Empirical Research in Vocational Education and Training

RESEARCH Open Access

Challenges of a cross-national computer-based test adaptation

Doreen Holtsch[1*], Silja Rohr-Mentele[1], Eva Wenger[1], Franz Eberle[1] and Richard J. Shavelson[2]

*Correspondence: doreen.holtsch@uzh.ch
[1] Institute of Education, University of Zurich, Kantonsschulstrasse 3, 8001 Zurich, Switzerland
Full list of author information is available at the end of the article

Abstract

Background: In an increasingly globalized world, the call for internationally comparable competence measurements has emerged. After several international studies on pre-college education, the focus has shifted to international assessments of vocational education and training (VET). VET researchers in Germany developed a computer-based test (ALUSIM) that measures the competence of German apprentices in internationally defined core commercial areas. Our own study deals with the adaptation of this test to Switzerland and a discussion of the challenges involved, with the aim of providing guidance for future adaptations. More specifically, despite commonalities between the German and Swiss VET systems, it is necessary to examine whether the contents and technical aspects of ALUSIM are appropriate for Swiss conditions in order to ensure validity and evidence based on test content.

Methods: Several methods were used to examine the criteria context, construct, and IT components of the German computer-based test ALUSIM in order to successfully adapt it for Swiss commercial VET. To this end we first analyzed and compared the German and Swiss commercial VET systems and commercial curricula (context) to assess whether the tasks of the test were also embedded in the Swiss curriculum and whether any specific Swiss commercial contents were not represented by the test. Second, we interviewed experts in the commercial area to learn more about representative commercial job requirements (construct). Finally, we interviewed apprentices and tested our initial adaptation of ALUSIM to the Swiss context in order to assess the test's IT requirements.

Results: The analysis revealed similarities between the German and Swiss VET and the construct 'commercial competence'. However, commercial work conditions and cultural characteristics differ between these countries and lead to different job requirements. Therefore, only a subset of ALUSIM tasks is valid for the construct 'commercial competence' in Switzerland. Thus, the addition of further tasks for Switzerland would more validly represent the construct 'commercial competence'. Moreover, IT components need to be adjusted because the technical implementation represents the measured construct.

Conclusions: The proceedings and findings of our adaptation study imply that context, construct, and IT components need to be analyzed before doing adaptations. Even when dealing with countries with a similar context such as comparable educational system and language, it is necessary to carefully examine and test an adaptation in advance. Therefore, creating a successful and internationally comparable adaptation is admittedly possible but challenging, costly and time-consuming.

Introduction

Megatrends such as globalization, demographic change, and multiethnic communities impact society. These trends are intertwined, inevitably affecting job and training markets (Franke 2014). With increased globalization comes mobility; working environment and professional life requirements are converging. Thus, workforce knowledge and skills must be internationally competitive, spurring global interest in internationally comparable competence measurements, among other things.

During the most recent decade, a number of studies have addressed how the competence of pre-college learners in compulsory education, college learners and adults are assessed and measured. The international measurement of competence in compulsory pre-college education has a long history, which is exemplified by the Program for International Student Assessment (PISA)[1] and Trends in International Mathematics and Science Study (TIMSS)[2]. Additionally, the OECD's (Organisation for Economic Co-operation and Development) Assessment of Higher Education Learning Outcomes (AHELO)[3] aimed to measure college students' generic skills and their discipline-specific skills in civil engineering and economics upon completing their bachelor's degrees. With regard to assessing of adults' competences, the Programme for the International Assessment of Adult Competencies (PIAAC)[4] measures core competences in literacy, numeracy and problem solving in technology-rich everyday life environments (OECD 2013, p. 4). However, none of these programs assesses the vocational competence required in a professional environment. Consequently, attention has turned to large-scale international assessments (LSA) in vocational education and training (VET).

The European Union developed different instruments to compare qualifications in VET more precisely and transparently than had been done in the past. Namely, the European Qualifications Framework (EQF) and the European Credit System for Vocational Education and Training (ECVET) can compare formal qualifications across national borders.[5] However, the EQF covers only formal reference levels and provides no evidence of actual and available vocational competence at the individual level. In fact, valid and reliable cross-national competence measurement in VET might provide an empirical basis for evaluating the EQF and ECVET classifications (Baethge et al. 2006b, pp. 8–9). In this regard, Seeber et al. (2010, p. 9) emphasize that the challenges of measuring professional competence in VET involve (a) creating complex professional assessment situations, (b) sampling tasks performed in the professional context and (c) creating innovative test methods to assess competence.[6] Thus, authentic and complex job situations are hardly assessable with paper-and-pencil tests alone and should therefore be

[1] PISA was sponsored by the Organisation for Economic Co-operation and Development (OECD) and assessed students' competence in grade 9. 'Around 510,000 students between the ages of 15 years 3 months and 16 years 2 months completed the assessment in 2012, representing about 28 million 15-year-olds in the schools of the 65 participating countries and economies.' (Organisation for Economic Co-operation and Development (OECD) 2014, p. 3).

[2] TIMSS was sponsored by the International Association for the Evaluation of Educational Achievement (IEA), mainly for students in fourth and eighth grade (e.g., Beaton et al. 1996)

[3] For more information, see: http://www.oecd.org/edu/skills-beyond-school/ahelo-main-study.htm. Accessed 3 Dec 2015. Neither Germany nor Switzerland participated in the AHELO feasibility study.

[4] PIAAC was sponsored by the OECD. Switzerland did not participate in PIAAC. See for more information Organisation for Economic Co-operation and Development (OECD) (2013) and http://www.oecd.org/skills/piaac/aboutpiaac.htm. Accessed 29 June 2016.

[5] For more information, see: http://www.ecvet-toolkit.eu/site/introduction/whatisecvet. Accessed 14 June 2016.

[6] We applied another counting (a), (b), (c) than in the original version of Seeber et al. (2010, p. 9).

simulated in either an assessment center or by means of a computer-based test to measure VET competence (Shavelson 2010, 2012). Baethge et al. (2006a, 2006b) and Baethge and Arends (2009) studied the feasibility of an international VET-LSA. Although there are cross-nationally defined core areas in business and administration,[7] Baethge and Arends (2009, p. 83) stated that finding similarities was more challenging than for industrial occupations. Simultaneously, Achtenhagen and Winther (2009) led the development of a test instrument that measures competence in the core commercial areas defined by Baethge and Arends (2009). Specifically, they developed the computer-based test ALUSIM, which measures industrial clerk apprentices' 'commercial competence' in Germany and is designed to be internationally compatible. The test simulates authentic business processes in a fictional company in the metal working industry, ALUSIM. Every commercial challenge (task) begins with a prompt (e.g., a video). Thereafter, test takers must manage common work situations, including sales, purchases, and production. An interactive computer desktop provides all the documents and tools that are required to complete the tasks (Achtenhagen and Winther 2009, Appendices IV and V; Winther and Achtenhagen 2009a). The development of an international VET-LSA was continued in a German research program 'Technology-based Assessment of Skills and Competencies in VET' (ASCOT)[8] between 2011 and 2014.

The Leading House project 'Learning and Instruction for Commercial Apprentices' (LINCA)[9] took the first step to expand the VET-LSA cross-nationally by implementing ALUSIM in German-speaking[10] Switzerland in 2011. The aim was twofold, i.e., (a) to implement a Swiss version of the computer-based test that measures the development of apprentices' commercial competence in a longitudinal study during their apprenticeship in 85 classes and 35 VET schools in Switzerland, and to compare the results cross-nationally, and (b) to take advantage of adapting and implementing an existing computer-based test, which would thereby save money and time.

