Rohr-Mentele and Forster-Heinzer
Empirical Res Voc Ed Train (2021) 13:18
https://doi.org/10.1186/s40461-021-00122-2

Empirical Research in Vocational Education and Training

RESEARCH Open Access

Practical validation framework for competence measurement in VET: a validation study of an instrument for measuring basic commercial knowledge and skills in Switzerland

Silja Rohr-Mentele[1*] and Sarah Forster-Heinzer[2]

*Correspondence:
silja.rohr-mentele@uzh.ch
[1] Institute of Education, University of Zurich, Zurich, Switzerland
Full list of author information is available at the end of the article

Abstract

Competence development and measurement are of great interest to vocational education and training (VET). Although there are many instruments available for measuring competence in diverse settings, in many cases, the completed steps of validation are neither documented nor made transparent in a comprehensible manner. Understanding what an instrument actually measures is extremely important, inter alia, for evaluating test results, for conducting replication studies and for enforcing adaptation intentions. Therefore, more thorough and qualitative validation studies are required. This paper presents an approach to facilitate validation studies using the example of the *simuLINCA* test. The approach to validation applied in this study was developed in the field of medicine; nevertheless, it provides a promising means of assessing the validity of (computer-based) instruments in VET. We present the approach in detail along a newly developed computer-based simulation (*simuLINCA*) that measures basic commercial knowledge and skills of apprentices in Switzerland. The strength of the presented approach is that it provides practical guidelines that help perform the measurement process and support an increase in transparency. Still, it is flexible enough to allow different concepts to test development and validity. The approach applied proved to be practicable for VET and the measurement of occupational competence. After extending and slightly modifying the approach, a practical validation framework, including the description of each step and questions to support the application of it, is available for the VET context. The computer-based test instrument, *simuLINCA*, provides insights into how a computer-based test for measuring competence in various occupational fields can be developed and validated. *SimuLINCA* showed satisfying evidence for a valid measurement instrument. It could, however, be further developed, revised and extended.

Keywords: Competence measurement, Validation, Validation framework, Computer-based simulation, Commercial apprenticeship

Introduction

The aim of (vocational) education is to equip students with the knowledge, skills and attitudes they need in order to be competent in their workplace and as members of society today and in the future. Valid competence measurement helps determine whether this goal is achieved. The last decade has seen a change in the way competence is measured in vocational education and training (VET), in particular with the implementation of computer-based simulations. The advantage of such tests is the possibility of developing tasks that are similar to workplace situations (reality) and practicable (Jude and Wirth 2007; Seeber et al. 2010; Winther 2010). However, Rüschoff (2019) points out deficits in the transparent validation of such tests. In a review of the methods of competence measurement in initial VET in Germany in the years 2001–2017, the author illustrates that around 11% of the conducted studies did not (or not clearly enough) mention the validation procedure. While content validity was analysed in all the studies, construct and criterion validity were not considered in 78% and 88% of cases, respectively. Rüschoff concluded that construct and criterion validity are underrepresented and that it is worth striving for a more thorough validation, including predictive validity. In order to achieve progress in the validation of measurement instruments in VET, it is important to follow guidelines during the validation process and to present the modelling of constructs, results and resources needed as transparently as possible. Such guidelines, for example, suggested by the American Educational Research Association (AERA) et al. (2014), are available, but Rüschoff's results indicate that they have to be more practical and/or easier to access.

The present paper builds on this demand by presenting and discussing a practical approach to validity, which is of particular interest for the development and validation of assessments in education and further training (VET/company). In the 'Introduction' section, we outline the concept of validity and approaches to validation. In the 'Method' section, we describe the practical approach to validation applied in this paper on a theoretical basis, and in the 'Results and discussion' section, we explicitly apply the individual steps of the approach using the example of competence measurement in commercial VET in Switzerland. Therefore, the present paper is not structured like a 'classical empirical' paper with research questions and underlying hypotheses; instead it takes a conceptual perspective to pursue two concrete goals: (a) we want to examine whether the approach is suitable for the VET field and (b) we want to give a concrete illustration of the process of development and validation of the computer-based instrument, *simuLINCA* for researchers interested either in adapting or in developing a (computer-based) competence measurement instrument.

Validity and approaches to validation

According to McClelland (1973), a test should not be trusted, nor is its use justified, unless there is evidence for its validity. But what validity means is not as clear as it seems. Commonly known is the division of validity into the 'types' criterion, content and construct validity (Cook et al. 2015; Rüschoff 2019). However, some (educational) measurement experts have a different understanding of validity, namely of validity being a unified concept. Messick (1975, 1989) introduced this 'unitary concept of validity', which

is arranged around the framework of a broad version of construct validity. Construct validity is to be understood as superordinate and includes the other validity types (Anastasi 1986). Although there is a large consensus favouring a unified concept of validity over a three-part one, the debate over what the *concept of validity* encompasses is ongoing (Hammond and Moss 2016). There is agreement that validity is about the 'evaluation of interpretive claims' (Kane 2016, p. 198) that have to be justified somehow. Validity should be addressed by providing evidence from multiple perspectives (Kane 2016; Shepard 2016). However, there is disagreement, for example, about whether the 'test use' and/or 'test consequences' have to be included in the concept of validity (Newton and Baird 2016). A compromise (Geisinger 2016; Shepard 2016) can be found in the *Standards of Educational and Psychological Testing* (AERA et al. 2014):

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. […] The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (p. 11)*[1]

Despite a heated debate and a preoccupation with the concept of validity, Brennan (2006, p. 8) noted that the *practice of validation* is frequently 'impoverished'. Gafni (2016) also emphasized that the practice of validation is often inadequate and occasionally not done at all. Cook et al. (2013) came to a similar conclusion. They conducted a review study that included the validation process of more than 400 simulation-based assessments in the field of medicine and concluded that there is a need for qualitatively better validation studies and thus for clearer and simpler instructions on the validation process.

A crucial step in the validation process is the *selection and collection of evidence*. There are various methods for this, inter alia,[2] (A) Messick (1989), for instance, considered the aspects content, substantive, structural, external, generalizability and consequential as relevant sources for (construct) validity. Nevertheless, Messick's approach was criticized because it would provide incomplete guidance in terms of prioritization or selection of evidence, which would consequently lead to an open-ended validation process (Kane 2013; Carney et al. 2019). (B) This objection was addressed by Kane (1992, 2001, 2013) with the introduction of an interpretation-use argument (IUA) prior to the actual evidence collection; the theory and/or construct of a test should be made explicit and underlying assumptions concerning the interpretation and use of test scores should be formulated. Depending on the aim of the measurement, Kane suggested referring the assumptions to the 'inference-categories' scoring, generalization, extrapolation and implications/decisions. (C) Similarly, AERA et al. (2014) proposed starting the validation process with an IUA. In contrast to Kane, AERA et al. (1999,2014) adopted Messick's validity aspects and 'renamed' them (test content, internal structure, relations to other

[1] This understanding of validity is also followed in this article.

[2] Carney et al. (2019, p. 12) identified three perspectives regarding argument-based approaches to validation: (1) 'validation via principled design', (2) 'common identified categories' and (3) 'chain of assumption/inferences'. There are also examples provided for each perspective. The AERA approach was assigned to the second perspective and Kane's approach to the third.

variables, response processes and consequences). Carney et al. (2019) recognized the difference between the two approaches insofar as the five sources of evidence are quite prescriptive and the focus is on test development, whereas Kane's (2013) 'inference-categories' paid more attention to practical implementation in a more open or contingent form.[3] (D) Cook and Hatala (2016) devised an eight-step approach to validation that could operate with the 'classical' understanding of validity (content, criterion, construct), with AERA's et al. and Kane's validity approach. Thus, Cook and Hatala did not present a new version of validation. What is new or appealing about their approach is that they transformed the written text into an easy-to-use framework.[4] This was achieved with the aid of consecutive steps and clear as well as informative headings. In our view, the approach of Cook and Hatala is particularly interesting for VET since it has been developed for the medical field on the one hand, which corresponds to the world of labour and not to that of general education. On the other hand, the approach to validation was devised for simulation studies. As we have applied this approach to the development and validation of *simuLINCA*, we will introduce the eight steps proposed in more detail.

