eliminated. As expected, the number of missing responses is quite low (1.41%) and all of them can be classified as "omitted responses", because all participants received exactly the same test and the missing responses were spread over the whole test not only over the last items. Because the number of missing responses correlates significantly with persons' abilities (Kendall's $\tau = -.243$, $p = .000$), we interpreted a missing response as an inability in item answering.

Our scoring guide includes a best-practice solution of the written exam as well as a description of each optional scoring category [see Appendix (Table 5)]. Twelve items were scored binary and for 14 items students could earn partial credits (three response categories). The appendix (Table 5) explains the scoring rules for each item in detail. In correspondence with the scoring guidelines, two trained raters coded the students' answers independently. These raters are research and teaching assistants on an expert level who teach in empirical research methods for students of HRE&M at the LMU. An interrater reliability Kappa (Fleiss and Cohen 1973) of .940 was attained. This accounts for a high level of agreement.

Instrument for an expert-based rating of the tasks' difficulties

For developing the *proficiency level model for CCER (RQ3)*, the seven experts introduced in chapter 3.2 were asked to evaluate the test items corresponding to the three characteristics assumed to determine their difficulty. A written questionnaire was used for this rating. It was largely performed on three- or four-point Likert scales [for the operationalization of the criterions for the tasks' difficulty see Appendix (Table 6)]. Before the rating, the experts participated in training to explain the design of the items and the meaning of the different degrees of the task criteria which had to be evaluated. The theoretical estimation of the test items' difficulty levels was carried out a priori and therefore independently from the performance test's results. In order to find consensual ratings, the responses given by each expert were discussed within the group of all raters and the research team (Kuckartz 2014; Wahl 1982). This procedure served to make sure that all experts correctly understood what the variables of the questionnaire aimed at. Slight adjustments of the original coding were made in response to the insights derived from the focus group discussion.

Methods of data analysis

For *empirically validating the theoretically assumed structural model for CCER (RQ1)* and for *examining the test instrument's quality (RQ2)*, the written students' exams were analyzed using psychometric models belonging to the IRT. RQ1: Two central Rasch-models were applied—a one-dimensional and a two-dimensional Partial-Credit-Model (PCM; Masters 1982; Adams et al. 1997)—by using the software ConQuest 3.0 (Wu et al. 2007).

The central advantage of Rasch-models—namely that individuals' ability parameters are estimated independently of the tasks used to compare the individuals—is only valid if there are equal discrimination values of all items in a test (RQ2a). To test this, median scoresplit analyses—Andersen-Likelihood-Ratio-Tests (Andersen 1973) and Wald-Tests (Koller et al. 2012, pp. 77–79)—were executed, using the eRm-package belonging to the software R (version 3.1.2; Mair and Hatzinger 2007).[3] The quality of the test items (RQ2b) was investigated by calculating and evaluating (i) the scaling of the individuals' ability parameters as well as the items' difficulty parameters, (ii) the EAP/PV (expected a posteriori/plausible values) reliability, (iii) the curve of the total test information function, and (iv) the wMNSQ (weighted Mean Square) values.

For the *expert-based determination of proficiency levels (RQ3)*—following Hartig (2007)—we chose an additive and linear regression model for the coherence between the item features (independent variables) and the IRT-based item difficulty parameters (dependent variable). The 50%-thresholds resulting from the IRT-scaling were used as item difficulty values.

Results and discussion

Empirical validation of the structural model: RQ 1

Based on the finding of the LeScEd study (Groß Ophoff et al. 2014, p. 266), where the one-dimensional model fitted the data best, a one-dimensional PCM was tested against a two-dimensional between-item-multidimensionality PCM in order to identify if the two expected dimensions of CCER are empirically distinguishable. Thereby, DIM1—the conceptual dimension—is described by 11 test items (1–6, 17–19, 25–26) and DIM2—the statistical dimension—by 15 test items (7–16, 20–24). According to our theoretical expectations, the information criteria BIC, AIC, and CAIC—which show lower values for the two-dimensional PCM—provide empirical evidence for a better fit of the two-dimensional model (cf. Table 2). This finding is confirmed by the Likelihood-Ratio-Test according to Martin-Löf (Glas and Verhelst 1995, pp. 86–89) which became significant on the 5%-level (Chi square = 25.90; df = 2; p = .000). The moderate correlation between the two latent dimensions of CCER (r = .678; covariance = .226) supports a two-dimensional solution. For the following analyses, the estimated parameters for the better fitting two-dimensional PCM are used.

Quality of the test instrument: RQ 2a) and 2b)

RQ 2a) refers to investigating whether the central Rasch-modeling assumption of equal discrimination parameters is fulfilled for all items. For performing the Andersen-Likelihood-Ratio-Tests and Wald-Tests we determined a significance level of 20%. For

[3] Scoresplit analyses [by using Andersen-Likelihood-Ratio-Tests (Andersen 1973) and Wald-Tests (Koller et al. 2012)] has a long tradition as well as a high power to examine this assumption (for a detailed discussion see Rasch 1961 and/or Glas and Verhelst 1995).

Table 2 Fit statistics for the one-dimensional PCM in comparison with the two-dimensional PCM

	One-dimensional PCM	Two-dimensional PCM	
Deviance (LR-test)	6073.97	6048.07	Difference = 25.90
Number of estimated parameters	41	43	
BIC	6280.75	6264.93	
AIC	6155.97	6134.07	
CAIC	6321.75	6307.93	

The degrees of freedom result from the difference between the estimated parameters

taking into account the alpha-error-cumulation, a Bonferroni correction (Abdi 2007) was conducted for the Wald-tests through dividing the defined significance level by the number of performed tests per dimension. The Andersen-Likelihood-Ratio-Test is not significant for the conceptual dimension ($p = .503 > .2$) but for the statistical dimension ($p = .003 < .2$). However, the results of the Wald-Tests show that no z value is significant. Consequently, our test allows a separated statement regarding task-difficulties and test persons' abilities.