The main reason that supports the decision to adapt was that the German and Swiss commercial dual VET systems exhibit extensive systematic overlap (Bundesamt für Berufsbildung und Technologie (BBT) 2011a, p. 5). The similarities in the official languages in Germany and German-speaking Switzerland[11] and the results of a feasibility study by Baethge and Arends (2009, pp. 71–84) led to the initial assumption that 'commercial competence' is highly comparable between these two countries; only minor adaptations of the German computer-based test were anticipated for Switzerland. However, Geisinger (1994, p. 304) previously reported difficulties with adapting tests within a given nation because of different cultural and/or life experiences. Therefore, when implementing an instrument internationally, 'the cross-cultural generalizability of the construct(s)' (Tanzer 2005, p. 236) must be ensured in order to obtain a valid assessment. In other words, the construct 'commercial competence' should be understood similarly in different countries (Hambleton 2001, pp. 165–166).

[7] The core areas include purchasing, sales and marketing, stock keeping, financials/accounting, customer service, and organizational activities. Customer communication is included in each core area (Baethge and Arends 2009, p. 84).

[8] For more information, see: http://ascot-vet.net. Accessed 25 June 2016.

[9] For more information, see PROJECT-website: www.linca.uzh.ch. Accessed 25 June 2016.

[10] Switzerland is a multilingual country. The four national languages are German, French, Italian, and Romansh. http://www.swissinfo.ch/eng/languages/29177618, Accessed 1 July 2016.

[11] The study focuses only on the German-speaking part of Switzerland.

Achtenhagen and Winther (2009) established the psychometric quality of ALUSIM (e.g., content, reliabilities), arguing that the simulation provided a valid assessment of the construct 'commercial competence' in Germany. Although Switzerland appears to be an appropriate or even 'best' case for cross-national adaptation of ALUSIM, the question of whether the test also allows valid assessment of the construct in Switzerland first had to be addressed.

This paper thus addresses the issue of whether an adaptation is possible and what steps must be taken before deciding to adapt a test. To this end, in this study, we report in detail on an investigation into the extent to which apprentices in Switzerland and Germany need the same competence to be prepared for the commercial workforce. Consequently, we address the following broad research question:

1. What considerations must be taken into account when attempting to adapt a computer-based test in VET cross-nationally to increase the likelihood of producing a valid measure of the construct 'commercial competence'?

   More specifically, we also address the following questions:
2. To what extent is 'commercial competence' comparable across the German and Swiss VET context?
3. Can a German computer-based test be adapted to measure 'commercial competence' in Switzerland, and if so, how?

An examination of the standards for cross-national test-adaptation is outlined in the following section. Then, the methodological procedures for and findings from curricula analysis, expert interviews, think-alouds with apprentices, and a measurement pilot study are presented. Finally, the challenges of adapting a computer-based instrument that measures competence in VET cross-nationally are discussed.

To facilitate the understanding of the processes of a cross-national computer-based test adaptation, we chronologically narrate the steps taken before proceeding to a complete test adaptation. Furthermore, we provide specific examples for clarity and evidence.

Even though there were no specific plans for applying the original and the adapted test for didactical purposes in VET or cross-national certification, our study provided insights into the adaptation processes involved. More specifically, this paper deals with a specific case, the challenges we faced and the decisions we made are immensely relevant for other research groups who intend to adapt a test. Given the focus on a particular case, the paper concludes with a discussion of generalization and with criteria and guiding questions for future cross-national adaptation projects and research.

Standards and findings for computer-based tests and test adaptations

Standards for tests and test adaptions

Computer-based test components Tanzer and Sim (1999, p. 260) defined the following as the general components that a test should contain: (1) construct definition, (2) specification of the test scope, (3) the instrument itself, (4) administration guidelines, (5) scoring rules and norm tables, and (6) guidelines for score interpretation. Additionally, with regard to documentation, a test must include (7) a test manual, (8) evidence of reliability and validity; (9) a description of the test-development process, (10) any test modifications; (11) description of test-developer(s)' competence; and (12) minimum qualifications that are required to properly administer the test.

However, the adaptation of a computer-based test calls for additional considerations. The International Test Commission (ITC) (2005a) provides test developers, publishers and users with the 'International Guidelines on Computer-Based and Internet Delivered Testing'. Computer-based test stakeholders should address important issues regarding the technology, quality, control, and security of computer-based and Internet testing. For example, test developers should define test hardware and software requirements and adequately document the applied usability testing of the system requirements (ITC 2006, p. 147, 2005a, p. 14). In turn, test users should have a substantial understanding of the technical and administrative test requirements and the system requirements for test takers (ITC 2006, p. 148, 2005a, p. 15). These issues and the criteria for computer-based test components should serve as the basis for planning test adaptation.

Validity evidence for construct *Validity* is the superordinate test characteristic to objectivity and reliability (Hartig et al. 2012, p. 144) and refers to the degree to which empirical and logical evidence and theory support the interpretations of test scores for proposed uses of tests (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) 2014, p. 11). In order to make valid interpretations of test scores, the scope of the test has to be described thoroughly. To do so requires an adequate working definition of the *construct* to be measured (Hartig et al. 2012, p. 147).

With respect to the scope of the test, the construct can be defined either operationally (holistically) or theoretically (analytically). Both cases are applicable to assess competence in VET. In the first case, the test items directly represent, i.e., are sampled from the domain specified by the assessed construct without specific theoretical assumptions (Hartig et al. 2012, pp. 147–152). The aim is defining and measuring the construct (e.g. competence) as closely to real-life performance in a criterion situation as possible (Shavelson 2010; Blömeke et al. 2015, pp. 3–4). Achieving this objective requires studying and sampling tasks from the criterion situation, e.g., the work-life or job situation (Shavelson 2010). The second case begins with establishing or referring to a theory regarding the hypothesized construct. The construct, then, is defined within a theoretical framework, from which test items are generated. Accordingly, the item scores are interpreted as reflecting the unobserved construct (Hartig et al. 2012, pp. 147–152; Blömeke et al. 2015). Consequently, this requires both a detailed theoretical basis for the construct and items that were developed in line with the theoretical considerations and justified with conclusive argumentation linking items to construct definition (Hartig et al. 2012, p. 150). AERA et al. (2014, p. 14) use the term 'evidence based on test content', which, among other things, refers to the adequacy with which a construct is represented by test content (e.g., themes, wording, item format, tasks, and questions). Multiple ways exist to evaluate whether the content is appropriate, for example by asking experts in that occupation to judge the representativeness of the items, by analyzing school training curriculum, and/or by interviewing and observing job incumbents regarding on-the-job demands (AERA et al. 2014, p. 14; Hambleton 2005, p. 7; Sireci et al. 2005; Shavelson 1991). Moreover, especially when using a test for a different purpose than its intended use (e.g., adapting a test), the original content must be examined very carefully to ensure its appropriateness for whether it includes irrelevant content or omits content that may be relevant to the adapting country (cf. Hartig et al. 2012, p. 152; Tanzer 2005; AERA et al. 2014, pp. 14–15).

Furthermore, while a construct is measured on an individual level it should also be analyzed against the context—the system—in which it is embedded (system level). That system includes the organization, norms and learning venues of VET, and the characteristics and relevance of the target group in VET (Achtenhagen and Baethge 2007, p. 66; Baethge et al. 2009, p. 13 et seq.).