Method: the practical approach to validation by Cook and Hatala

Strictly speaking, Cook and Hatala's approach is not just about validation but rather about an assessment model including the validation process. Usually, modelling, development and practical issues are not part of the validation procedure. However, Shepard (2016), amongst others, stated that the test format of a measurement instrument is determined by a clear idea of test utilization. Thus, there is no real separation of development and validation since they are intertwined. The eight proposed steps of the validation framework are:

1. Define the construct and proposed interpretation
2. Make explicit the intended decision(s)
3. Define the interpretation-use argument, and prioritize needed validity evidence
4. Identify candidate instruments and/or create/adapt a new instrument
5. Appraise existing evidence and collect new evidence as needed
6. Keep track of practical issues including cost
7. Formulate/synthesize the validity argument in relation to the interpretation-use argument
8. Make a judgment: does the evidence support the intended use? (Cook and Hatala 2016, p. 6)

These eight steps are interdependent and build on each other. Describing the construct to be measured as precisely and theoretically soundly as possible and specifying the interpretations, uses and decisions to be taken on the basis of test results (Steps 1 and 2) are the foundation to formulate the IUA (Step 3). In the IUA, the often implicit existing assumptions are formulated explicitly (e.g. test items are authentic and the test can

[3] For a detailed insight and comparison of the two approaches, the article by Carney et al. (2019) is recommended.

[4] In the literature, the terms 'approach' and 'framework' are sometimes used synonymously. We understand 'approach' as open and general access, whereas we understand 'framework' as a more restrictive, prescriptive and precisely defined procedure, like the steps proposed by Cook and Hatala. A 'framework' would therefore be part of an approach.

be objectively administered) and it is determined what evidence is essential for which assumption. In the context of testing, countless assumptions can be made, but not all are equally relevant. Therefore, as many assumptions as possible should first be identified and then the most problematic and challenging should be prioritised. It is not known at this stage whether there is already evidence of validity, but at least it is clear what evidence is needed to 'prove' the assumptions. In Step 4, the search begins for a suitable measurement instrument that can fulfil the assumptions formulated in Step 3. It is only in Step 5 that concrete evidence is sought, which is generally known as validation. In this step, the IUA formulated in Step 3 is now (empirically) 'tested'. Planning this step thoroughly is extremely important with regard to other constructs that are assessed along with the competence measurement. Often needed validity evidence is not considered in advance and therefore not collected, which makes it impossible to test assumed relations with other constructs as most data collection has already been done and not repeatable. Step 6 emphasises that valid testing also depends on contextual and financial factors, which should be documented to show other research groups what resources are needed for the whole validation process and for the implementation of the test. Being aware of challenges in advance and providing alternatives (e.g. offline simulation), ensures that testing conditions are similar and test results comparable. In addition, knowing the cost helps focus and possibly limit the instrument and validation to the most important aspects. In Step 7, the evidence for validity collected in Step 5 is compared and evaluated with the explicitly formulated assumptions in the IUA (Step 3). Possible shortcomings with regard to the proof of validity are explicitly mentioned. Finally, on the basis of Steps 1–7, a final conclusion (Step 8) is drawn as to the extent to which the present test instrument can withstand the intended use. Furthermore, plans to close remaining gaps concerning validity evidence are presented.

Changes to the framework

Even though this framework is clearly structured and includes the most important facets of the validation process, there is one step missing. Hence, an indispensable condition for the implementation of suitable measurement instruments (especially in VET) is that the characteristics of the context in which the measurement should be applied is taken into account (Holtsch et al. 2016). It is, for example, important to be aware of target characteristics in advance (e.g. age, language skills). Therefore, we added Step 0, 'Define the Measurement Context and Target Group' to the framework. Furthermore, we changed the wording of Steps 1 and 2 to (1) 'Define the Construct' and (2) 'Make Explicit the Intended Interpretation, Use and Decision(s)'. It does not become entirely clear how the construct and interpretation are distinguishable in the example provided by Cook and Hatala (2016, p. 6). To correspond to the validity definition provided by the AERA et al. (2014), we included 'interpretation' and 'use' in Step 2. We have also changed the order of the steps, as in our opinion, the formulation of the validity argument (previously Step 7, now Step 6) should take place immediately after the collection of evidence (Step 5).

Results and discussion

Application of the framework in the validation study *simuLINCA*

Against the background of a lack of instruments for measuring competence in Swiss commercial VET and of the need for a careful validation process, we follow and illustrate in detail each of the eight steps as proposed by Cook and Hatala (2016) as well as the additional Step 0 to develop and validate *simuLINCA*.

Step 0: Define the measurement context and target group

In Switzerland, dual (commercial) VET plays a special role in educating young people to become competent members of the workforce and society. Two thirds of young adolescents enter dual VET after compulsory schooling, and one out of five school-leavers serves a commercial apprenticeship. The apprenticeship can be completed at one of three levels, all of which lead to a Federal VET Diploma (EFZ): B-profile (basic education), E-profile (advanced education) and M-profile (advanced education with a federal vocational baccalaureate). In this study, commercial apprentices of the E- and M-profile at the end of their training make up our target group. Their commercial training is organized by the three learning venues—vocational school, company and industry courses. In order to achieve the training goals, learning objectives are formulated at a Bloom taxonomy level of 1–6 (Bloom et al. 1956) for all learning venues. The *school curriculum* includes the subjects economics and society (E&S), information, communication and administration (ICA) and languages. Despite the fact that apprentices learn their profession in one of 21 commercial branches (SKKAB [Swiss Conference of Commercial Training and Examination Branches] 2020)[5]and consequently receive branch-specific training in their company and industry courses, all of them, irrespective of their branch, attend commercial vocational school together. This is a unique situation in the context of Swiss commercial VET and known as the 'all-branch concept' (Rosenheck 2010). The idea behind this concept is that after the completion of their apprenticeship, trained commercial apprentices should be able to work in all commercial branches and, hence, broadly in the workforce and society. Having been devised by practitioners, the (branch-specific) *company curriculum* is characterized by a practical orientation and is thus relevant with respect to on-the-job training as well (Zbinden-Bühler and Volz 2007). The company curriculum consists of eight competence areas: (1) handling material, data or services, (2) advising customers, (3) carrying out orders, (4) performing marketing and public-relations tasks, (5) carrying out human-resources administration tasks, (6) executing financial processes, (7) performing administrative and organizational activities and (8) using knowledge gained in one's own branch and company (BBT [Federal Office for Professional Education and Technology] 2017, p. 2).[6]

Step 1: Define the construct

The construct to be measured by *simuLINCA* is *basic commercial knowledge and skills* (synonymous with basic commercial competence). Referring to a competence definition by Shavelson (2010, p. 44), we understand *competence* as the capability to perform

[5] All of the 21 branches are listed here: https://www.skkab.ch/berufsinformationen/branchen/.

[6] The branch-specific curricula came into effect on 1 January 2012 and were last updated on 1 January 2017.

successfully in a specific context, which is closely related to the use of knowledge and skills in real life as in our example of commercial apprentices of all branches and training profiles E and M. Within this two-dimensional understanding of competence, knowledge is seen as the necessary basis and understood to be as important as the skills for solving a problem or performing successfully (Preckel and Holling 2006). The concept of skills refers to the practical side of the activity, the ability to do something well. While knowledge and skills are separable from an analytical point of view, they are closely intertwined in practice. In the context of VET, the school is, generally speaking, responsible for the 'transmission' of knowledge whereas the company provides the setting for the acquisition of pertinent skills (Sloane 2006).

Research activities in VET regarding the structure of professional competence show empirically that professional knowledge can be distinguished from action-related knowledge/skills (Nickolaus 2018). In the commercial domain, for example, Achtenhagen and Winther (2009) and Winther (2010) were able to empirically subdivide competence into comprehension-based and action-based dimensions. Furthermore, it has been shown that professional knowledge can often be divided into sub-dimensions that are commonly used to structure knowledge (Rosendahl and Straka 2011; Nickolaus and Seeber 2013; Nickolaus 2018). Sub-dimensions, as for example, case situations, can also be identified for action-related knowledge/skills (Achtenhagen and Winther 2009; Winther 2010; Rosendahl and Straka 2011).

Differing from previous research activities in the commercial field, the construct of the present project is to be understood independently of a specific commercial occupation (e.g. industrial clerk) and is comprised of basic commercial knowledge and skills that apprentices of all commercial branches should possess. In line with previous research (Nickolaus 2018), our construct (or competence structure model) can be described as having two dimensions: (1) *Basic commercial knowledge* covers dealing with commercial requirements based on knowledge acquired at vocational school, in particular in the lectures on business and administration (BA) (e.g. knowledge about tax rates). It reflects the 'school curriculum'. It is basic because it relates to a non-academic level and is relevant to apprentices of all branches and for both profiles. (2) *Basic commercial skills* are needed to fulfil non-branch-specific tasks in a commercial workplace (e.g. writing an e-mail to a customer). These skills are mainly acquired in the company and reflect the 'companies' curricula'.