RQ 2b) examines whether the test items allow for a reliable and valid measurement of CCER. For all test persons and for all test items, (i) the *scaling* of the *individuals' abilities* and of the *items' difficulties* can be illustrated by a Wright map regarding the two scales (cf. Fig. 3; Wilson 2005, pp. 90–98). Based on a maximum of 40 the test score value which was achieved on average is 23.35 (SD = 5.69). The *ability parameters* (EAP/PV-estimators) range from −.952 to .900 logits for the conceptual dimension and from −1.664 to 1.547 logits for the statistical dimension. They are normally distributed (Kolmogorov–Smirnov test; DIM1: $p = .564$; DIM2: $p = .402$). The *difficulty parameters* define the latent variable of conceptual competencies on a scale from −1.506 to .748 logits and the latent variable of statistical competencies on a scale from −1.809 to .663 logits. Correspondingly there is a lack of items with a very high degree of difficulty. This effect was to be expected, because in order to regulate the CCER exam failure rate a considerable number of items of an easy and moderate difficulty had to be included.

With regard to the two dimensions, which are assumed based on the empirical analysis, the (ii) *EAP/PV reliability*—which is comparable with Cronbach's alpha (Adams and Wu 2002, p. 152)—shows moderate values of .548 for the conceptual and .737 for the statistical dimension. It is assumed that the low number of items—which is a consequence of the explained high-stakes testing conditions—is responsible for the moderate reliability values. Since the reliability value only expresses how precise the measurement is with respect to the complete ability spectrum, the (iii) *Wright map* is considered additionally. Their advantage compared with the EAP/PV-reliability value is that the Wright map indicates how accurate the measurement is regarding different ability areas. As Fig. 3 illustrates, with the expectation of high ability parameters, the item difficulties and student abilities corresponding well. This supports a precise measurement as the test also covers the ability level of students with a very high and a very low degree of CCER in a differentiated way.

In order to examine (iv) potential Differential Item Functioning (DIF)-effects with respect to the test persons' gender, an Andersen-Likelihood-Ratio-Test (Glas and

Verhelst 1995) as well as Wald-Tests were calculated. The test results show that no DIF-effect exists for any item.

The (v) *wMNSQ values* of the 26 test items are located within a range of .89–1.15 [cf. Appendix (Table 7)]. Hence, all items show a good to very good fit since they are situated within the strict interval of .80 ≤ wMNSQ ≤ 1.20 postulated for the PISA study (OECD 2014, p. 151). The corresponding t values range from −1.0 (> −1.96) to 1.7 (<1.96) and are therefore non-significant. As all items are of a high quality, no item has to be excluded from the test.

Our study is oriented towards Messick's (1989, 1995) concept of validity which integrates different validity evidences. Accordingly, the test instrument designed for measuring CCER meets the requirements of (1) *content validity* as great importance was attached to the design of authentic test tasks derived from a real empirical research study. To ensure that the constructed tasks are valid with regard to content-related aspects, we discussed them with experts (lecturers) of the addressed course on research methods within focus groups. These focus groups were also carried out for ensuring (2) *substantive validity* through discussing the cognitive solution processes (including the tasks' scoring) intended to be activated by the test tasks with the experts. Furthermore, (3) *psychometric validity* is indicated through the good values of the presented fit indices. Finally, (4) in a first access regarding the test instrument's *external validity*, we considered the relationship of the grades the HRE&M students achieved in their course on research methods and the IRT-based ability parameters for the one-dimensional model. The grades—determined by the course teachers (independently from the research group)—describe a combination of reviewing empirical academic literature and autonomously performing empirical research projects. The Spearman rank correlation coefficient ($r = -.756$, p value = .00) indicates that performances within the course and performances measured by the CCER test point in the same direction. That means, students who attained a high (low) ability parameter in the test on CCER also displayed a good (weak) performance concerning the course grade (whereby, the smaller the number of the grade, the better the performance). The effect size of the relationship can be interpreted as meaningful. But, it has to be noted, that the validity criterion of the course grade is not able to differentiate between the performance regarding the conceptual and the statistical dimension. Ergo, it does not allow a separate validation for the two dimensions. To sum up, based on the examined quality measures (i)–(v) as well as on the considerations addressing the different aspects of validity concerning the instrument for measuring CCER, a successful test construction can be assumed.

Level model: RQ 3

Our last RQ focuses on defining levels of CCER by task characteristics that significantly determine the item difficulty. Therefore, in a first step we determined the level of agreement between the experts by using the interclass-correlation (ICC) (Shrout and Fleiss 1979). The ICC values confirm a strong agreement between the raters concerning all three characteristics (cognitive process: ICC = .871; complexity: ICC = .818; familiarity: ICC = .854). Table 3 outlines the estimation of the common predictive power of all task features (adjusted R^2) and the identified predictive