Adaptation standards When adapting a test, the ITC (2005b) 'International Guidelines on Test Adaptation' must be considered as they are general standards for test development and adaptation (cf. Moosbrugger and Höfling 2012, pp. 210–211). The guidelines for translating and adapting tests inform test adaptors about what requirements must be met to ensure validity and how differences in test versions can be (statistically) examined. The guidelines are primarily applied in large-scale international assessments of performance such as PISA, TIMSS or tests in the field of psychology (Hambleton 2005, 2001; Moosbrugger and Höfling 2012, pp. 210–211). The 22 most recent guidelines are divided into four categories: (C) context, (D) test development and adaptation, (A) administration and (I) documentation/score interpretation (ITC 2005b, pp. 7–8; Hambleton 2001). An important requirement for test adaptation is found in the second guideline, C.2: 'The amount of overlap in the construct measured by the test or instrument in the populations of interest should be assessed' (ITC 2005b, p. 7). Hambleton (2005, p. 26) explains that the measured construct should be similar in form and frequency.

AERA et al. (2014); Hartig et al. (2012); and Hambleton (2005) provide theoretical implications for adapting a test and validly assessing a given construct. However, we found little *strategic or practical information or criteria* for determining whether a test is 'adaptable' or one has to start afresh. To identify these criteria we turned to other research groups that had adapted tests within the same language and/or had adapted computer-based tests. Reports from these groups yielded valuable information and experience regarding how the validation and adaptation processes might work.

Findings from computer-based instrument adaptation studies

Similar to our study, Lokan and Fleming (2003) attempted to adapt the US-American computer-based career guidance system (System for Interactive Guidance Plus, SIGI Plus) for use in Australia. Although there are surface similarities between Australia and US, the educational pathways and occupational conditions in these countries differ (Lokan and Fleming 2003, p. 167). The adaptation of the occupational terms, spelling and vocabulary was considered a simple and basic task (Lokan and Fleming 2003, p. 171). The 'difficult part' was reviewing the 261 offered occupations to adapt the details (e.g., deleting or renaming the occupations such as 'Career Planning and Placement Director' versus 'Careers Counsellor' [Lokan and Fleming 2003, p. 174)] and making the provided job range, and job descriptions (e.g., salary, workplace conditions) representative of Australia throughout the instrument (Lokan and Fleming 2003, p. 171 et seq.). Further, the technical implementation and evaluation of the changes that were made was exceptionally laborious (Lokan and Fleming 2003, pp. 175–176). The authors came to realize that if they had recognized the level of complexity involved in adaptation, they likely would have opted out of adapting the US-American SIGI Plus (Lokan and Fleming 2003, p. 176).

Fitzgerald (2005) described the challenge of adapting and validating a computer software-skill certification test for cross-national use for Microsoft. This process was performed in four phases. In the Development phase, the English test version was developed (Fitzgerald 2005, p. 199). During the Prelocalization phase, the item drafts were reviewed to anticipate potential problems in the respective countries (Fitzgerald 2005, pp. 199–201). In the Localization phase, the items were translated, and the content was adapted with regard to the cultural and linguistic conditions of the target population (Fitzgerald 2005, pp. 201–207). During the Postlocalization phase, extensive technical reviews were performed, and the exam was delivered (Fitzgerald 2005, pp. 207–212). She found that computer-based tests required much more adaptation effort than other tests, particularly when testing software skills. One challenge in testing software skills is that the tests must be developed quickly because software develops rapidly. The author suggested that tests should be reviewed in their computer-based form to ensure their quality. Such a review includes, for example, the cross-national comparison of software interfaces with the use of screen shots (Fitzgerald 2005, p. 196).

The experiences of Lokan and Fleming (2003) and Fitzgerald (2005) lead us to three main realizations: First, in the labor market, job requirements and their detailed descriptions are important and may vary even when the same label is used. Second, despite similarities in language and job descriptions, there may be differences within the job-situation. Third, for technology-based testing, IT representation and implementation must be authentic. This authenticity requires conscientious monitoring and several feedback loops in the field until adaptation is successfully achieved.

Conclusions for adaptation studies

Validation theory (Hartig et al. 2012; Blömeke et al. 2015), test standards and guidelines [such as those provided by AERA et al. (2014) and ITC (2005b)], and the empirical experience available from Lokan and Fleming (2003) and Fitzgerald (2005) stress that it is essential that we evaluate whether tasks, items, simulated environment and setting are contextually (cross-national context), contentwise (construct equivalence), and technically (IT quality) appropriate when adapting a computer-based test in VET. Therefore a to-be-adapted test must be evaluated against test standards, test adaptation standards and ITC standards criteria before proceeding with the adaptation. Apart from the aforementioned standards and findings, there is still little practical experience with regard to complex adaptation studies on computer-based tests in VET.

Adaptation study

The aim of our adaptation study was to investigate the equivalence of the context and construct in Swiss and German commercial VET, as well as the equivalence of the characteristics of the IT components before adapting ALUSIM. Consequently, our study focused on the content-related validity evidence for the construct 'commercial competence' and not on statistical criteria. We did so because test adaptation requires representative and appropriate test content before collecting statistical evidence. Therefore, we analyzed the equivalence of the German and the Swiss commercial VET context and construct from both an analytical perspective (e.g., by analyzing the theoretical construct assumptions made by test developers) and a holistic (e.g., analyzing commercial work-situations) perspective.

To address the study's demand for content-related validity evidence, we:

1. Described in detail the German context (e.g., the VET system), the German commercial competence construct, and ALUSIM's IT components;
2. Described the Swiss commercial VET system and analyzed it against German VET system;
3. Described the Swiss VET curriculum for commercial apprentices' and analyzed it against the German curriculum;
4. Interviewed commercial sector experts;
5. Conducted think-alouds with Swiss apprentices on selected ALUSIM tasks, and
6. Collected pilot data on the Swiss apprentices' performance on selected and adapted ALUSIM tasks.

Methodologically, for the purpose of examining *context*, we compared the characteristics of the two VET systems (steps 1 and 2). *Construct equivalence* was addressed by ALUSIM's context, construct and IT description, an analysis of the VET curricular artifacts, detailed interviews on job tasks with commercial sector experts, apprentices' think-alouds while performing the simulated tasks, and apprentices' task performance collected in the field (steps 1 and 3–6). The *IT components* of ALUSIM were evaluated with a detailed IT description, apprentices' think-alouds, and apprentices' task performance (steps 1, 4, 5 and 6).

We used this combination of steps, because a single method would be insufficient to adequately address all three criteria. Moreover, every method provides evidence regarding the equivalence of contexts, constructs, and IT components. Table 1 summarizes the combinations of criteria and methods employed.

Step 1. A description of the computer-based test ALUSIM

The first step in adapting the German computer-based test of apprentices' 'commercial competence' for use in Switzerland was to describe the nationally developed and to-be-adapted test and to closely examine its context, construct, and IT components. Therefore, the to-be-adapted test ALUSIM is described in detail below [Table 2; for descriptive criteria, see Tanzer and Sim (1999)].

Context ALUSIM was developed for industrial clerk apprentices in the German dual VET system. The industrial clerk apprenticeship is one of more than 50 formal commercial apprenticeships in Germany (Schapfel-Kaiser and Brötz 2010, p. 28). It is the fifth most chosen among all commercial apprenticeships and is the most favored apprenticeship among graduates that had passed the upper-secondary level[12] (Statistisches Bundesamt (destatis) 2015). German apprentices for industrial clerks complete compulsory education and most have an upper-secondary level qualification (more than 80% of the

[12] Upper-secondary level qualification means here Abitur and is a German High School degree.