Step 2: Make explicit the intended interpretation, use and decision(s)

The main purpose of the *simuLINCA* instrument is to measure basic commercial knowledge and skills in the longitudinal project 'Learning and Instruction for Commercial Apprentices' (LINCA) (Holtsch and Eberle 2018). A by-product is that the test results provide apprentices as well as their VET teachers and trainers with an objective determination of the status quo by showing them whether the apprentices are able to deal successfully with representative tasks that pertain to all commercial branches and whether they can answer questions concerning BA that are typical on the final exams. Such information could encourage teachers and/or trainers to take measures in order to prepare effectively for the transition from VET to the labour market. However, the aim of the test is not to replace the qualification procedures or to influence selection processes; the

Table. 1 Assumptions made for test interpretation and use (to be tested in the subsequent steps)

Sources of evidence				
A) Test Content	**B) Response Processes**	**C) Internal Structure**	**D) Relations to other Variables**	**E) Consequences**
Assumptions				
Tasks are authentic and relevant to all branches and correspond to real-life workplace situations	Commercial apprentices understand the skill-based tasks and the knowledge-based items	The two-dimensionality of the construct can be confirmed empirically	Apprentices with higher marks in E&S and ICA perform better than apprentices with lower marks	An instrument for the measurement of basic knowledge and skills in commercial VET is available
Knowledge-based items are relevant for the final exams and representative of BA	Commercial apprentices know where to find material and where to look for information	The test measures basic commercial knowledge and skills reliably	Apprentices with higher scores on the intelligence test perform better than apprentices with lower scores	Feedback is appropriate for a determination of the status quo
			M-profile apprentices perform better than E-profile apprentices	The instrument is not used for qualification or selection processes

aim and use of *simuLINCA* should be clearly communicated to other stakeholders (e.g. school deans, government) in order to avoid misuse of the instrument.

Step 3: Define the interpretation-use argument, and prioritize needed validity evidence

Table 1 presents the most important assumptions about the construct and test use (Steps 1 and 2). We structured our assumptions with the aid of the five sources of evidence suggested by the AERA et al. (2014). For example, as we noted in Step 1 that the test should contain tasks that can be solved independently of the branch. One of the assumptions to be tested later was therefore 'tasks are authentic and *relevant to all branches*' (Table 1, source: test content). Furthermore, we assumed with regard to the internal test structure that our construct could be divided into the two aforementioned dimensions of basic commercial knowledge and basic commercial skills (Step 1). Consequently, we had to check our proposed construct along the data empirically. Additionally, we expected apprentices with higher marks in E&S to perform better than apprentices with lower marks (source: relations to other variables). Therefore, we needed to ask for the marks in E&S in order to be able to test this assumption.

Step 4: Identify candidate instruments and/or create/adapt a new instrument

In a next step, the existing testing instruments will be reviewed, focusing on whether they are consistent with the underlying assumptions (Table 1). In this way, it is possible to decide whether these instruments are suitable or whether a new instrument needs to be developed.

Step 4.1 Candidate instruments

At the beginning of our research project in 2012, only the instrument ALUSIM (Achtenhagen and Winther 2009; Winther 2010) measured a similar construct. However, due

to substantial country-related differences in the conception of commercial VET and the fact that it focuses primary on industrial apprentices (violating the all-branch assumption), it has proven to be impossible to adapt ALUSIM satisfactorily to the Swiss context (Holtsch et al. 2016). In consequence, the *simuLINCA* simulation was devised (Mentele et al. 2014).

Step 4.2 Test development of the new instrument

In order to represent what apprentices of all branches learn at the two learning venues, the curricula of both vocational school and company were used to develop the tasks and items. The development process of *simuLINCA* involved (a) *the selection of competence areas*, (b) *the design of the simulation environment and the formulation of skill-based tasks*, (c) *the formulation of knowledge-based items* and (d) *the digital construction of the instrument*:

a. Selection of competence areas: It was to decide which competence areas (see Step 0) were to be covered by the test in order to meet our assumptions of respecting the 'all-branch concept'. To this end, we conducted both a quantitative and a qualitative analysis of the similarities and differences concerning the educational goals set for commercial apprentices in all branch-specific curricula. The quantitative curriculum analysis focused on the number of performance targets per competence area. It showed that four of the eight competence areas were included in all curricula. Subsequently, the wording of the targets in qualitative respects was analysed, and it was revealed that overall, most of the targets were identically formulated. In summary, this means that the differences between the branches do not manifest themselves in the formulation of the performance targets but in whether or not they form part of the curriculum.
b. Design of the simulation environment and formulation of skill-based tasks: To represent working life in the simulation, the fictional company called LINCA was created. Its organization is reminiscent of Switzerland's two largest retailers. LINCA is a holding and is subdivided into four organizational units: *holding*, *retail*, *travel agency* and *banking*. All created tasks for measuring basic commercial skills reflect the results of the curriculum analysis and include the four common competence areas mentioned above. The starting point of the sub-tests of each organizational unit was always a workplace situation. The situations were introduced by means of video vignettes, audio recordings or written work assignments. Given the task structure, the response format of the skill-based tasks was open-ended. For example in one task, a customer wants to book a language stay and asks for an offer. The apprentices have to check dates, language schools, flight schedules and exchange rates in order to submit an offer to the customer.
c. Formulation of knowledge-based items: We developed items that measured basic commercial knowledge based on the results of a document analysis of the school curriculum, course books and qualification procedures. Furthermore, the items needed to relate thematically with the skill-based tasks. Although the content was connected, the knowledge-based items could be answered without the context of the skill-based

Table. 2 Tasks and items of *simuLINCA*

Organizational unit	Skill-based tasks	# of items	Knowledge-based items	# of items	Total # of items
Holding Department LINCA	Identify information in the balance sheet, mission statement, and organizational chart and process it	3	Balance sheet, legal and corporate form	2	5
Retail	Carry out a price calculation	5	VAT, exchange of goods, overhead, cash discount, discount, bonus	12	42
	Enter an order into the system	7			
	Notify of a delay in delivery	6			
	Record a business transaction	6	Business transaction in textbook form	6	
Travel Agency	Issue an offer	11	Taxes, profitability analysis	3	18
			Profitability analysis	4	
Banking	Analyse and evaluate customer consultation and formulate further steps	5	Marketing, human resources, negative shareholder equity, share capital	6	11
Total		43		33	76

tasks. The response format of the knowledge-based items consisted of mainly forced-choice items and some open-ended items (e.g. 'What does AIDA stand for?').

d. Digital construction of the instrument: *SimuLINCA* had to be programmed from scratch. In order to facilitate the test procedure, the test could be taken online. Once a task had been completed, however, it was not possible to go back and make revisions. One important reason for this decision was that, in real working life, a sent e-mail or letter cannot be cancelled or reversed.

After development and revisions, the final test was comprised of 76 items, 43 of which were skill-based and 33 knowledge-based. All items could be answered independently of each other so as to ensure local stochastic independence (Lord and Novick 1968). Table 2 provides an overview of the contents of *simuLINCA*. The test duration is 90 min.

Step 5: Appraise existing evidence and collect new evidence as needed

The sources of validity evidence and underlying assumptions (Table 1) were approached with four different methods (presented in Table 3). The test was first applied in two pilot studies and then revised before application in the main study. Therefore, we have three different empirical samples consisting of commercial apprentices (Table 3, notes). Depending on the source of validity evidence, one or more methods were used in order to check for validity. In the next sub-sections, evidence of validity for all five sources are presented in more detail.

Table. 3 Methodical approach and sources of validity evidence for *simuLINCA*

Methodical approach	Sample	Data	Sources of validity evidence				
			A) Test Content	B) Response Processes	C) Internal Structure	D) Other Variables	E) Consequences
I. Document analysis		Tasks and items allocated to contents of school and branch curricula	√				
II. Pilot Study I:Interviews with apprentices in winter 2013	A[a]	20 interviews at workplace (think alouds, ≈ 150 min), recorded on tape, no incentive 20 questionnaires	√	√			
III. Pilot Study II: test application in winter 2013	B[b]	Test data of 102 apprentices Qualitative feedback (open format)	√		√		
IV. Main study in spring 2015	C[c]	Test data of 1365 apprentices					
A) *SimuLINCA*					√		
B) Frequent tasks*			√				
C) Context (CAT, marks)*					√	√	
D) Feedback*							√

* Was assessed/collected during the study of the longitudinal project LINCA (Holtsch and Eberle 2018)

[a] Sample A: N = 20 apprentices, final year of training, employed by companies that are part of the ten largest branches. Age: 18.64 (SD = 1.74). Sex: 60% female. 70% E-Profile

[b] Sample B: N = 102 apprentices, 4 vocational schools, 6 classes, half a year before the end of the training. Sex: 61.8% female. 75.5% E-profile

[c] Sample C: N = 1,365 apprentices, 82 randomly drawn commercial classes. Age: 19.1 years (SD = 1.5). Sex: 63.2% female. 52.5% E-profile

Step 5.1 Validity source A) test content

As Table 3 illustrates, the validity of the source *test content* was approached by means of four different methods (document analysis, first and second pilot study, main study), which are presented and discussed in more detail below.