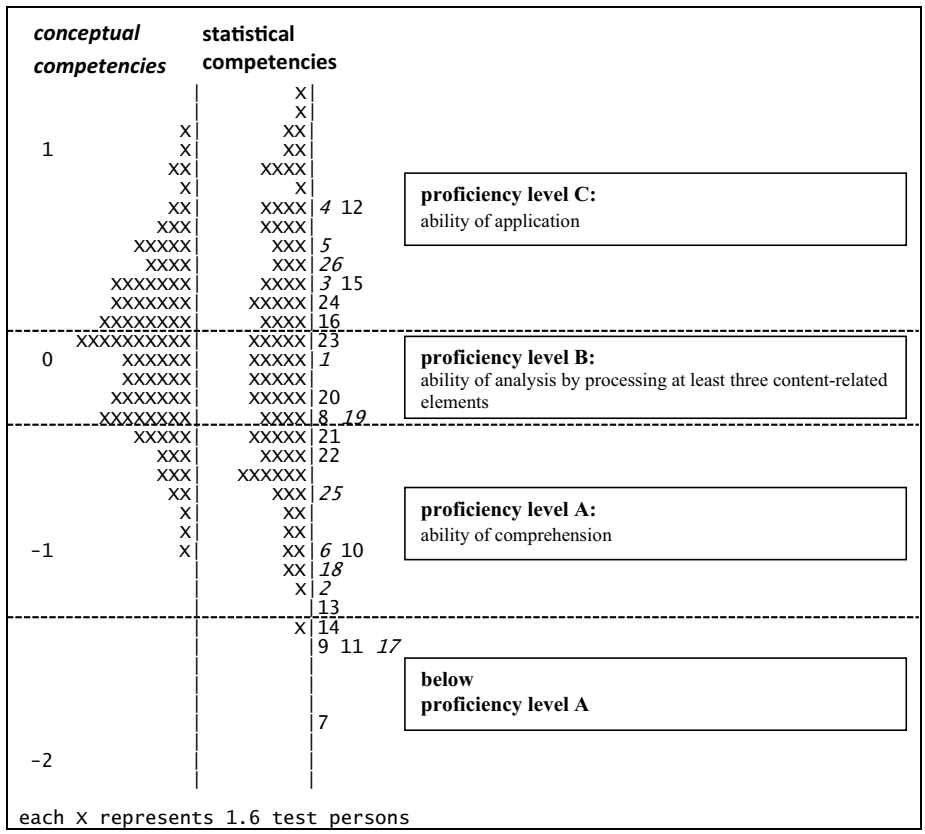

Fig. 3 Level model for CCER (based on the Wright map using the two-dimensional PCM); distribution of the person and item parameters

effects of each task characteristics based on a multiple regression analysis (all prerequisites are met).[4]

The specified model explains 77% of the total variance ($F = 12.965$, $p = .000$), which constitutes a high proportion. The four task features marked in italic significantly influence the item difficulty on an alpha level of 10% and were therefore used for the level modeling. The "familiarity" of the situations seems to be neither conducive to nor hindering of the task solution. Three proficiency levels could be determined: (A) 'the ability to comprehend research-methodological terms and concepts', (B) 'the ability to analyze research-methodological situations by processing several content-related elements', and (C) 'the ability to apply research-methodological concepts and procedures' (cf. Fig. 3). For all three levels, items covering the conceptual dimension and items covering the statistical dimension of CCER were constructed successfully.

In Table 4 the proficiency levels are defined by the respective logit values for the thresholds. Additionally, the table presents the proportional allocation of the test persons to the proficiency levels. It has to be emphasized that the empirically defined three-stage level model has an explanatory power for 100% of the test persons concerning the conceptual dimension and for 98.71% referring to the statistical dimension. The

[4] The Variance Inflation Factor (2.539–5.573) and Tolerance (.179–.394) indices for all predictors do not show any critical values. Therefore, no multicollinearity between any variables exists (Bühner and Ziegler 2009, pp. 681–682).

Table 3 Regression coefficients (standardized and non-standardized) for the specified model

Adjusted $R^2 = .770$	b (non-std.)	Standard error	Beta (std.)	t value	p value
Constant	−1.779	.315		−5.643	.000
Cognitive process: comprehension	.451	.252	.251	1.791	*.090*
Cognitive process: analysis	.917	.349	.559	2.629	*.017*
Cognitive process: application	1.226	.379	.682	3.239	*.005*
Complexity: 2 elements	.125	.254	.069	.490	.630
Complexity: at least 3 elements	.625	.328	.412	1.907	*.073*
Familiarity: moderate learning opportunities	.382	.299	.199	1.276	.218
Familiarity: few learning opportunities	.309	.283	.189	1.092	.289

N = 26 (number of test items for both dimensions of CCER); reference categories for the three characteristics: cognitive process: retrieval, complexity: one element has to be processed, familiarity: many learning opportunities are provided

Table 4 Allocation of the test persons (N = 155) to the proficiency levels

Proficiency level	Level threshold (in logits)	Proportion of students on the conceptual levels (in %)	Proportion of students on the statistical levels (in %)
Below Proficiency level A		.00	1.29
Proficiency level A (ability of comprehension)	−1.328	27.10	36.77
Proficiency level B (ability of analysis by processing at least three content-related elements)	−.237	28.39	17.42
Proficiency level C (ability of application)	.072	44.52	44.52

following description of the levels is based on analyzing the specific requirements of the tasks.

After having graduated from the course on research methods, almost 100% of the HRE&M students are able to comprehend essential concepts and procedures of empirical research-methods (*proficiency level A*). We interpret this level as the criterion for passing the bachelor degree in HRE&M studies. Item 18 (item difficulty = −1.063 logits), for instance, addresses the ability to recognize and establish that the author's decision to apply a written survey with a closed-ended response format is reasonable (e.g., because a larger number of persons can be questioned when using a closed format). Hence, in accordance with Marzano and Kendall (2007, p. 40), the learner has to mix "new knowledge"—meaning the information contained within the presented extracts of the study—"and old knowledge residing in the learner's permanent memory" to solve tasks of this level.

The major part of the students (72.91% for the conceptual and 61.94% for the statistical dimension) even reaches the level of conducting analyses of research-methodological situations by linking a crucial number of content-related elements (*proficiency level B*). Examining the items assigned to this level, analytical processes such as "specifying", "matching", and "classifying" (Marzano and Kendall 2008, pp. 18–19) are needed. For example, to solve item 20 (item difficulty = −.274 logits) (outlined in section

"Research questions") correctly, two mental processes are relevant: (a) matching—as scientific authors' statements have to be compared with statistical values; and (b) specifying—as the test persons have to "identify […] principles that apply to a specific situation" (Marzano and Kendall 2007, p. 50).

A proportion of 44.52% regarding both dimensions is even able to achieve the ability of applying research-methodological concepts and procedures (*proficiency level C*). These students are able to apply mental processes of knowledge utilization such as "experimenting", "investigating", "decision making", and "problem solving" (Marzano and Kendall 2007, p. 51). Item 26 (item difficulty = .413 logits), for instance, prompts the students to create an experimental design based on a follow-up research question to evaluate the effectiveness of responsibility training and its impact on academic achievement. Regarding the process of experimenting, this task requires "testing hypotheses for the purpose of understanding some physical or psychological phenomenon" (Marzano and Kendall 2008, p. 20).