Table 1 Criteria and methods (steps) of the adaptation study. Own development

Method (step)	Criterion		
	Context	Construct	IT components
1. Description of the computer-based test	✓	✓	✓
2. VET system analysis	✓		
3. Curricula analysis (school, company)		✓	
4. Expert interviews		✓	✓
5. Think-alouds with apprentices		✓	✓
6. Piloting with apprentices		✓	✓

✓ applicable

participants in the study conducted by Achtenhagen and Winther (2009, p. 46). The training for 'industrial clerk' accounted for 14.7% of all commercial apprenticeships in Germany in 2015.[13]

Industrial clerks are educated and trained at three learning venues. They attend a vocational school 2 days per week, they work at a company 3 days per week, and they participate regularly in specific branch courses during their working time at the companies. The apprenticeship usually takes three years to complete.[14] After the apprenticeship, graduates typically work as industrial clerks in industrial companies and rarely change to a different branch, such as to a bank.

Construct The German commercial competence construct distinguishes domain-linked and domain-specific competences (Winther 2010, p. 31 et seq., 79 et seq., 250 et seq.; Winther and Achtenhagen 2009a). In accordance with Gelman and Greeno (1989), the domain-linked part subsumes generic occupational competences (economic literacy and economic numeracy) (Winther and Achtenhagen 2009a, p. 92). The domain-specific part of competence involves the handling of typical job requirements within a domain (e.g., closing a sale) (Winther and Achtenhagen 2009a, pp. 92–93, 99). This latter competence dimension is reflected in the composition of the computer-based test, as it assesses domain-specific tasks in the commercial area (Winther and Achtenhagen 2009a, p. 99 et seq.).

Construct definition and the process of test development are well documented (e.g., Achtenhagen and Winther 2009; Winther 2010; Winther and Achtenhagen 2009a).[15] The computer-based test predominantly addresses business processes and tasks that are typical for a businessperson or an industrial manager. Winther and Achtenhagen (2009a, p. 98) justified the authenticity, content appropriateness, and practical adaptation of the test by, for example, drawing context and tasks for industrial clerks from an existing company. Therefore, the content of the test is particularly appropriate for typical on-the-job tasks as opposed to tasks found in the school curriculum (Winther 2010, p. 207; Achtenhagen and Winther 2009).

[13] Email information from Statistisches Bundesamt, H204-Berufsbildung, Frau Renth on December 2, 2015.

[14] For more information, see: http://www.bibb.de/govet/de/2361.php; http://berufenet.arbeitsagentur.de (Industriekaufmann) (Accessed 1 July 2016).

[15] Various procedures were used to develop the content and methods of the computer-based test ALUSIM, e.g., a structured content analysis of guidelines, observations, and interviews with employees (Achtenhagen and Winther 2009; Winther 2010, pp. 204 et seq.).

IT components The computer-based test and IT materials were provided on an offline CD-ROM. The IT components were well documented (cf. Table 2). Although ALUSIM was initially intended for use in cross-national VET assessment (Achtenhagen and Winther 2009), its documentation did not address adaptation to other commercial jobs in Germany or to similar jobs in other countries (e.g., content equivalence; reliability for different target groups). ALUSIM was implemented exclusively in one target group (industrial clerks) in Germany. Consequently, adaptation information for other target groups and test components (e.g., a detailed test administrator manual) was not available for ALUSIM and had to be developed during the adaptation process to be used for LINCA. For example, the to-be-adapted computer-based test was designed in a way that the program code contained German names (e.g., city names) and did not permit direct substitution of other names. Therefore, the German names and addresses had to be replaced with Swiss names and addresses, and the program code had to be adapted as well.

This description of ALUSIM's context, construct, test components, and technical documentation (Table 2) served as the basis for comparison with the Swiss VET and was used to judge the cross-national equivalence of the context and construct. From this point

Table 2 Characteristics of the German computer-based test for measuring commercial competence of industrial clerks. Achtenhagen and Winther (2009); Tanzer and Sim (1999); Winther and Achtenhagen (2009a)

Target group of original test	Industrial clerks at the end of their VET Often with upper-secondary level qualification
Test scope	Test of commercial VET in Germany Industrial clerks Realistic and authentic work tasks
Contents	Business processes Task 1: sales Task 2: purchases Task 3: production
Construct definition	Vocational competence in the field of business and administration and in particular domain-specific knowledge and skills
IT components	Computer-based test requiring the following Internet browser System programs (e.g., Microsoft Office, Adobe Acrobat Reader, Adobe Flash Player, Media Player) Hardware for audio and video playback (e.g., headphones) Program code
Test development	A structured content analysis of official guidelines (e.g., vocational training regulations, framework curriculum, educational textbooks) Observations and interviews at workplaces of selected businesses Analyses of students' record books and of interviews Interviews with experts
Relevant test components	Competence of developers ✓ Process of test development ✓ Evidence for reliability/validity ✓ Test modification Test manual/test administrator manual Administration guidelines Scoring rules ✓ Score interpretation Norm tables
Test application	Sample size N = 264 apprentices/N = 7 schools Implementation at one single measurement point at the end of the apprenticeship No test time limit, between 4 and 5 h

✓ available

forward, we took ALUSIM's context and construct as a given and compared it with the Swiss VET context, construct, and IT requirements for purposes of the adaptation study.

Step 2. Commercial VET system analysis

The aim of the next step within the adaptation process was to analyze the to-be-adapted test against the VET system of the target country. To this end, we compared the equivalence of the contexts, namely the German and Swiss VET systems (system level).

Context The target group in Switzerland for measuring 'commercial competence' was commercial apprentices, in which apprentices combine a common formal education in a vocational school (one to two days) with on-the-job training at the company (three to four days),[16] and specific branch courses during their working time at the company. An apprentice may specialize in one of 21 branches (e.g., bank, insurance company, travel agency).[17] There are two aspiration levels in the commercial Swiss VET, namely the practically oriented E-profile (certification) and the more academically oriented M-profile (federal vocational baccalaureate certification).[18] Apprentices in both profiles generally begin a commercial apprenticeship with a degree from a compulsory school between the ages of 15–16 years and the apprenticeship usually takes three years to complete.

In vocational school, the main subject for developing 'commercial competence' is 'Economics and Society' (E&S). Concerning the curriculum, on the one hand, there is a common national curriculum for all VET schools for all 21 branches of commercial apprenticeship. On the other hand, a national company curriculum for each of the 21 branches of commercial apprenticeship defines the branch-specific knowledge and job skills that should be acquired in the companies and branch-specific courses. Among the 21 branches, 'industry' is the fourth largest branch but represented with only 5% of commercial indentures (Bundesamt für Statistik 2015).[19]

After completing the apprenticeship, it is possible to change branches and work as a clerk in any one of the other 20 commercial branches. This structure contrasts with the German VET system with its unalterable context in which the construct 'commercial competence' unfolds. Table 3 provides a comparison of the target groups in the German computer-based test and the Swiss to-be-adapted test. The criteria of comparison for the context include characteristics of the VET systems, which provide, in turn, a framework for analyzing the construct and implementing the computer-based test.

Context equivalence findings The Swiss and German contexts were compared based on criteria including organization, aims, learning venues and branch relevance for VET (Table 3). The German and Swiss VET organizational structures and learning venues appear to be comparable.[20] One main difference is that the Swiss commercial VET prepares apprentices for all the commercial branches within one single apprenticeship pro-

[16] Apprentices in the E-profile attend VET school 1 day and work for 4 days a week in the company in their 2nd and 3rd year of apprenticeship.