Evidence from Document Analysis (I) Aim: The aim of the document analysis was to ensure that the test covered important contents of commercial VET and respected the 'all-branch concept'. The document analysis was an essential step in the development of the test as it supported the content definition of the skill-based tasks and knowledge-based items. This demonstrates the important interplay between development and validation.

Method: The method was rechecked regarding how the tasks corresponded to the performance targets of all of the 21 branches (branch-specific curricula) and whether all knowledge-based items could be mapped to the learning objectives of the school curriculum.

Results: Most of the skill-based tasks were found to be relevant for all branches. Only minor revisions were necessary. With regard to the knowledge-based items, they corresponded to the school curriculum.

Evidence from the First Pilot Study (II) Aim: The aim of the structured interviews was, inter alia, to find out whether the apprentices considered the tasks and the items of the test to be manageable and authentic.

Method: First, the apprentices (*Sample A*) completed and/or commented on the test version of the simulation (think alouds) (Ericsson and Simon 1984). Second, they rated the tasks in terms of difficulty on a Likert-type scale ranging from 1, 'not difficult at all', to 6, 'very difficult'. Furthermore, they had to state whether they had already dealt with such tasks and items in one of the three learning venues.

Results: The rating of the difficulty of the skill-based items ranged from $M = 1.84$ ($SD = 1.12$) to $M = 4.47$ ($SD = 1.46$). Half of the apprentices stated that they had done similar tasks at their workplace before, and they assessed most of the skill-based tasks as realistic. The rating of the difficulty of the knowledge-based items varied between $M = 1.69$ ($SD = 0.63$) and $M = 5.00$ ($SD = 1.20$). The tasks or items covered an intended range from easy to difficult. For almost all test items the apprentices specified that they had to deal with similar tasks and items at vocational school and their previous training situations.

Evidence from the Second Pilot Study (III) Aim: One aim of the second pilot study *(Sample B)* was to evaluate apprentices' spontaneous comments on the simulation in terms of test content.

Method: The apprentices were given 270 min to complete *simuLINCA* and give feedback on the simulation in an open format.

Results: Most of the participants reported that the tasks connected to everyday work and that the test provided a balanced mixture of knowledge-based items and practical tasks. The apprentices confirmed that the documents as well as the tasks had been realistic and appropriate in terms of difficulty. Nevertheless, some of the tasks and items had to be restructured and/or simplified in some places because they had been (too) difficult and had not been clearly formulated.

Evidence from the Main Study (IV) Aim: An analysis of the routine activities of apprentices was carried out in order to clarify whether the test content corresponded to real working tasks.

Method: The commercial apprentices of *Sample C* had to write down the three most important activities they typically perform at their training company. In a first step, the activities mentioned were coded according to the eight competence areas (see Step 0). In a second step, the activities themselves were counted in a frequency analysis.

Table. 4 Model characteristics of *simuLINCA*

Model characteristics	1 PL—IRT model		
	One-dimensional	Two-dimensional	
		Skill-based	Knowledge-based
N	1365	1365	
N items	76	43	33
EAP/PV reliability	0.877	0.857	0.778
WLE reliability	0.892	0.874	0.721
Item separation reliability	0.997	0.997	
Deviance	135,839.23	134,985.85	
Estimated parameters (number)	116	118	
AIC	136,071.23	135,221.85	
Variance	0.391	0.481	0.368

Calculation on the basis of the Gauss-Hermite Quadrature with 225 nodes

Results: Shortly before the completion of the apprenticeship, 41.0% of the apprentices ($N=1350$) dealt with administrative and organizational tasks, 28.4% handled orders, and 24.5% executed financial processes on a routine basis. The tasks that the apprentices ($N=1353$) reported to occur most often were telephone service (19.7%), accounting (13.6%) and payment and invoices (13%).

Step 5.2 Validity source B) response processes

Evidence from the First Pilot Study (II) Aim: In order to deal successfully with the skill-based tasks of *simuLINCA*, the apprentices (*Sample A*) had to know exactly what they were asked to do and where to find the requisite information within the simulation. As for the knowledge-based items, it was important to ensure that the apprentices understood what was asked of them.

Method: In interviews (think alouds), the apprentices stated whether and how they understood the questions and the work assignments of the simulation and described how they would approach them. In addition, they were asked to explain a selection of basic terms such as 'holding' and 'net method'. The apprentices' account of how they would approach the tasks allowed for drawing conclusions about their cognitive processes.

Results: The apprentices understood what (working) processes were required and where they could find relevant information so as to be able to deal with the tasks and items successfully. Furthermore, the knowledge-based items were clearly formulated so that the apprentices knew what they were expected to do (irrespective of whether they solved the task correctly).

Step 5.3 Validity source C) internal structure

Evidence from the Main Study (IV) Aim: Checking whether the test measures the construct reliably and whether the assumption of two-dimensionality (skill-based vs. knowledge-based) could be confirmed empirically (*Sample C*) was of interest.

Method: We analysed the item property and the dimensionality of *simuLINCA* in ConQuest 4.0 (Wu et al. 2015) and compared the fit of a one-dimensional 1 PL-IRT

partial-credit model with a two-dimensional partial-credit model (Table 4). The items were selected on the basis of infit statistics (mean squares, MNSQ) (Wright and Linacre 1994) and *t*-values. Wilson (2005) recommended analysing the MNSQ and *t*-values in combination and excluding items only if both values point to a misfit as *t*-values are prone to become significant in a large sample and where a large number of items is concerned.

Results: In the one-dimensional model, the *t*-values of 16 items proved to be greater than 2. As the MNSQ of the 76 items fell into a satisfying value range (0.87 to 1.17); no items were excluded. The EAP/PV reliability of the latent abilities was 0.88, the WLE reliability was 0.89 and the AIC was 136,071.23. In the two-dimensional model, distinguishing the dimensions 'skill-based' and 'knowledge-based,' the *t*-values of 13 items were greater than 2, but the MNSQ of the items of all task types fell into a satisfying value range (0.85 to 1.20). The EAP/PV reliabilities of the latent abilities were 0.86 for the skill-based items and 0.78 for knowledge-based items. The WLE reliabilities were 0.87 and 0.72, and the AIC was 135,221.85. The two dimensions correlated to the extent of $r = 0.613$.

According to Adams and Wu (2010, p. 3), the one-dimensional model is a sub-model of the two-dimensional model and the deviance of the two models 'is distributed as a chi-square with two degrees of freedom.' As the deviance of the two-dimensional model is 853.38 smaller than the deviance of the one-dimensional model, and as the AIC of the two-dimensional model is also smaller than the AIC of the one-dimensional model, the two-dimensional model proved to fit the data better (Table 4).

The Wright Map (Fig. 1) of the two-dimensional model indicates that some test items (especially skill-based items) were too difficult for the apprentices. Overall, we consider item distribution and reliability to be satisfactory.

Step 5.4 Validity source D) relations to other variables

Evidence from the Main Study (IV) Aim: As there is some empirical evidence that cognitive skills as well as prior knowledge are relevant to the development of professional knowledge and skills (e.g. Rosendahl and Straka 2011; Helm 2015; Atik and Nickolaus 2016), the aim was to determine whether there were such correlations in our data as well.

Method: Apprentices' cognitive knowledge (measured by means of the CAT test[7]) was introduced into the model as a *domain-unspecific predictor*. Further, the marks they achieved in the school subjects ICA and E&S[8] in the previous school year were introduced as *domain-specific predictors*. Given the existing empirical evidence, we expected that higher CAT scores, and in particular better marks in ICA and E&S, would have a positive effect on the test scores obtained in *simuLINCA*. Furthermore, M-profile apprentices were assumed to score higher than E-profile apprentices because the M-profile is (cognitively) more demanding than the E-profile.