Conclusions

Discussion and limitations

The results of the study show that we succeeded in designing a reliable and valid test instrument for assessing (prospective) VET-educators' competencies in consuming empirical research. With regard to the competence structure, our results indicate that the two considered dimensions—frequently referred to in practical applications (e.g., Renkl 1994; Onwuegbuzie 2001; Dunn et al. 2007)—also become empirically evident. However, with the used approach, it cannot be excluded that empirically a model with more than two dimensions will fit the data better than with two dimensions. Existing finding of the LeScEd group assume a one-dimensional solution. We explain this deviation mainly on the basis of different test conditions (low-stakes vs. high-stakes testing) and a different handling of missing values. While under low-stakes testing conditions omitted items are often ignored (e.g., Groß-Opphoff et al.), under high-stakes testing conditions we have evidence that omitted items could be reduced to the fact that the participant does not know the answer. Despite the strong restrictions of the high-stakes testing approach (small sample size and a limited number of test tasks) the values of quality (wMNSQ and t values, assumption of equal discriminability, and test information curve) can be interpreted as sound; only the EAP/PV reliability shows moderate values. Based on the measures for the item quality, no item has to be excluded, and therefore our high standard of theoretically based content validity is fulfilled within the final item pool. In our opinion this positive result is attributable to our decision to follow ambitious standards for test designs and validation (Curriculum-Instruction-Assessment Triade, Evidence-Centered Design, high-stakes testing). But, this decision is also linked to the limitation that the study lacks generalizability, which constitutes a further crucial criterion of validity according to Messick (1995), but is not primarily being dealt with within this article. To provide a generalizable evaluation of prospective VET-educators' CCER, it would be interesting to analyze how the test participants perform in CCER follow-up tests. Additionally, the test on CCER has to be implemented as a high-stakes testing exam for other research methodological training courses within different institutions and in different courses of studies which prepare future VET-educators. These institutions should comprise selected

universities, whose module descriptions for courses on research methods we analyzed. The results of the CCER level model specification show that two of the three defined task characteristics (cognitive processes and complexity) are able to explain nearly 100% of the prospective VET-educators' CCER abilities. Besides the generalizability aspect, further limitations are (1) a constrained pool of items, (2) limited criteria for external validity, (3) constrained statements regarding the test fairness, and (4) the test focus on quantitative research methods: (1) Only a constrained selection of situations requiring CCER could be presented within the test. This is based on the high-stakes testing conditions in form of a real exam and on the limited test time. In future large-scale research designs additional CCER test situations—that prompt further statistical procedures (e.g. cluster analyses)—should be included for instance as a multi-matrix design. Furthermore, to measure also the abilities of very high-performing students sufficiently, items with a very high degree of difficulty have to be integrated into the test on CCER. (2) First indications for confirming the external validity of our test instrument are provided. However, in order to make a separate statement regarding the external validity of both scales (the conceptual and the statistical scale) for assessing CCER, external criteria for the two dimensions are necessitated. (3) Due to constrained possibilities for collecting demographic information under high-stakes conditions and related reasons of anonymity, the test fairness could only be examined for the covariate 'gender'. (4) So far, CCER focus on quantitative research aspects. As a consequence, an operationalization of the qualitative part constitutes a crucial desideratum for further research.

Implications

From the point of capturing and promoting the development of CCER during prospective VET-educators' studies, it is relevant to scale this competence according to features that might determine the difficulty of corresponding tasks. As not all students achieved the learning goal to apply research-methodological concepts and procedures, there is a necessity for a stronger focus on teaching activities in order to inspire learning processes which support the development of abilities to master application tasks. The identification of significant task characteristics can help to design learning environments as well as test tasks. The constructed test tasks vary systematically with regard to the identified characteristics determining the item difficulty. Apart from applying them for assessing CCER they can also be used as learning tasks in order to (further) develop CCER. As hardly any valid tests are available in the field of higher education, usually a minimum score value of 50% of the overall achievable score for passing an exam is used. On the basis of a substantial proficiency level model, a criterion for passing learning goals could be provided. As a consequence, grades could be based on such an a priori defined criterion instead of a more arbitrary defined minimum score value. The superordinate objective must be to implement validated test instruments with defined criterion-based proficiency levels in the form of an adaptive test design.

Authors' contributions

Both authors contributed substantially to this work. They designed the study, modeled the test items, implemented them, and participated in drafting and discussing the manuscript at all stages. Both authors read and approved the final manuscript.

Authors' information

Michaela Wiethe-Körprich studied Human Resource Education and Management at Ludwig-Maximilians-University in Munich from 2007–2011. Since 2012 she has been a research and teaching assistant at the Institute for Human Resource

Education and Management, Munich School of Management, Ludwig-Maximilians-University in Munich. Sandra Bley studied Human Resource Education and Management at Georg-August University in Göttingen from 2001–2006. She holds a Master of Business Research (MBR, 2008) and a doctoral degree (Dr. oec publ. 2010) from Ludwig-Maximilians-University in Munich. Since 2011 she has been a senior researcher at the Institute for Human Resource Education and Management, Ludwig-Maximilians-University in Munich.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The data supporting the authors' findings are provided by the authors on request.

Ethics approval and consent to participate
It is confirmed that the study was performed according to the ethical principles that are relevant for writing scientific studies.

Appendix

See Tables 5, 6 and 7.

Table 5 Instruction and scoring rules of the CCER test items

Item		Scoring
1	Please derive the two main research questions from the abstract of the study by Koch (2006)	0 = no research question is derived correctly 1 = one of the research question is derived correctly 2 = both research questions are derived correctly
	You can find some selected sentences of Koch's study below. To which part of the introduction do these sentences refer? Parts of an introduction: A) Definition/defining the research topic B) State of the art C) Relevance of the problem D) Research gap	
2–5	Sentence 1: '*A long duration of study, subject changes and dropouts are quite characteristic for university studies in Germany [...].* '	0 = identified aspect is wrong
	Sentence 2: '*In psychology, responsibility is closely related to perceived control and self-efficacy [...]. These two constructs then influence study success [...].*'	1 = identified aspect is correct
	Sentence 3: '*A person is self-determined if he takes responsibility for his own actions. We assume that self-determined students are more likely to demonstrate intrinsic motivation for their studies. This in turn seems to be desirable in pedagogical contexts [...].*'	
	Sentence 4: '*Schlenker* et al. *[...] introduce a social-psychological model of personal responsibility and examine its application to study success.*'	
6	What is the difficulty in operationalizing a latent construct? Please explain	0 = no aspect of difficulty is explained correctly 1 = one aspect of difficulty is explained correctly 2 = two aspects of difficulty are explained correctly
7, 9, 11, 13	Table 1: factor-analyzes "personal responsibility" (main component analysis, varimax rotation, N = 133)	0 = information is wrong, insufficient or not denoted
	Which information is given to the reader by the marked numbers: "1"/"2"/"3"/"4"?	1 = information is denoted correctly