[17] All 21 branches are listed here: http://www.kfmv.ch/de/1404/Links-Ausbildungsbranchen.htm. Accessed 7 Jan 2016.

[18] Federal vocational baccalaureate means 'Berufsmatura', which is a higher education entrance qualification.

[19] The three largest branches are 'provision of service and administration' (35.6%), 'public administration' (15.6%) and 'banking' (13.3%) and encompass 64.5% of all commercial jobs.

[20] Rothe (2001) provides a detailed comparison of educational systems in Austria, Germany and Switzerland.

Table 3 Comparison of commercial VET in Germany and Switzerland. ***Germany*** **Kultusministerkonferenz der Länder (KMK) (2002); Hoeckel and Schwartz (2010); Statistisches Bundesamt (destatis) (2015);** ***Switzerland*** **BBT (2011b); Hoeckel et al. (2009); Holtsch and Eberle (2016); Online: http://www.igkg.ch; http://www.kvschweiz.ch; http://www.ehb-schweiz.ch; http://www.skkab.ch/en/documents. Accessed 30 Dec 2015**

	Germany	Switzerland
Organization	Dual system (combination of apprenticeship in a company and vocational education in a VET school) Apprentices are employees of a company Duration of apprenticeship: typically 3 years	
Learning venues	Companies, branch courses, VET schools	
Specific characteristics of education	Branch-specific education in the vocational education school, training company and branch courses	'General' commercial education for all 21 branches in commercial vocational school Branch-specific education in the training company and branch courses
Aim of education	Train apprentices in both school and at the company to work in the industrial branch Commercial work in an industrial company	Train apprentices In school to become economic-civic citizens and In school and at the company to become commercial employees in one of the commercial branches Commercial work in one of the 21 commercial branches For M-profile: reach higher education entrance qualification (e.g., at a university of applied sciences)
Norms (Curricula)	A national company's curriculum for each commercial apprenticeship A national single curriculum for VET school for each of more than 50 specialized commercial apprenticeships	A national company curriculum for each of 21 branches of commercial apprenticeship A common national curriculum for VET school for all 21 branches of commercial apprenticeship
Characteristics target group of the test	Age 18+ Graduates with upper-secondary level qualification	Age 15–16 Graduates of compulsory school
Relevance of commercial apprenticeship in industrial companies	Fifth most-chosen commercial apprenticeship, and the most favored apprenticeship among graduates 14.7% of all commercial apprenticeships in Germany in 2015[a]	'Industry' is the fourth largest commercial branch 5% of commercial indentures

[a] Email information from Statistisches Bundesamt, H204-Berufsbildung, Frau Renth on 2 December 2015

gram. By contrast, the commercial apprentices in Germany learn one specialized occupation out of 50 different commercial occupations. Additionally, Swiss commercial VET schools prepare apprentices to become economic-civic citizens, an aspect not found explicitly in German commercial VET.

While the industrial branch is very important in Germany it is not as important and therefore less representative of commercial apprenticeship in Switzerland. Moreover, the majority of German apprentices for industrial clerks typically enter the apprenticeship with a higher education entrance qualification (upper-secondary level qualification) at an older age—18 or 19 years. By contrast, Swiss apprentices begin their commercial VET at the age of 15–16 years and can reach a higher education entrance qualification only in the M-profile.

In conclusion, the targeted German and Swiss commercial apprentice systems appear comparable due to considerable contextual overlap. However, a main difference does emerge: Switzerland's general commercial branch VET stands in contrast to Germany's branch-specific commercial VET. To determine whether this difference matters when adapting ALUSIM, additional empirical evidence for construct equivalence is required. Although minor adaptations of the computer-based test were expected, construct equivalence should nevertheless be ascertained with the aid of a detailed curricula analysis.

Step 3. Curricula analysis

Curricula constitute one factor that determines apprentices' opportunities to learn and develop competences. The common school's and the branch-specific companies' curricula are important formal sources of competence development because they define the content to be taught and learned, including its depth and breadth at all learning venues. Therefore, the 'commercial competence' construct that is reflected in the German computer-based test must be compared against the Swiss curricula. Another aim of this analysis was to reveal that part of the Swiss content that fosters apprentices' 'commercial competence' as a whole.

Method To assess the Swiss apprentices' opportunities to develop 'commercial competence' in the three main computer-based tasks of (1) sales, (2) purchases, and (3) production, we analyzed the curricula from the Swiss common commercial vocational schools and the branch-specific companies. Because of the three learning venues in Switzerland and Germany the analysis turned out to be more complicated than that of curricula in general education. Hence, we consulted the following Swiss material: (a) 'E&S'-curriculum and 'Information, Communication, and Administration' (ICA)-curriculum for the VET school, and (b) companies' curricula for Commercial Employee with Federal VET Diploma (Bundesamt für Berufsbildung und Technologie (BBT) 2011b).[21]

The analysis was performed in two directions. The first direction was to examine whether the content of ALUSIM could be found in the Swiss curricula *(forward Analysis)*. The guiding question was as follows: To what extent do ALUSIM's three main computer-based test tasks represent the curricula of Swiss commercial education? This part of the analysis sought to determine whether we could link ALUSIM's tasks and corresponding documents to the Swiss vocational school's curricula for E&S and ICA and companies' curricula[22] for the E-profile.[23]

The second direction aimed to gain insight into the construct of 'commercial competence' in Switzerland. Therefore, we also had to ascertain the content that represents 'commercial competence' in the Swiss curricula (*backward analysis*). The guiding question in this part of the analysis was as follows: To what extent was Swiss commercial content represented in ALUSIM? By comparing the content of the E-profile curricula with the content covered in ALUSIM, we sought to determine which facets of 'commercial competence' in the Swiss curricula were not assessed by the original test. As with

[21] Additional analysis of information provided by http://www.igkg.ch; http://www.kvschweiz.ch; http://www.ehb-schweiz.ch. Accessed 30 Dec 2015.

[22] The companies' curricula vary for all 21 branches.

[23] We focus on the documents of the E-profile because the curriculum defines the minimum objectives of the commercial apprenticeship. The M-profile curriculum has similar objectives but at a higher and more detailed level.

the procedure in the forward analysis, the ALUSIM's tasks were analyzed from the perspective of Swiss vocational school's and companies' curricula content. In Table 4, we compared the common compulsory objectives (C) and compulsory-elective objectives (CE) of each branch that are not covered by ALUSIM.

Results regarding construct equivalence The overall analysis of the VET curricula revealed that nearly all of the ALUSIM's tasks could somehow be connected to the curricula of Swiss schools and companies. However, this would not automatically result in the students' ability to perform those tasks successfully because of minor but crucial differences between the Swiss and German curricula.

From a *forward perspective*, subtask (1.2) 'managing customer orders and acting accordingly' in task (1) 'sales' is partly covered by the Swiss school system in the subject ICA. In the first and second semester, they learn how to prepare, answer and organize emails as well as to decide which means of communication (telephone, Internet, email) is suitable for a given situation (BBT 2011b, learning objectives ICA 1.4.1.1, 1.4.1.8). Another example is subtask (3.1) 'calculation of the machine hour rate' in task (3) 'production', which is partially linked to the Swiss VET school curricula in the second semester, when the apprentices learn calculations and accounting in E&S and are trained to prepare calculation tables in ICA (BBT 2011b, learning objectives E&S 1.5.1.2 and ICA 1.4.5.1). However, the term 'machine hour rate' is unfamiliar to Swiss apprentices.