[7] The revised short form of the cognitive ability test (CAT) 4th to 12th grade (Heller and Perleth 2000), which included the subtests V3 (word analogies, 20 items), Q2 (series of numbers, 20 items) and N2 (character analogies, 25 items), was applied in November/December 2012 in *Sample C*.

[8] The analysis was based on marks as reported by the apprentices. We could not verify whether this information was correct. In the Swiss education system, marks range from 1 (fail) to 6 (excellent), with 4 representing the passing mark. Dickhäuser and Plenter (2005) found that the accuracy of self-reported school marks and the teachers' evaluation highly correlate and that the reported school marks were, on average, higher than they had actually been.

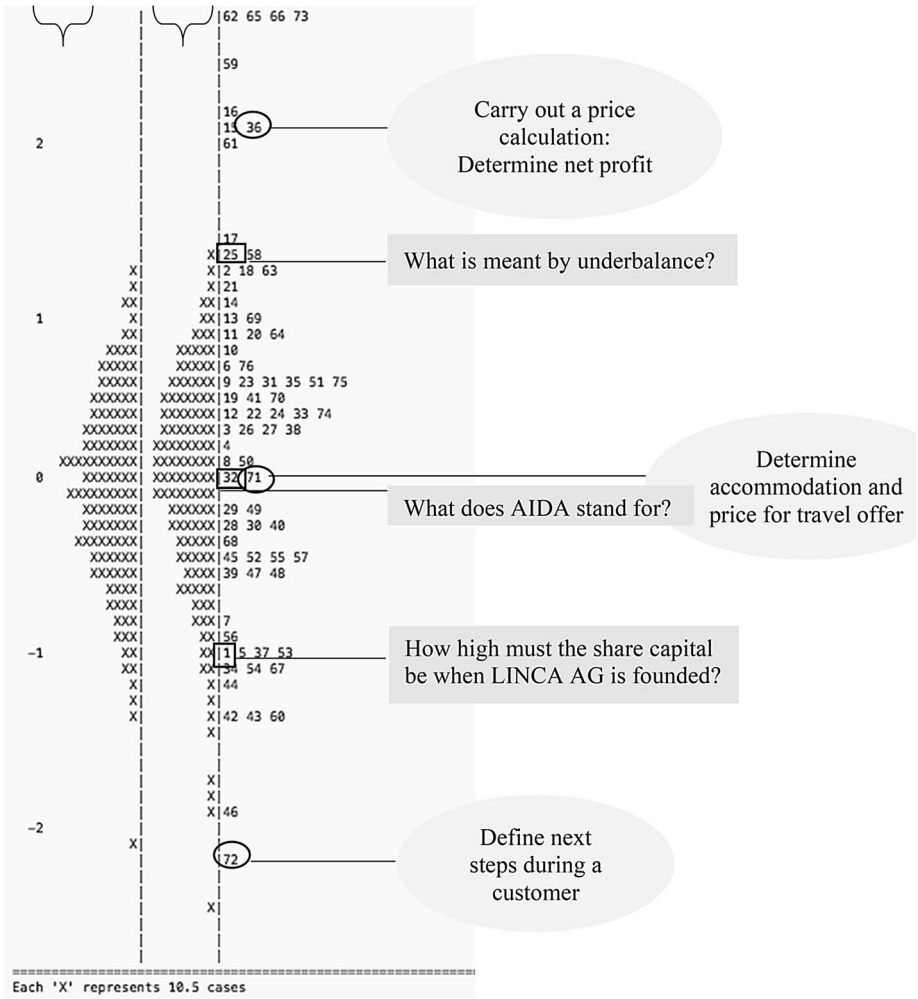

Fig. 1 Wright Map of the two-dimensional model. Items 1–33 belong to the first dimension [1–25 are dichotomous, 26–33 are partial credit]. Items 34–76 belong to the second dimension [34–49 are dichotomous, 50–76 are partial credit]). Ovals = sample items for skill-based dimension. Rectangles = sample items for knowledge-based dimension

A structural equation model (SEM) was calculated by means of the statistics software Mplus, Version 7.2 (Muthén and Muthén 2012). The latent CAT variable is a latent second-order variable that parcels items at the lowest level. The E-profile apprentices acted as the reference group of the dichotomous profile variable. In order to avoid a substantial bias in the parameter estimates, which may arise if the hierarchical data structure is not respected (Asparouhov 2005), a pseudo-maximum-likelihood estimator (PML) (type = complex) (Muthén and Muthén 2012, p. 249) was implemented in the SEM. This procedure took the hierarchical data structure into account (Muthén and Muthén 2012).

Results: The goodness of fit indices were satisfying ($\chi^2 = 190.046$, $df = 67$, $p < 0.01$; CMIN = 2.83; CFI = 0.987; RMSEA = 0.037; SRMR = 0.021). Confirming the

expectations, the context factors correlated significantly with the test score of *simuLINCA*. The previously achieved marks in ICA proved to have the strongest influence on basic commercial knowledge and skills ($\beta = 0.269$; $p < 0.001$), followed by the profile (E-profile apprentices, $\beta = -0.237$; $p < 0.001$). In line with expectations, M-profile apprentices scored higher on *simuLINCA* on average than E-profile apprentices. Furthermore, the E&S marks achieved in the previous year correlated significantly with the results of *simuLINCA* ($\beta = 0.208$; $p < 0.001$). The same applied to cognitive knowledge (CAT; $\beta = 0.093$; $p < 0.05$). As this effect proved to be rather small, however, we concluded that previously obtained marks (domain-specific predictors) were more relevant with regard to the apprentices' achievement in *simuLINCA* which is in line with previous research (e.g. Rosendahl and Straka 2011).

Step 5.5 Validity source E) consequences

Evidence from the Main Study (IV) Aim: Test results should be made available to apprentices *(Sample C)* after they have taken the test and should provide a status quo on their ability to successfully complete representative commercial tasks and to answer BA-type final exam questions. Furthermore, it should also be possible to make the results accessible (in an anonymised form) to other VET stakeholders.

Method: Two months after testing, the apprentices had the opportunity to download a PDF file with their test scores for both dimensions compared to the average scores of their peers and those of the entire sample. Their VET teachers in E&S received the average scores of the class performance in comparison to the whole sample (Table 3, Sample C).

Results: The form of feedback seems to be neither useful nor attractive to apprentices, which is implied by the fact that only a small number of apprentices downloaded the file.

Step 6: Formulate/synthesize the validity argument in relation to the interpretation-use argument

The assumptions formulated in the IUA (Step 3, Table 1) were compared and evaluated with the collected evidence for validity for each source (Step 5).

Test content

The assumptions concerning test content could be met. Nevertheless, in order to increase the authenticity of the test, further measures could be taken. Although the process of item selection was careful, well-founded and based on competence areas that are relevant to all branches, the items covered only a part of the knowledge and skills that the apprentices are expected to have acquired by the end of their apprenticeship. With regard to the representativeness in terms of the breadth of competence testing and construct representation, a larger item pool should be developed.

Response processes

The interviews showed that the participating commercial apprentices understood the tasks and items and that they were able to process available information cognitively. However, it became apparent during the second pilot study that some tasks/items were nevertheless too demanding. In order to gain deeper insights into the response

processes, it would be useful to analyse, for example, eye movements, solution duration or the use of documents.

Internal structure

The data provided empirical support for the two-dimensionality of the construct, but the one-dimensional model showed, in terms of quality criteria, an acceptable fit as well. This implies that both models are tenable in principle. This is interesting for theoretical considerations on the assumed competence structure model, as it may indicate that both dimensions, even if distinguishable, belong to the same construct. It remains to be clarified to what extent the two dimensions have a differentiating effect, whether they can be predicted by different predictors or whether they operate in the same way. It is also unclear at this point whether the two dimensions also differ because of the underlying task formats (open vs. closed) (Blömeke et al. 2015).

The test measures basic commercial knowledge and skills with satisfactory reliability. As the evaluation of test results (e.g. Wright Map) showed, the test contained a relatively high number of tasks that were too difficult for the apprentices, especially concerning skills. On the one hand, this might indicate that the design of the test was too demanding. In a next test version, more tasks of varying levels of difficulty should be included. On the other hand, it seems that the participating apprentices had not yet developed certain knowledge and skills that they were expected to possess at the end of their training.

Relation to other variables

As expected, M-profile apprentices and apprentices with good marks in ICA and E&S and higher CAT scores perform better on *simuLINCA* than E-profile apprentices and apprentices with lower marks and lower CAT scores respectively.