Table 5 continued

Item		Scoring
8, 10, 12, 14	Table 1: factor-analyzes "personal responsibility" (main component analysis, varimax rotation, N = 133)	0 = information is wrong, insufficient or not assessed
	Please assess the information given by the marked numbers "1"/"2"/"3"/"4"?	1 = information is assessed correctly
15	Factor analysis provides us with various quality measures to assess latent constructs. Please name two quality measures and explain under which conditions the measure is assessed as "well fulfilled" (an approximate value is sufficient if you want to specify numbers)!	0 = no measure is outlined correctly AND no condition is explained *OR* One measure is outlined correctly BUT associated condition is wrong (explained) 1 = one measure is outlined correctly AND the associated condition is correctly explained *OR* Two measures are outlined correctly BUT for neither of them the associated condition(s) are (explained) correctly 2 = two measures are outlined correctly AND for both of them the associated condition(s) are (explained) correctly *OR* Two measures are outlined correctly BUT only for one of them the associated condition(s) are (explained) correctly
16	The quality of the presented factor analysis cannot be assessed, if the assessment is solely based on Table 1. Which aspects are missing for a complete assessment of the quality? Please outline two aspects!	0 = no aspect is outlined (correctly) 1 = one aspect is outlined correctly 2 = two aspects are outlined correctly
17	Data collection is carried out in Koch's (2006) study by means of a written survey with a closed answer format. Which other data collection methods do you know? Please name four additional methods!	0 = one or no method is outlined correctly 1 = three or two methods are outlined correctly 2 = four methods are outlined correctly
18	How do you assess the decision in this study to use a written survey with a closed answer format? Please justify your assessment with two arguments!	0 = no assessment and/or no argument (on the grounds) 1 = assessment and one correct argument (on the grounds) 2 = assessment and two correct arguments (on the grounds)
19	The summary (Koch 2006, p. 1) states: "It is claimed, that personal responsibility for study success […] has a positive effect on study management and performance." Would you prefer a correlation or a regression analysis to examine this statement? Please justify your decision with two arguments!	0 = no decision and/or no argument (on the grounds) 1 = decision and one correct argument (on the grounds) 2 = decision and two correct arguments (on the grounds)
20–23	Based on the presented correlation matrix included in Koch's (2006) study, a researcher interpreted the following statement:	
	(1) "Final university examination grade and commitment to the studies are two independent criteria for success."	0 = no or incorrectly marked value AND no or incorrect explanation
	(2) "Personal responsibility for your own studies and study success are positively correlated."	1 = correctly marked value BUT no or incorrect explanation
	(3) "The two sub dimensions "clarity of purpose" and "significance" are dependent factors."	*OR*
	(4) "Measuring accuracy of the sub dimension "significance" can be assessed as good."	Incorrectly marked value BUT correct explanation

Table 5 continued

Item		Scoring
	Please mark the value that resulted in the given statement (1)—(4) within the presented output (e.g. by circling) and explain in one sentence why the argument is derived correctly!	2 = correctly marked value AND correct explanation
24	Correlation analyses were supplemented by regression analyses. These analyses also show a positive effect of personal responsibility and study success. Nevertheless, the author argue in the conclusion part, that the results do not have any predictive character. That means variance differences in study success cannot traced back casually to variances in taking personal responsibility for their own study	0 = wrong or no aspect is denoted and justified
	Why could the authors come to this conclusion? Please denote a central aspect and justify your explanation!	1 = correct aspect is denoted BUT wrong or not justified
		2 = aspect is denoted and justified correctly
25	In a following research project, Koch and colleagues developed a training to strengthen personal responsibility of students for their studies	0 = no or one column is sketched correctly
	Please sketch a suitable experimental plan in the usual matrix format for evaluating the effectiveness of the new developed training!	1 = two or three columns are sketched correctly
		2 = all four columns are sketched correctly
26	Please describe each component of a (quasi-) experimental design for evaluating the effectiveness of the newly developed training!	0 = no or one component is described correctly
		1 = two or three components are described correctly
		2 = four or five components are described correctly

Table 6 Criteria for the tasks' difficulty, including their operationalization and examples

Criterions for the tasks' difficulty including their operationalization	Examples
(I) *Kind of cognitive process according to the Cognitive System of the 'New Taxonomy of Educational Objectives' by Marzano and Kendall* (2007; 2008)	(1) A student who is able to reproduce the concept of 'reliability' has attained the level of retrieval
Four response categories:	(2) A student who is able to decide which data collection method is suitable depending on the presented investigation context has achieved the level of comprehension
(1) Retrieval	
(2) Comprehension	
(3) Analysis	(3) A student located on the level of analysis is able to assign presented extracts from a study's problem definition to its typical elements
(4) Application	
	(4) A student who is able to set up a research design which is based on a presented research objective or question of a concrete study attained the level of knowledge utilization
(II) *Complexity concerning the number of content-related/ curricular elements (=solution-relevant variables) (Adams and Wu* 2002)	(1) A task that only demands to describe what the term 'nominal scale level' means contains a low complexity
Three response categories:	(2) A task which requires a decision if a correlation or a regression analysis is fitting better in order to answer a presented research question shows a moderate complexity
(1) Only one isolated content-related element has to be processed	
(2) Two content-related elements have to be processed	
(3) At least three content-related elements have to be processed	(3) A task prompting to set up a context-related experimental design which has to include all relevant components—such as pretest, treatment, posttest, experimental group, and control group – and which requires considering the concept of randomization involves a high complexity