Concerning the Swiss companies' curricula the forward analysis revealed that the main task (1) 'sales' and (2) 'purchase' are partly covered by the companies' objective '1.1 Managing material, data or services'. Furthermore, (1) sales could be linked to '1.3 Carrying out an order', which is a compulsory objective in companies in all six branches. However, the abovementioned objective '1.1 Managing material, data or services' is represented only in two commercial branches and the main task (3) 'production' could not be linked to either the school's or the companies' curricula.

From a *backward perspective,* the objectives of the E&S-curriculum, which are similar for all 21 commercial branches, aim to develop both apprentices' commercial and economic-civic competences (BBT 2011b, learning objective 1.5). However, VET schools develop primarily apprentices' domain-linked knowledge and skills (e.g., knowledge of marketing or market pricing) whereas companies primarily develop apprentices' domain-specific commercial knowledge and skills (BBT 2011b).

Concerning the Swiss companies' curricula, the backward analysis revealed that part of the curricula was not represented at all or not with the same relevance as in ALUSIM. In Switzerland, objectives in the six largest commercial branches are compulsory (C) for advising customers, carrying out an order, and performing administrative and organizational activities.

These objectives cover important job requirements in all six branches but are not represented in ALUSIM. In contrast, managing material, data or services; marketing and PR; human resource administration tasks; and executing financial processes are not compulsory objectives (CE) in the six largest branches (see Table 4).

Although there are overlaps in German and Swiss job requirements, the curricula analysis revealed important deviations between the constructs of 'commercial competence' in Germany and Switzerland. Because of the differing job requirements in Switzerland

Table 4 Companies' curricula objectives among the six largest commercial branches. BBT (2011b); http://www.skkab.ch/en/documents. Accessed 3 Jan 2016, http://skkab.cmsbox.ch/de/leistungsziele/ausbildungs-und-pruefungsbranchen. Accessed 3 Jan 2016

Branch (6 largest)		Provision of ser-vice and admini-stration		Public admini-stration		Banking		Industry		Trusts and estates		Private insurance	
Indicative objectives		**C**	**CE**	**C**	**CE**	**C**	**CE**	**C**	**CE**	**C**	**CE**	**C**	**CE**
1.1	Managing material, data or services	1	2	2	–	–	–	1	4	–	–	–	–
1.2	Advising customers	3	–	4	–	3	–	3	–	3	–	3	–
1.3	Carrying out an order	3	–	8	–	3	–	2	2	2	–	2	–
1.4	Marketing and PR	–	4	1	–	1	–	–	3	–	–	1	1
1.5	Performing human resource administration tasks	–	2	1	2	–	–	–	5	1	–	–	2
1.6	Executing financial processes	1	2	4	–	1	–	–	4	1	–	–	2
1.7	Performing administrative and organizational activities	6	–	7	–	6	–	4	3	6	–	4	6
1.8	Using knowledge gained in own branch and company	4	–	Integrated		5	–	2	1	8	–	11	–

Numbers numbers of objectives

– no objectives

C compulsory objectives

CE compulsory-elective objectives

and Germany, the implementation of ALUSIM would result in measuring only a part of 'commercial competence' in Switzerland. Guided interviews with experts were conducted with the goal of learning more about typical job requirements in Switzerland.

Step 4. Expert interviews

The aim of the on-the-job expert interviews was to evaluate ALUSIM's tasks with regard to the Swiss construct 'commercial competence' and to learn more about the applicability of ALUSIM's tasks to Swiss commercial job requirements.

Method The expert interviews were conducted to evaluate whether ALUSIM's individual tasks could be used in the different Swiss branches. Although the guided interviews were structured, the interviewers gave the experts sufficient latitude and time to add additional points and comments they deemed important for each individual task in the computer-based test (Bortz and Döring 2006, p. 314).[24] Consequently, a few sections of the Achtenhagen and Winther (2009, Appendix II; Winther 2010, pp. 207–209) interview guidelines were modified to increase the suitability of the manual for adaptation and to conform to the Swiss context and conditions. For example, the interviewees were to assess whether the situation presented in ALUSIM was relevant for their branch. The key questions in the original and the adapted versions are presented in Table 5. The interviews were conducted between January 20 and February 1, 2012. Two trained research assistants interviewed all the experts in single interviews. The interviews were recorded, transcribed and analyzed.

Sample The sample was limited to seven experts in the supervision and instruction of apprentices. Every interviewee was required to have some fundamental experience in supervising apprentices and substantial knowledge of all the operations of the branch. The companies were located in the surrounding areas of Zurich.[25] One expert from each of the seven branches with the highest number of apprentices was selected; these branches encompass approximately 85% of all the apprentices.[26] Additionally, the company sizes varied from small to medium to large.

Results regarding construct equivalence The majority of the interviewees thought that, at a minimum, the ALUSIM tasks (1) 'sales' and (2) 'purchase' covered typical commercial requirements in Switzerland. The interviewees evaluated the first sequence of (1) 'sales' as an authentic job requirement (e.g., interviews 1–5, and 7). Moreover, they appreciated the branch-specific transfer of knowledge and skills, as demonstrated in the following example:

> *So, here again I like the fact that the apprentices also need to do things that are a bit more tangible than the tasks in the bank, where money is (only) an abstract concept. I like that very much. (Interview 3)*

[24] 'Guided' or 'Leitfaden' means that the interviewer prepares questions prior to the interview, which gives the data collection a structure that enables comparisons between different participants.

[25] In each case, we contacted the headquarters of the company to ensure the cooperation of the most suitable contact person not only for the interviews but also (potentially), e.g., for future cooperation.

[26] 1st provision of service and administration, 2nd public administration, 3rd banking, 4th industry, 5th private insurance, 6th trust and estates, 7th trading.

Table 5 Interview objectives and questions for developing and adapting the computer-based test. ***Germany*** **Achtenhagen and Winther (2009, Appendix II); Winther (2010, pp. 207–209);** ***Switzerland*** **Eberle et al. (2012b)**

	Interview objectives (selection) by Achtenhagen and Winther (2009)	Additional interview objectives for this adaptation study
A		
Questions describing the workplace	What are three typical tasks in everyday work? What documents are indispensable to the department? What are the work steps of those tasks? What data are received and forwarded? How often are those tasks performed, and how much time do they require?	Same questions as on the left and: Are there additional tasks that apprentices must complete that are not part of their main tasks?
B		
Questions evaluating the job situation (computer-based task)	How realistic do you deem the presented situation is in your everyday working life or in your apprentices' everyday working lives? To what extent must the characteristics of the situation be changed to make them more realistic? What type of situation would represent your daily work more realistically?	Same questions as on the left and: How must the following characteristics be changed to make them more realistic for the Swiss context: opening question, implied work steps, required and useful tools and documents, work results, other matters? In which year of the apprenticeship do you typically teach that situation?