Consequences

The assumptions regarding the consequences could not be fully confirmed. An instrument is available for measuring knowledge and skills in commercial VET, but the feedback is not fully suitable for determining the status quo. It turned out that the form of feedback did not appeal to the apprentices. One problem seems to be that test participants did not receive immediate feedback as the coding was time-consuming. The time delay could have been a reason why the participants did not download the PDF file. To reduce the temporal distance, the coding should be automatized as far as possible to make instant feedback possible.[9] It could also have been a problem of form. Since only one test score was reported for each dimension, apprentices had little concrete information about their status quo and the information remained abstract. Meaningful feedback would need to be more detailed. A possibility would be to give apprentices feedback on their level of competence, including the type of tasks and items they can use it to solve.

[9] For an example in the commercial domain, see Egloffstein et al. (2016).

Step 7: Keep track of practical issues including cost

Step 4 (development of new instrument) and Step 5 (evidence collection) provided information on practical issues, which can be subsumed under the aspects of test administration, technical implementation and cost.

Step 7.1 Aspects of test administration

Prior to the testing, the test administrators had to ensure that the *computer infrastructure* was working properly (e.g. no immediate need for system updates, sound system of each computer turned on, stable internet connection). Therefore, contacting the IT department in advance was crucial for a smoothly running test. During the testing, an administrator tool with which administrators were able to check whether test takers had been able to log in and how many tasks they had dealt with at any given point proved to be important.

Step 7.2 Aspects of technical implementation

During the pilot studies, problems with the online version occurred. It took the data (e.g. videos) too long to load. Therefore, it was necessary to ensure that the loading of the data took place before the apprentices started the test (keyword buffering) and to have a printed back-up version (e.g. written dialogues and all documents on paper) at hand. In addition, it proved to be crucial that no further software installations were necessary. Therefore, the test ran on a functioning browser (e.g. Chrome, Firefox).

Step 7.3 Aspects of cost

The main costs—besides the personnel costs (test designers)—were the programming costs (approximately CHF 30,000).[10]Another costly item can be filming and editing of videos, especially if actors are hired to increase authenticity. Since we did this ourselves, we cannot give a specific number. However, creating the final videos and audio recordings took about one work week. Additionally, creating authentic documents was very time-consuming (approximately one month of working time). To contact the IT departments in advance required increased personal resources. Furthermore, the need to code some of the open items (six tasks/items) by hand led to additional costs. For each task/item, two people coded for about two work weeks.

Step 8: Make a judgment: does the evidence support the intended use?

On the basis of the evidence summarized above, we conclude that with *simuLINCA*, an instrument is now available for measuring basic commercial knowledge and skills of Swiss apprentices. The definition of the construct indicated a need to include, on the one hand, questions about basic concepts of BA and, on the other hand, basic commercial tasks of the type that apprentices must deal with in real life. The validity argument showed, inter alia, that *simuLINCA* considers the learning objectives of both learning venues (school and company) and is judged to contain realistic and authentic (working) tasks. Consequently, the instrument meets the demands of the proposed

[10] CHF 30,000 equal approximately 33,340 US dollars or 27,340 Euros (exchange rate as on 30.05.2021).

interpretation-use argument. Nevertheless, the test instrument and also the underlying construct description could benefit from further development, revision and extension. To describe the *construct* (*internal structure*) in more detail, a competence level model for basic commercial knowledge and skills should be developed. Therefore, how a meaningful categorization (knowledge and skills for the respective competence areas) can be made should be considered. This could be taken into account in the development of a broadened item and task pool *(test content*). Concerning the *relation to other variables*, we decided on two domain-unspecific context factors (CAT, profile) and two domain-specific ones (marks). However, there would also be arguments in favour of considering other context factors, such as other simulations in the commercial area or accounting tests (e.g. Guggemos 2016). Especially with regard to *consequences*, that is, the feedback of the status-quo and the implementation, *simuLINCA* could be improved. The feedback of the results needs to be communicated in a more attractive form. Therefore, we will further address the question of how to give an instant feedback as well as to produce an anonymous overview of the test results for other stakeholders. This involves different technical features, such as automatic coding of open answers as well as descriptors for competence levels. In order to implement *simuLINCA* in vocational schools or companies, for example, as a learning tool, the test design would have to be adapted. The test could consist of individual modules (e.g. the eight competence areas) and include branch-specific tasks. Furthermore, an adaptive format is conceivable. In order to make diagnostic statements, the use of 'cognitive diagnostic models' (DiBello et al. 2007) should be considered. In contrast to unidimensional item response theory (IRT) models, cognitive diagnostic models can provide an explanation for why a test taker does not perform well based on skills that could not be applied (Henson et al. 2008).

All in all, *simuLINCA* contributes to evaluating whether VET equips future commercial staff with the knowledge and skills that they need in their workplace and thus to ensuring that the objective of training a skilled commercial workforce is achieved. The results of the competence measurement are not only important for the tested persons themselves but especially for designers of workplace and school learning (VET teachers, trainers, researchers).

Conclusion

The aim of the paper was to present and implement the approach to validation proposed by Cook and Hatala (2016) for valid competence measurement in VET using the example of the development and validation of a measurement instrument for commercial VET in Switzerland.

As has been argued, validation of developed instruments is often not reported and remains hidden from interested researchers. The strength of the presented—now nine-step—approach is that it provides practical guidelines that help perform and structure the measurement process. Nonetheless, the approach is kept flexible enough to allow the integration of distinct concepts. For example, the evidence-centered model (Mislevy et al. 2003) or the assessment triangle (Pellegrino et al. 2001) could be taken into account in the development part or the setting up of assumptions could be adapted to one's own ideas. In order to structure our assumptions, we used the five sources of evidence provided by AERA et al. (2014). However, there are other methods and frameworks

available that could also be combined. For instance, applying Kane's (2013) inferences-categories would set a focus on the implementation of the test, which would be a nice add-on. Therefore, we suggest thinking about a consolidation of the inferences-categories by Kane (2013) and the sources provided by AERA et al. (2014). Carney et al. (2019) discussed this idea and emphasized that further research should be done on this topic.

Applying the framework supports the increase of transparency of the measurement process, that is, the documentation and reporting of results. In such a way, research groups and/or test users can assess the suitability of an instrument for their test purposes or the effort required for their own validation study. Since the framework not only considers content aspects, but also takes resource issues into account, it is particularly suitable for computer-based instruments. The differences in test development and validation of paper–pencil and computer-based formats especially manifest themselves concerning the required resources (e.g. programming, costs).

A disadvantage of the framework is that performing the nine steps means that the explanations for each step can become very detailed, and the overview can be lost. In our case, the presentation of the evidence collection (Step 5) is very dense. Additionally, for an article to be published, it must not exceed a certain word length. Therefore, it would be useful to create a condensed version and make additional materials available on an online platform. This would help maximize transparency and is in accordance with the open scientific thinking.

As Cook and Hatala (2016) created their approach for the occupational competence field (medicine) and for validating computer-based instruments, no major changes to the framework were necessary for VET and occupational competence measurement. Content-wise, it was only necessary to add Step 0. Since VET systems vary widely from country to country, it is important to describe the measurement context and target group more precisely than is necessary in the medical field. After extending and slightly modifying Cook and Hatala's framework, Table 5 gives an overview of the aims and essential questions one could ask during each step and provides in such a way a practical and easy-to-use validation framework for competence measurement in VET.

To advance competence research in VET, researchers would be required to publish articles on the development and the validation process of test instruments. This is the only way to assess what kind of constructs were actually measured and to evaluate the extent to which results are comparable. Thus, carefully conducted and transparent measurement procedures would also facilitate (international) replication studies and adaptation processes of measurement instruments. Moreover, such a trend would lay the groundwork for an in-depth discussion on the concept of validity (three-part or unified concept) and on the validation process (advantages and disadvantages of different frameworks and approaches) in the VET sector. As another starting point, it would be desirable to extend Rüschoff's (2019) review on German borders by compiling the

Table. 5 Practical validation framework for competence measurement in VET

Step to validity	Aims	Essential questions
0. Define the measurement context and target group	To make the context (e.g. learning objectives, VET system) of the study explicit and to understand the characteristics of the context and the target group	What is the context of the study? What are the characteristics of the context? Who is the target group? Do special requirements have to be considered when testing the sample?
1. Define the construct	To make the underlying theoretical concepts and their functioning or relation transparent	How is the construct defined? On which theoretical considerations/model is the construct based on? Which dimensions of the construct should be measured? How does the construct operate (with regard to other related constructs)?
2. Make explicit the intended interpretation, use and decision(s)	To formulate which decisions are taken on the basis of the test results, e.g., entrance or final examination, determination of status quo, selection or qualification procedure	What are the test results used for? What should the test results not be used for? What decisions are made based on the test results? Who receives the test results and in what form?
3. Define the interpretation-use argument, and prioritize needed validity evidence	To formulate comprehensive assumptions regarding the test construction, which will be tested in Step 5 using appropriate methods Different frameworks can be used for evidence collection. The AERA et al. (2014) proposed *test content, response processes, internal structure, relations to other variables* and *consequences*	Which is the most important validity evidence? ***Test Content*** What is expected in terms of what the test is supposed to measure? ***Response Processes*** What is expected of the participants to think or understand when completing the test? ***Internal Structure*** What is expected with regard to the dimensionality of test content (one-dimensional, multi-dimensional)? How reliable is the test? ***Relations to Other Variables*** How are the achieved test scores expected to be related to the outcomes of other constructs (e.g. intelligence, prior knowledge, motivation)? ***Consequences*** What intended and unintended consequences (impact, decisions, actions) does the test entail?