Table 6 continued

Criterions for the tasks' difficulty including their operationalization	Examples
(III) *Degree of familiarity* *(~curricular weighting of the task contents; degree of routine with regard to the respective task context)* *(e.g., Blum et al. 2003)* Three response categories: (1) High degree of familiarity = many learning opportunities (2) Moderate degree of familiarity = moderate number of learning opportunities (3) Low degree of familiarity = few learning opportunities *Details concerning the operationalization:* Nine content areas, which depict all contents instructed during the course on empirical research methods were defined →Regarding each item, an index consisting of three criteria was calculated in order to assess the amount of learning opportunities: (a) How many lecture slides are dealing with the relevant content area? (referring to the first element of the course—the lecture) (b) How extensive was the respective content area treated within the instruction? (referring to the first and the second element of the course—the lecture and the tutorial moderated by a lecturer) (i) = low instructional extent (ii) = high instructional extent (c) Did the test persons have the opportunity to participate proactively in a case-related hands-on-application concerning the respective content area? (referring to the third element of the course—the project work in small groups supported by advanced students) (i) = no hands-on-application was performed (ii) = hands-on-application was performed	(1) A large proportion of the course's instruction was spent on performing several in-depth and hence intensive exercises to handle situations belonging to the curricular area of evaluating the adequateness of methods for collecting data (2) A moderate proportion of the course's instruction was spent on in-depth exercises with regard to interpreting statistical outputs for explorative factor analyses (3) The curricular area of regression analysis was treated very superficially during the course (no hands-on applications and repetitions of the respective contents/methods were provided)

Assumption ad (I): the solving probability decreases with an increase in the kind of cognitive process which is necessary to master the respective task; assumption ad (II): less test persons are able to solve an item addressing many different content-related/curricular elements that have to be linked than an item designed to capture only one or few elements of the complex structure of research methods; assumption ad (III): the solving probability is lower for items which are directed at a quite unfamiliar situation compared to items that display familiar situations

Table 7 Item fit statistics

Items	Difficulty parameter	Standard error	wMNSQ	CI	t value
1	−.034	.136	1.00	(.81, 1.19)	.1
2	−1.215	.196	1.02	(.81, 1.19)	.2
3	.287	.171	.98	(.92, 1.08)	−.4
4	.748	.179	1.02	(.88, 1.12)	.4
5	.512	.174	1.02	(.90, 1.10)	.5
6	−1.029	.209	.98	(.80, 1.20)	−.2
7	−1.809	.234	.95	(.72, 1.28)	−.3
8	−.335	.181	1.03	(.89, 1.11)	.5
9	−1.442	.213	.89	(.78, 1.22)	−1.0
10	−.992	.194	.96	(.84, 1.16)	−.5
11	−1.484	.215	.96	(.77, 1.23)	−.3
12	.663	.186	1.06	(.87, 1.13)	.9
13	−1.244	.204	.90	(.81, 1.19)	−1.0
14	−1.401	.211	.96	(.79, 1.21)	−.3
15	.316	.127	1.09	(.83, 1.17)	1.0
16	.104	.136	1.13	(.82, 1.18)	1.5
17	−1.506	.367	.99	(.79, 1.21)	−.0
18	−1.063	.155	1.01	(.78, 1.22)	.2
19	−.336	.142	.97	(.80, 1.20)	−.3
20	−.274	.120	.97	(.83, 1.17)	−.4
21	−.420	.124	1.01	(.82, 1.18)	.1
22	−.517	.126	.93	(.82, 1.18)	−.7
23	.035	.115	1.15	(.83, 1.17)	1.7
24	.237	.133	1.07	(.83, 1.17)	.8
25	−.722	.116	1.04	(.79, 1.21)	.4
26	.413	.142	.95	(.81, 1.19)	−.4

wMNSQ weighted mean square, *CI* confidence interval

Received: 11 August 2016 Accepted: 8 March 2017
Published online: 27 March 2017

References

Abdi H (2007) Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks, pp 103–107

Adams RJ, Wu ML (2002) PISA 2000 technical report. OECD, Paris

Adams RJ, Wilson M, Wang W-C (1997) The multidimensional random coefficients multinomial logit model. Appl Psychol Meas 21(1):1–23

Andersen EB (1973) A goodness of fit test for the Rasch model. Psychometrika 38(1):123–140

Baumert J, Kunter M (2006) Stichwort: Professionelle kompetenz von lehrkräften. Z Erziehwiss 9(4):469–520

Bley S (2017) Developing and validating a technology-based diagnostic assessment usingthe evidence-centered game design approach: an example of intrapreneurship competence. Empirical Res Voc Ed Train 9(6):1–32. doi:10.1186/s40461-017-0049-0

Bley S, Wiethe-Körprich M, Weber S (2015) Formen kognitiver Belastung bei der Bewältigung technologiebasierter authentischer Testaufgaben – eine Validierungsstudie zur Abbildung vonberuflicher-Kompetenz. Zeitschrift für Berufs- und Wirtschaftspädagogik 111(2):268–294

Blömeke S, Felbrich A, Müller C, Kaiser G, Lehmann R (2008) Effectiveness of teacher education: state of research, measurement issues and consequences for future studies. Int J Math Educ 40:719–734

Blömeke S, Suhl U, Kaiser G (2011) Teacher education effectiveness: quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. J Teach Educ 62(2):154–171

Blömeke S, Gustafsson J-E, Shavelson RJ (2015) Beyond dichotomies: competence viewed as a continuum. Z Psychol 223(1):3–13