The third task (3) 'production' was generally considered unrealistic and too branch-specific (e.g., interviews 1, 3, 4, and 6). Overall, although the interviewees evaluated the tasks positively, they doubted the representativeness of some subtasks and the entire test tasks' range for commercial job requirements for Switzerland. Furthermore, the interviewees stressed that the tasks mainly focused on table reading and calculation skills (e.g., interviews 1, 2, and 4). The following quote for subtask (2.2) 'comparing suppliers' offers' out of task (2) 'purchase' illustrates this impression:

> *This is possible but doesn't really apply to the learners because commercial apprentices are often not present when we conduct inquiries concerning supplier suitability. A similar example where our apprentices are rather involved is dealing with marketing applications such as adjusting opening hours. This involves gathering justification for why we need to do that and giving precise reasons for doing so. This example goes in the same direction to some extent. However, I cannot remember involving apprentices when dealing with suppliers as suggested by this example. (Interview 7)*

According to the interviewees, ALUSIM did not address a sufficient number of business tasks carried out in Switzerland, thus leading to construct underrepresentation concerning commercial apprenticeship and typical job requirements in Switzerland (e.g., Interview 4). They also suggested that company size might influence job requirements (e.g., Interview 1). Larger companies tend to have a greater division of labor, more specific tasks and customer contact that varies among email, in-person and/or telephone contact. In smaller companies, employees tend to be more involved in multiple tasks and entire processes. Therefore, the interviewees suggested adding additional tasks to ALUSIM, such as event organization (e.g., Interviews 2 and 5), or intensifying tasks focusing on customer communication (e.g., Interviews 1, 2 and 5). Finally, the experts recommended changing the visual design of the documents to comply with Swiss standards, and adding Internet research and presentation tasks. These latter tasks are representative of the tasks required in Switzerland, according to the experts.

Results regarding IT components Unfortunately, it was not possible to provide the interviewees with the computer-based test in its authentic online form. Instead, the experts were provided with printed paper–pencil documents. Hence, they were not able to comment on the functionality and IT implementation of ALUSIM.

In conclusion, the expert interviews confirmed the results of the curricula analysis, namely that two of the three main ALUSIM tasks should be adaptable to Switzerland. However, the task (3) 'production' is embedded in a production department, which is less representative of Swiss commercial activities. The curricula analysis and expert interviews enabled us to identify three subtasks[27] of the ALUSIM tasks—namely, out of (1) 'sales': (1.1) 'processing a customer order' and (1.2) 'writing an email to a customer'; and out of (2) 'purchase': (2.1) 'asking for an offer'—which the apprentices should be able to complete in their first year of apprenticeship. These subtasks were called 'anchor tasks' and were to be piloted in their original form in think-alouds with apprentices. Furthermore, the expert interviews revealed that the addition of tasks concerning customer communication and organization would enhance the content representativeness of the Swiss computer-based test. Nevertheless, these additional tasks would extend beyond the purpose of the adaptation and may be considered only after the adaptation study.

Although we knew how experts evaluate ALUSIM's representativeness and authenticity, we still did not know what thinking processes and performances ALUSIM evoked in apprentices. Moreover, we still had to clarify how ALUSIM could be implemented technically in Switzerland, whether the apprentices considered the test authentic and whether they were able to manage the technical aspects of the test.

Step 5. Think-alouds with apprentices

The first aim of the think-alouds was (a) to determine whether the apprentices were able to successfully complete the anchor tasks, namely (1.1) 'processing a customer order', (1.2) 'writing an email to a customer' and (2.1) 'asking for an offer' and whether they considered the documents and videos authentic. As the original ALUSIM test was performed within 4–5 h (see Table 2), there was no specific time limit for our three anchor tasks. Therefore, another aim of the pilot study was to see (b) how long apprentices needed to carry out the anchor tasks (we expected 1 h or less).[28] A further point of interest was to determine the technical implementation requirements of the computer-based test.

[27] We decided to implement domain-specific subtasks for two reasons: domain-specific tasks are crucial for commercial competence and they represent the core areas of the feasibility study.

[28] For the piloting study and for the first data collection in the main study commercial VET schools provided us with 60 min for implementing the computer-based test. For the think-alouds we provided additional time.

Method The apprentices carried out the three anchor tasks (order, email and asking for an offer) on June 21, 2012. Therein, participants were asked to talk or 'think' out loud while working on the tasks. According to Häder (2015, pp. 402–403) the resulting response taps into participants' mental processes in problem solving. Think-alouds are useful in identifying difficulties in understanding a given task, in understanding instructions, and in producing an appropriate solution.[29] After the think-alouds, we explicitly asked participants to comment on the documents to obtain feedback on cultural and linguistic differences (content and format). Five research assistants documented the apprentices' think-alouds and comments.[30]

Sample Nine commercial apprentices working at a training company in the largest and most representative commercial branch, 'provision of service and administration' (31.0% of all commercial indentures), volunteered to participate. The apprentices were either in their first or third year of training. Eight were female and one was male; five learned in the E-profile, and four learned in the M-profile.

Results regarding construct equivalence The apprentices were only able to perform two of the three anchor tasks, namely (1.1) 'processing a customer order' and (1.2) 'writing an email to a customer' within 60 min. They needed an additional 30 min to perform the anchor task (2.1) 'asking for an offer'.

Table 7 provides the apprentices' results for anchor task (1.1) 'processing a customer order' compared with the results of '6. Piloting with apprentices', which will be addressed in the following section. The results presented in Table 7 demonstrate that the apprentices were able to process a customer order. However, they did not completely understand the test setting as they did not intuitively recognize what to do ('What is my role in this simulation?' Participant 2). The apprentices highlighted the lack of clear instructions on and explanations about the simulation ('What do I have to do? There was no question in the video!' Participant 4). Furthermore, they did not fully understand the purpose of the documents and videos ('Why am I getting this information?' Participant 3). Hence, the participants suggested including an introduction to the computer simulation at the beginning and instructions on how often they were allowed to watch the videos and look at the documents.

As noted above, we selected anchor tasks for the think-alouds that the apprentices could perform in their first year, based on curricula analysis and expert interviews. With regard to evidence based on test content, we therefore expected little feedback on the suitability of the tasks because the apprentices were expected to be familiar with the content and the material. However, the apprentices' comments revealed that some of the terms (e.g., 'sales tax', 'T-Euro', 'EBIT'), formats (e.g., layout of letters, emails and faxes), and details (e.g., currency EURO-CHF, use of 'ß' versus 'ss') had to be adjusted

[29] There are two strategies for the think-aloud method. The first is called 'Concurrent Think-Aloud', in which the participants are asked to report their thoughts and considerations with solving the problem, which can be quite challenging. The second is called 'Retrospective Think-Aloud', in which the participants recapitulate their thoughts after reaching a solution. This strategy is easier for the participants but entails the possibility of the interviewees being too focused on justifying their answer instead of reporting on the process of their decision-making (Häder 2015, pp. 402–403).

[30] The think-alouds were documented with the aid of a video for the entire group and paper–pencil minutes for each participant and each team.

to authentically simulate their commercial work environment. The apprentices also stated that they were used to working with specialized software such as SAP but could also work with Microsoft Office. Apart from this input, the apprentices noted that the original videos were in High German, and the apprentice in the video addressed her supervisors by their last names. In Switzerland, Swiss German is commonly spoken in companies, and supervisors and employees/apprentices call one another other by their first names.

Results regarding IT components There were no technical problems with the offline version of ALUSIM. However, preparing the computers with the offline ALUSIM version and the moderation during the testing was both very time-consuming and challenging. All apprentices were able to work with the computer (Windows), as would naturally be expected of commercial apprentices. Moreover, they were accustomed to working with Microsoft Office software programs. The apprentices found it challenging, however, to handle several documents, although they stated that the captions of the different files were well chosen and helpful. The apprentices would have appreciated extra support (e.g., 'It would be great if we could have two screens.' Participant 1).