Table. 5 (continued)

Step to validity	Aims	Essential questions
4. Identify candidate instruments and/or create/adapt a new instrument	To check already available instruments with regard to the fit of the above formulated intentions and assumptions To adapt or develop a new instrument with regard to the requirements	What kind of assessment tool is the best to measure the construct? Are there already validated instruments available that meet the purpose of the intended measurement in part and are adaptable to the context and target group? What are the requirements of test construction with regard to the previous considerations (e.g. context, target group, construct definition, internal structure)?
5. Appraise existing evidence and collect new evidence as needed	To appraise existing evidence and to collect new evidence according to the interpretation-use argument (Step 3)	Is there existing evidence for the validity of items/tasks? And if so, is more evidence needed to for example fit the target group? What kind of method is the best to gather the evidence? In what form will the evidence be documented, analysed and reported?
6. Formulate/synthesize the validity argument in relation to the interpretation-use argument	To compare the proposed assumptions with the evidence found	Was it possible to collect sufficient evidence for all assumptions? Is there a need to review the instrument further, to change or revise some tasks? Is there a need to gather more evidence? If so, what evidence and by what method? Are evidence gaps recognizable?
7. Keep track of practical issues including cost	To report on resources required for test administration, costs and technical aspects	What resources are needed to implement the instrument? How much does it cost to develop the instrument (e.g. programming, media, personal)? How much does it cost to gather evidence? What kind of problems might arise during the implementation? What is helpful to implement the instrument? How much does it cost to apply the test? What kind of equipment is needed? Do test administrators need special knowledge and skills?
8. Make a judgment: does the evidence support the intended use?	To make a final judgement on whether the available evidence is in accordance with the intended use of the test	Is there enough evidence available to support the intended test use? How can evidence gaps be tackled/approached? What are next steps?

existing research on (computer-based) competence measurement. It lends itself to use the review process of Cook et al. (2013) as the basis for such an extended review and to structure it along the nine-steps approach to validation.

Abbreviations
AERA: American Educational Research Association; BA: Business and administration; CAT: Cognitive ability test; CFI: Comparative fit index; CMIN: Chi-square statistics; df: Degree of freedom; E&S: Economics and society; EFZ: Federal VET diploma; ICA: Information, communication, and administration; IRT: Item response theory; IUA: Interpretation-use argument; MNSQ: Mean squares; PML: Pseudo-maximum-likelihood estimator; RMSEA: Root mean squareerror of approximation; SEM: Structural equation model; SRMR: Standardized root mean square residual; VET: Vocational education and training.

Acknowledgements
We are grateful for the valuable feedback and support from Prof. Dr. Doreen Holtsch and Prof. Dr. Franz Eberle. We would like to thank the SERI for funding.

Authors' contributions
SRM conceptualization, data collection, processing and analysis, writing and editing original manuscript. SFH conceptualization, data analysis, writing parts of the original manuscript, editing the manuscript. Both authors read and approved the final manuscript.

Authors' information
Silja Rohr-Mentele is a doctoral candidate at the Institute of Education at the University of Zurich. Her research interests are (technology-based) competence modeling and measuring, and learning in commercial vocational education and training.

Dr. Sarah Forster-Heinzer is the Head of Secondary Education II at the University of Teacher Education Lucerne. Her research interests are competence measurement and development, professional ethos, social interaction and communication, research on vocational education and training.

Funding
This research has been funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) through its Leading House 'Learning and Instruction for Commercial Apprentices' (LINCA). The funding body had no influence on the design of the study, collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
This study uses data from the Leading House Project 'Learning and Instruction for Commercial Apprentices' (LINCA), a data set that is provided by the Institute of Education, Secondary School Teaching with a focus on Business Education, at the University of Zurich. The Institute of Education at the University of Zurich is responsible for granting access to this data source. The data protection guidelines of the data provider do not permit the publication of data files.

Declarations

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

Author details
[1]Institute of Education, University of Zurich, Zurich, Switzerland. [2]University of Teacher Education Lucerne, Frohburgstrasse 3, 6002 Luzern, Switzerland.

Received: 23 February 2021 Accepted: 26 July 2021
Published online: 12 August 2021

References
Achtenhagen F, Winther E (2009) Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz—am Beispiel der Ausbildung von Industriekaufleuten. Bericht an das Bundesministerium für Bildung und Forschung (K350600) [Construct validity of simulation tasks: computer-aided measurement of professional competence—using the example of the training of industrial clerks]. In: Seminar für Wirtschaftspädagogik, Georg-August-Universität Göttingen, Göttingen

Adams R, Wu M (2010) Multidimensional models. https://www.acer.org/files/Conquest-Tutorial-7-MultidimensionalModels.pdf. Accessed 13 Jan 2021

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999) Standards for educational and psychological testing. American Educational Research Association, Washington, D.C.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (2014) Standards for educational and psychological testing. American Educational Research Association, Washington, D.C.