Blum W, Neubrand M, Ehmke T, Senkbeil M, Jordan KA, Ulfig F, Carstensen CH (2003) Mathematische Kompetenz. In: Prenzel M, Baumert J, Blum W, Lehmann R, Leutner D, Neubrand M, Pekrun R, Rolff H-G, Rost J, Schiefele U (eds) PISA 2003. Waxmann, Münster, pp 47–92
Bouley F, Wuttke E, Schnick-Vollmer K, Schmitz B, Berger S, Fritsch S, Seifried J (2015) Professional competence of prospective teachers in business and economics education—evaluation of a competence model using structural equation modelling. Peabody J Educ 90(4):491–502
Brown NJS, Furtak EM, Timms M, Nagashima SO, Wilson M (2010) The evidence-based reasoning framework: assessing scientific reasoning. Educ Assess 15(3/4):123–141
Bühner M (2011) Einführung in die test- und fragebogenkonstruktion. Pearson Studium, München
Bühner M, Ziegler M (2009) Statistik für Psychologen und Sozialwissenschaftler. Pearson, Hallbergmoos, pp 681–682
Darling-Hammond L, Bransford J (2005) Preparing teachers for a changing world: what teachers should learn and be able to do. Jossey-Bass, San Francisco
de Ayala RJ (2009) The theory and practice of item response theory. The Guilford Press, New York
de Ayala RJ, Plake BS, Impara JC (2001) The impact of omitted responses on the accuracy of ability estimation in item response theory. J Educ Meas 38(3):213–234
Dunn D, Smith RA, Beins B (2007) Best practices for teaching statistics and research methods in the behavioral sciences. L. Erlbaum Associates, Mahwah
Egeln J, Gottschalk S, Rammer C, Spielkamp A (2002) Hohe Zahl an Spinoff-Gründungen aus der Wissenschaft. ZEW Gründungsreport 2(2):3–4
Embretson SE (2002) Generating abstract reasoning items with cognitive theory. In: Irvine S, Kyllonen P (eds) Generating items for cognitive tests: theory and practice. Lawrence Erlbaum Associates, Publishers, Mahwah, pp 35–60
Fichten W (2010) Forschendes Lernen in der Lehrerbildung. In: Eberhardt U (ed) Neue impulse in der Hochschuldidaktik: Sprach- und Literaturwissenschaften. VS-Verlag für Sozialwissenschaften, Wiesbaden, pp 127–182
Fleischer J, Koeppen K, Kenk M, Klieme E, Leutner D (2013) Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. Z Erziehwiss 16(Sonderheft 18):5–22
Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33(3):613–619
Gesellschaft für Empirische Bildungsforschung (2012) *Satzung der Gesellschaft für Empirische Bildungsforschung*. http://www.gebf-ev.de/über-die-gebf/satzung/. Accessed 1 Oct 2015
Glas CAW, Verhelst ND (1995) Testing the Rasch model. In: Fischer GH, Molenaar IW (eds) Rasch models: foundations, recent developments, and applications. Springer, New York, pp 69–95
Groß Ophoff J, Schladitz S, Lohrmann K, Wirtz MA (2014) Evidenzorientierung in bildungswissenschaftlichen Studiengängen. In: Drossel K, Strietholt R, Bos W (eds) Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen. Waxmann, Münster, pp 251–275
Groß Ophoff J, Schladitz S, Leuders J, Leuders T, Wirtz MA (2015) Assessing the development of educational research literacy: the effect of courses on research methods in studies of educational science. Peabody J Educ 90(4):560–573
Groth RE (2007) Toward a conceptualization of statistical knowledge for teaching. J Res Math Educ 38(5):427–437
Hartig J (2007) Skalierung und definition von Kompetenzniveaus. In: Beck B, Klieme E (eds) Sprachliche Kompetenzen Konzepte und Messung DESI-Studie. Beltz, Weinheim, pp 83–99
Hartig J, Frey A (2013) Sind Modelle der Item-Response-Theorie (IRT) das "Mittel der Wahl" für die Modellierung von Kompetenzen? Z Erziehwiss 16(Sonderheft 18):47–51
Hartig J, Klieme E (2006) Kompetenz und Kompetenzdiagnostik. In: Schweizer K (ed) Leistung und Leistungsdiagnostik. Springer, Berlin, pp 127–143
Holland JL (1959) A theory of vocational choice. J Couns Psychol 6:35–45
Interstate Teacher Assessment and Support Consortium (2011) Model core teaching standards: a resource for state dialogue. Council of Chief State School Officers, Washington
Jahed J, Bengel J, Baumeister H (2012) Transfer von Forschungsergebnissen in die medizinische Praxis. Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes Germany) 74(11):754–761
Janesick VJ (2006) Authentic assessment primer. Peter Lang, New York
Kauertz A, Fischer HE (2008) Schwierigkeitserzeugende Merkmale physikalischer Testaufgaben. In: Höttecke D (ed) Kompetenzen, Kompetenzmodelle, Kompetenzentwicklung: Jahrestagung in Essen 2007. Lit, Berlin, pp 218–220
Klahr D, Dunbar K (1988) Dual search space during scientific reasoning. Cogn Sci 12:1–48
KMK (2004). Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004). http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf. Accessed 21 Aug 2014
Koch S (2006) Persönliche Verantwortung für den Studienerfolg. Dimensionen und Korrelate. Z Pädag Psychol 20(4):243–250
Koller I, Alexandrowicz R, Hatzinger R (2012) Das Rasch-Modell in der Praxis: Eine Einführung mit eRm. Facultas, Wien
Kuckartz U (2014) Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung (2. Aufl.). Beltz/Juventa, Weinheim
Mair P, Hatzinger R (2007) Extended Rasch modeling: the eRm package for the application of IRT models in R. J Stat Softw 20(9):1–20
Marzano RJ, Kendall JS (2007) The new taxonomy of educational objectives. Corwin Press, Thousand Oaks
Marzano RJ, Kendall JS (2008) Designing and assessing educational objectives: Applying the new taxonomy. Corwin Press, Thousand Oaks
Masters GN (1982) A rasch model for partial credit scoring. Psychometrika 47(2):149–174
Messick S (1989) Validity. In: Linn RL (ed) Educational measurement. American Council on Education, New York, pp 13–103
Messick S (1995) Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 50(9):741–749

Mislevy RJ, Haertel GD (2006) Implications of evidence-centered design for educational testing. Educ Meas Issues Pract 25(4):6–20
Mislevy RJ, Wu P-K (1996) Missing responses and IRT ability estimation: omits, choice, time limits, and adaptive testing. Princeton, NJ
National Research Council (2012) Education for life and work: developing transferable knowledge and skills in the 21st century. The National Academies Press, Washington
OECD (2014) PISA 2012 technical report. OECD, Paris
Onwuegbuzie AJ (2001) Statistics anxiety and the role of self-perceptions. J Educ Res 94(6):323–330
Oser FK (1997) Standards in der Lehrerbildung—Teil 1 Berufliche Kompetenzen, die hohen Qualitätsmerkmalen entsprechen. Beitr zur Lehrerbildung 15(1):26–37
Pellegrino JW (2010) The design of an assessment system for the race to the top: a learning sciences perspective on issues of growth and measurement. Educational Testing Service, Princeton
Pellegrino JW, Chudowsky N, Glaser R (2001) Knowing what students know: the science and design of educational assessment. Natl Acad Press, Washington
Pohl S, Gräfe L, Rose N (2014) Dealing with omitted and not-reached items in competence tests: evaluating approaches accounting for missing responses in item response theory models. Educ Psychol Meas 74(3):423–452
Ponocny I (2001) Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika 66(3):437–459
Rasch G (1961) On general laws and the meaning of measurement in Psychology. University of California Press, Berkeley
Renkl A (1994) Wer hat Angst vorm Methodenkurs? Eine empirische Studie zum Streßerleben von Pädagogikstudenten in der Methodenausbildung. In: Olechowski R, Rollett B (eds) Theorie und Praxis: Aspekte empirisch-pädagogischer Forschung, quantitative und qualitative Methoden. Peter Lang, Frankfurt am Main, pp 178–183
Rost DH (2007) Interpretation und Bewertung pädagogisch-psychologischer Studien. Eine Einführung. Beltz, Weinheim
Schweizer K, Steinwascher M, Moosbrugger H, Reiss S (2011) The structure of research methodology competency in higher education and the role of teaching terms and course temporal distance. Learn Instr 21(1):68–76
Shavelson R (2008) "The Collegiate Learning Assessment," Ford policy forum 2008: forum for the future of higher education, 20. http://net.educause.edu/forum/fp08.asp
Shavelson RJ (2010) On the measurement of competency. Empir Rese Vocat Educ Train 1:43–65
Shavelson RJ (2012) Assessing business-planning competence using the Collegiate Learning Assessment as a prototype. Empir Res Vocat Educ Train 4(1):77–90
Shrout PE, Fleiss JL (1979) Interclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420–428
Shulman LS (1987) Knowledge and teaching: foundations of the new reform. Harv Educ Rev 57:1–22
Slavin RE (2007) Educational research in an age of accountability. Pearson, Boston
Slavin RE (2008) Perspectives on evidence-based research in education—what works? Issues in synthesizing educational program evaluations. Educ Res 37(1):5–14
Stark R, Mandl H (2001) Entwicklung, Implementation und Evaluation eines beispielbasierten Instruktionsansatzes zur Förderung von Handlungskompetenz im Bereich empirischer Forschungsmethoden. Forschungsbericht Nr. 141. München
Tremblay K, Lalancette D, Roseveare D (2012) Assessment of higher education learning outcomes, AHELO, feasibility study report, volume 1—design and implementation. OECD. http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf. Accessed 23 Jun 2015
van Merriënboer JJG, Kirschner PA (2013) Ten steps to complex learning: a systematic approach to four-component instructional design. Routledge, London
Wagner C, Maree D (2007) Teaching research methodology: implications for psychology on the road ahead. S Afr J Psychol 37:121–134
Wahl D (1982) Handlungsvalidierung. In: Huber GL, Mandl H (eds) Verbale Daten: Eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung. Beltz, Weinheim, pp 259–274
Weber S, Achtenhagen F (2009) Forschungs- und evidenzbasierte Lehrerbildung. In: Zlatkin-Troitschanskaia O, Beck K, Sembill D, Nickolaus R, Mulder R (eds) Lehrprofessionalität: Bedingungen, Genese, Wirkungen und ihre Messung. Beltz, Weinheim, pp 477–487
Weber S, Trost S, Wiethe-Körprich M, Weiß C, Achtenhagen F (2014) Intrapreneur: an entrepreneur within a company—an approach on modeling and measuring intrapreneurship competence. In: Weber S, Oser FK, Achtenhagen F, Fretschner M, Trost S (eds) Becoming an entrepreneur: professional and VET learning. Sense Publishers, Amsterdam, pp 279–302
Weinert FE (2001) Concept of competence: a conceptual clarification. In: Rychen DS, Sagalnik LH (eds) Defining and selecting key competencies. Hogrefe & Huber Publishers, Seattle, pp 45–65
Wiethe-Körprich M, Trost S (2013) To conduct empirical research as a student – measurement of research methodology competency. Presentation on the 15th Biennial conference EARLI, Munich, 27 August–1 September 2013
Wilson MR (2005) Constructing measures: an item response modeling approach. Lawrence Erlbaum Associates Publishers, Mahwah
Winther E, Achtenhagen F (2009) Skalen und Stufen kaufmännischer Kompetenz. Zeitschrift für Berufs- und Wirtschaftspädagogik 105(4):521–556
Wise SL, DeMars C (2005) Low examinee effort in low-stakes assessment: problems and potential solutions. Educ Assess 10(1):1–17
Wise SL, DeMars C (2006) An application of item response time: the effort-moderated IRT model. J Educ Meas 43(1):19–38
Wolf R, Zahner D, Benjamin R (2015) Methodological challenges in international comparative post-secondary assessment programs: lessons learned and the road ahead. Stud High Educ 40(3):471–481
Wu ML, Adams RJ, Wilson MR (2007) ACER ConQuest. ACER Press, Camberwell
Wuttke E (2001) Wie relevant ist die Forschung für die Praxis? Überlegungen zu Forschungsmethoden und der Rezeption von Forschungsergebnissen. Zeitschrift für Berufs- und Wirtschaftspädagogik 97(Beiheft 16):30–41
Zurstrassen B (2009) Kompetenzorientierte Lehrerbildung in den sozialwissenschaftlichen Unterrichtsfächern: blühende Landschaften in der sozialwissenschaftlichen Lehrerbildung von morgen? J Soc Sci 8(2):32–45