Anastasi A (1986) Evolving concepts of test validation. Annu Rev Psychol 37:1–15. https://doi.org/10.1146/annurev.ps.37.020186.000245
Asparouhov T (2005) Sampling weights in latent variable modeling. Struct Equ Model Multidiscip J 12:411–434. https://doi.org/10.1207/s15328007sem1203_4
Atik D, Nickolaus R (2016) Die Entwicklung berufsfachlicher Kompetenzen von Anlagenmechanikern im ersten Ausbildungsjahr [The development of professional competences of plant mechanics in the first year of training]. Zeitschrift für Berufs- und Wirtschaftspädagogik 112:243–269
BBT [Federal Office for Professional Education and Technology] (2017) Bildungsplan Kauffrau/Kaufmann EFZ vom 26. September 2011 für die betrieblich organisierte Grundbildung (Stand am 1. Mai 2017) [Training plan for commercial apprentices for in-company training]. Bundesamt für Berufsbildung und Technologie, Bern
Blömeke S, Gustafsson J-E, Shavelson R (2015) Beyond dichotomies: competence viewed as a continuum. Zeitschrift für Psychologie 223:3–13. https://doi.org/10.1027/2151-2604/a000194
Bloom B, Englehart M, Furst E et al (1956) Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. Longmans, Greene
Brennan RL (2006) Perspectives on the evolution and future of educational measurement. In: Brennan RL (ed) Educational measurement, 4th edn. Praeger, Westport, pp 1–16
Carney M, Crawford A, Siebert C et al (2019) Comparison of two approaches to interpretive use arguments. Appl Meas Educ 32:10–22. https://doi.org/10.1080/08957347.2018.1544138
Cook DA, Hatala R (2016) Validation of educational assessments: a primer for simulation and beyond. Adv Simul 31:1–12. https://doi.org/10.1186/s41077-016-0033-y
Cook DA, Brydges R, Zendejas B et al (2013) Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. Acadamic Med 88:872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf
Cook DA, Brydges R, Ginsburg S, Hatala R (2015) A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 49:560–575. https://doi.org/10.1111/medu.12678
DiBello LV, Roussos L, Stout W (2007) Review of cognitively diagnostic assessment and a summary of psychometric models. In: Rao CR, Sinharay S (eds) Handbook of statistics. Elsevier, Amsterdam, pp 979–1030
Dickhäuser O, Plenter I (2005) "Letztes Halbjahr stand ich zwei". Zur Akkuratheit selbst berichteter Noten [On the accuracy of self-reported school marks]. Zeitschrift für Pädagogische Psychologie 19:219–224. https://doi.org/10.1024/1010-0652.19.4.219
Egloffstein M, Brandt S, Eigenmann R et al (2016) Modellierung und Erfassung domänenspezifischer Problemlösekompetenz von Industriekaufleuten—Produkte und Entwicklungsperspektiven des Projekts DomPL-IK [Modeling and measurement of domain-specific problem solving competence of industrial clerks—products and development perspectives of the DomPL-IK project]. In: Dietzen A, Nickolaus R, Rammstedt B, Weiss R (eds) Kompetenzorientierung. Berufliche Kompetenzen entwickeln, messen und anerkennen. W. Bertelsmann Verlag, Bielefeld, pp 149–171
Ericsson KA, Simon HA (1984) Protocol analysis: verbal reports as data. The MIT Press, Cambridge
Gafni N (2016) Comments on implementing validity theory. Assess Educ Princ Policy Pract 23:284–286. https://doi.org/10.1080/0969594X.2015.1111195
Geisinger KF (2016) Intended and unintended meanings of validity: some clarifying comments. Assess Educ Princ Policy Pract 23:287–289. https://doi.org/10.1080/0969594X.2016.1158150
Guggemos J (2016) Modellierung und Messung von Kompetenz im Externen Rechnungswesen [Modeling and measurement of competence in external accounting]. Verlag Dr. Hut, München
Hammond JW, Moss PA (2016) Validity theory in measurement. In: Peters MA (ed) Encyclopedia of educational philosophy and theory. Springer, Singapore, pp 1–5
Heller KA, Perleth C (2000) Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4–12+ R). Hogrefe, Göttingen
Helm C (2015) Determinants of competence development in accounting in upper secondary education. Empir Res Vocat Educ Train 10:1–36. https://doi.org/10.1186/s40461-015-0022-8
Henson RA, Templin JL, Willse JT (2008) Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika 74:191. https://doi.org/10.1007/s11336-008-9089-5
Holtsch D, Eberle F (eds) (2018) Untersuchungen zu Lehr-Lernprozessen im kaufmännischen Bereich. Ergebnisse aus dem Leading House LINCA und Schlussfolgerungen für die Praxis [Studies on teaching-learning processes in the commercial sector. Results from the Leading House LINCA and conclusions for practice]. Waxmann, Münster
Holtsch D, Rohr-Mentele S, Wenger E et al (2016) Challenges of a cross-national computer-based test adaptation. Empir Res Vocat Educ Train 18:1–32. https://doi.org/10.1186/s40461-016-0043-y
Jude N, Wirth J (2007) Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen [New opportunities in technology-based assessment of competencies]. In: Hartig J, Klieme E (eds) Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Bundesministerium für Bildung und Forschung (BMBF), Bonn, pp 49–56
Kane MT (1992) An argument-based approach to validity. Psychol Bull 112:527–535. https://doi.org/10.1037/0033-2909.112.3.527
Kane MT (2001) Current concerns in validity theory. J Educ Meas 38:319–342
Kane MT (2013) Validation as a pragmatic, scientific activity. J Educ Meas 50:115–122
Kane MT (2016) Explicating validity. Assess Educ Princ Policy Pract 23:198–211. https://doi.org/10.1080/0969594X.2015.1060192
Lord FM, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley, Reading
McClelland DC (1973) Testing for competence rather than for "intelligence." Am Psychol 28:1–14. https://doi.org/10.1037/h0034092

Mentele S, Heinzer S, Lekic C, Holtsch D, Eberle F (2014) Entwicklung eines computerbasierten Instrumentes LINCA zur Erfassung des kaufmännischen Wissens und Könnens von Lernenden in der deutschsprachigen Schweiz [Development of the computer-based instrument LINCA for the measurement of commercial knowledge and skills]. Universität Zürich, Institut für Erziehungswissenschaft, Abteilung Lehrerinnen- und Lehrerbildung Maturitätsschulen, Zürich

Messick S (1975) The standard problem. Meaning and values in measurement and evaluation. Am Psychol 30:955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick S (1989) Validity. In: Linn RL (ed) Educational measurement. American Council on Education and National Council on Measurement in Education, Washington, D.C.

Mislevy RJ, Almond RG, Lukas JF (2003) A brief introduction to evidence-centered design. ETS Res Rep Ser 2003:i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Muthén LK, Muthén BO (2012) Mplus user's guide, 7th edn. Muthén & Muthén, Los Angeles

Newton PE, Baird J-A (2016) The great validity debate. Assess Educ Princ Policy Pract 23:173–177. https://doi.org/10.1080/0969594X.2016.1172871

Nickolaus R (2018) Kompetenzmodellierungen in der beruflichen Bildung—eine Zwischenbilanz [Competency modeling in vocational education and training—an interim assessment]. In: Schlicht J, Moschner U (eds) Berufliche Bildung an der Grenze zwischen Wirtschaft und Pädagogik: Reflexionen aus Theorie und Praxis. Springer Fachmedien Wiesbaden, Wiesbaden

Nickolaus R, Seeber S (2013) Berufliche Kompetenzen: Modellierungen und diagnostische Verfahren [Professional Competencies: modeling and diagnostic procedures]. In: Frey A, Lissmann U, Schwarz B (eds) Handbuch Berufspädagogische Diagnostik. Beltz, Weinheim, pp 155–180

Pellegrino JW, Chudowsky N, Glaser R (2001) Knowing what students know. The Science and Design of Educational Assessment. National Research Council, Washington, D.C.

Preckel F, Holling H (2006) Die Rolle von Intelligenz und Begabung für Handlungskompetenz am Beispiel beruflicher Hochbegabung [The role of intelligence and aptitude for professional competence using the example of professional giftedness]. Bildung und Erziehung 59:167–178

Rosendahl J, Straka GA (2011) Effekte personaler, schulischer und betrieblicher Bedingungen auf berufliche Kompetenzen von Bankkaufleuten während der dualen Ausbildung [Effects of personal, school and company conditions on the professional competence of bank clerks during dual training], ITB-Forschungsberichte No. 51. Universität Bremen, Bremen. https://elib.suub.uni-bremen.de/edocs/00102039-1.pdf. Accessed 13 Jan 2021

Rosenheck M (2010) Kaufmännische Bildung in der Schweiz: Ausbildungsberuf Kaufleute zwischen Allrounder und Splitterberufen [Commercial training in Switzerland: commercial training between all-rounders and specialist professions]. Zeitschrift des Bundesinstituts für Berufsbildung. 31–33

Rüschoff B (2019) Methoden der Kompetenzerfassung in der beruflichen Erstausbildung in Deutschland. Eine systematische Überblicksstudie [Methods of competence assessment in initial vocational training in Germany. A systematic review]. Bundesinstitut für Berufsbildung, Bonn

Seeber S, Nickolaus R, Winther E et al (2010) Kompetenzdiagnostik in der Berufsbildung: Begründung und Ausgestaltung eines Forschungsprogramms. Berufsbildung in Wissenschaft und Praxis 1:1–15

Shavelson RJ (2010) On the measurement of competency. Empir Res Vocat Educ Train 2(1):41–63

Shepard LA (2016) Evaluating test validity: reprise and progress. Assess Educ Princ Policy Pract 23:268–280. https://doi.org/10.1080/0969594X.2016.1141168

SKKAB [Swiss Conference of Commercial Training and Examination Branches] (2020) Ein Beruf, viele Branchen [One profession, many branches]. https://www.skkab.ch/berufsinformationen/branchen. Accessed 13 Jan 2021

Sloane PFE (2006) Weiterbildung des betrieblichen Ausbildungspersonals [Further training of company trainers]. In: Euler D (ed) Facetten des beruflichen Lernens. h.e.p. Verlag, Bern, pp 449–499

Wilson M (2005) Constructing measures: an item response modelling approach. Lawrence Erlbaum Associates, Mahwah

Winther E (2010) Kompetenzmessung in der beruflichen Bildung Bildung [Measuring competencies in vocational education and training]. W. Bertelsmann, Bielefeld

Wright BD, Linacre MJ (1994) Reasonable mean-square fit values. Rasch Meas Trans 8:370

Wu ML, Adams RJ, Wilson MR (2015) ACER ConQuest. Version 4.0. Generalised item response modelling software. ACER Press, Camberwell

Zbinden-Bühler A, Volz C (2007) Analyse des beruflichen Handlungsfeldes zur Entwicklung kompetenzorientierter Bildungspläne auf der Basis von Situationsbeschreibung [Analysis of the vocational field of action for the development of competence-oriented curricula on the basis of a description of the situation]. Empirische Pädagogik 21:322–339

Publisher's Note