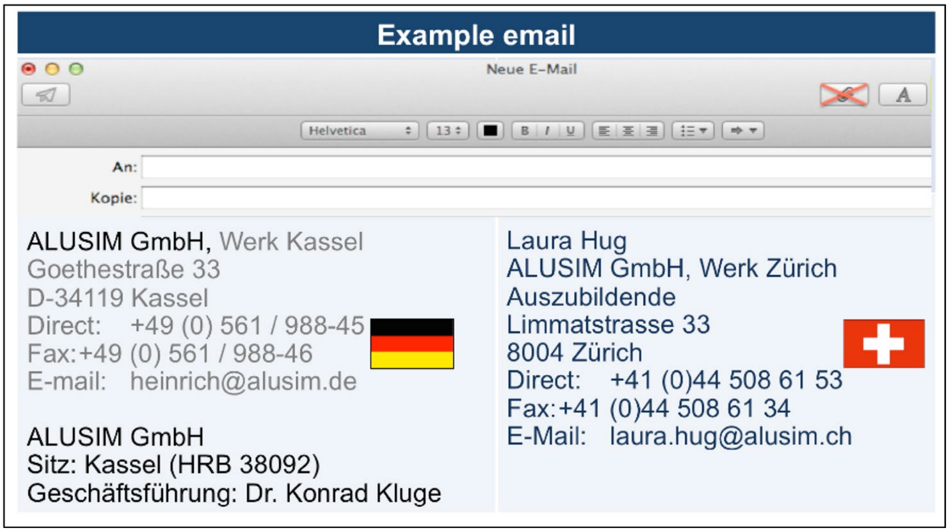
Based on the results of the think-alouds, we concluded that creating a test administrator manual and writing a detailed introduction at the beginning of the test were essential. Another main consideration for further work was to improve technical implementation of the test in an online version for pilot testing with apprentices and the intended main study in 35 VET schools and 85 classes.

As a result of our analysis of the German ALUSIM program package, we decided to completely revise ALUSIM's IT representation (e.g., documents, forms,), response capture (data storage with MySQL of students' written responses) and webpage programming. In response to other think-aloud findings, we adapted contents to Swiss conditions and developed new components for the Swiss simulation (see examples in Fig. 1; Table 6). A video production studio re-produced and synchronized all videos for the anchor tasks from German to (a) Swiss German and (b) Swiss Standard German. These components, other adaptations and the technical changes required a pilot field study with a larger group of apprentices at several VET schools.

Step 6. Piloting with apprentices

The aim of the piloting with apprentices was to obtain additional evidence on the construct equivalence of the three adapted anchor tasks, namely (1.1) 'processing a customer order', (1.2) 'writing an email to a customer' and (2.1) 'asking for an offer'. Further points of interest were to examine the generalizability of the computer simulation's technical functionality and implementation at different VET schools and to determine whether and how it might technically be implemented in an online version.

Method An adapted online version of the computer-based test was implemented for the piloting with apprentices (November 2012). A video in either Swiss German or Swiss Standard German was shown before each anchor task. For example, the videos showed trainers in a company explaining a task to an apprentice. The apprentices were given



An:

Kopie:

ALUSIM GmbH, Werk Kassel
Goethestraße 33
D-34119 Kassel
Direct: +49 (0) 561 / 988-45
Fax:+49 (0) 561 / 988-46
E-mail: heinrich@alusim.de

ALUSIM GmbH
Sitz: Kassel (HRB 38092)
Geschäftsführung: Dr. Konrad Kluge

Laura Hug
ALUSIM GmbH, Werk Zürich
Auszubildende
Limmatstrasse 33
8004 Zürich
Direct: +41 (0)44 508 61 53
Fax:+41 (0)44 508 61 34
E-Mail: laura.hug@alusim.ch

Fig. 1 Example of email adaptation. *Germany* Winther (2010, p. 219); *Switzerland* Eberle et al. (2012a)

Table 6 Examples of term and formal adaptations. *Germany* Achtenhagen and Winther (2009); Winther (2010, pp. 218–219); *Switzerland* Eberle et al. (2012a)

	Germany	Switzerland
Documents	Aktiva Passiva Latente Steueraktiva/Steuerpassiva	Aktiven Passiven Latente Steueraktiven/Steuerpassiven
Names and terms in videos	Frau Fox Fertigungssteuerung A-Kunden	Frau Hug Produktionsabteilung Stammkunden
Address	ALUSIM GmbH,Werk Kassel Herrn Heiner Kolbe Goethestraße 33 D-34119 Kassel	Marc Bucher ALUSIM GmbH, Werk Zürich Produktionsabteilung Limmatstrasse 33 8004 Zürich
Telephone	+49 (0) 561 988-45	0041 (0)31 971 15 20/21
Email	heinrich@alusim.de	bucher@alusim.ch
Form of letters	Sehr geehrte Frau Glüsing *wir bedanken uns …*	Sehr geehrte Frau Gasser Vielen Dank für …

Some examples are cited from unpublished material provided by ALUSIM's authors

60 min to perform the three anchor-tasks.[31] Test administrators observed and documented occurrences (e.g., apprentices' motivation, questions, and IT problems) during the piloting.

To obtain additional information about the construct equivalence of the anchor tasks, the apprentices were given a paper–pencil questionnaire and were asked for their feedback after performing each anchor task. More specifically, the apprentices were asked to

[31] Although the apprentices participating in the think-alouds were not able to perform the anchor tasks within 60 min, we still decided to keep the time limit of 60 min, because commercial VET schools provided us with only 60 min to implement the computer-based test during (a) the piloting study and (b) the first data collection in the main study. Additionally, we expected the apprentices to perform the anchor tasks faster due to the technical and administrative improvements.

respond to the following three questions on a 7-point Likert-type rating scale for each anchor task: How easy/difficult was the task? (1 = very easy to 7 = very difficult); How explicit were the instructions for the task? (1 = very explicit to 7 = very vague) and How authentic was the task? (1 = very authentic to 7 = not authentic at all).

Furthermore, the apprentices reported in an open format how often they had performed these tasks in their training company and how difficult, authentic, and comprehensible the documents (i.e., email, fax, and annual report) in the task were. They could also make further remarks about the computer-based test.

Sample The initial sample included five classes (three E-profile classes with N = 20, N′ = 24, N″ = 24 and two M-profile classes with N = 23, N′ = 22 apprentices) in their first year of training at three different schools for commercial apprentices. However, the computer-based test did not work in one class due to technical problems, such as slow-running computers or non-functioning Internet browsers. In the other four classes, (two E-profile classes N = 20 and N′ = 24 and two M-profile classes N = 23 and N′ = 22), the simulation was successfully implemented. Because of additional problems with data-saving processes during the piloting, we collected data for only 54 of the original 113 apprentices.

Results regarding construct equivalence Table 7 presents the empirical results for items out of anchor task (1.1) 'processing a customer order'. We present the percent of correct responses for items 1, 3, 4 and 5. These items are comparable with the think-aloud results

Table 7 Empirical results for Swiss anchor task (1.1) processing a customer order (% correct)

	Step 5: Think-alouds with apprentices (N = 9)[a]	**Step 6: Pilot study with apprentices (N = 54)**	**Time 1 main study (ALUSIM test) (N = 1756)[b]**	**Time 2 main study (LINCA test) (N = 1365)[c]**
Identify correct customer ID (item 1)	100	63.0	83.8	76.2
Identify correct amount of product/item (item 3)	100	85.2	90.8	91.4
Identify correct article number (item 4)	100	81.5	89.3	91.8
Identify correct customer specific price per item (item 5)	33.3	25.9	52.7	No equivalent item in LINCA

[a] Four apprentices worked alone; the remaining apprentices worked in groups of two or three

[b] After the adaptation study the three adapted anchor-tasks of ALUSIM were implemented in 85 classes in the E-profile and M-profile from November to December 2012 in Switzerland. Although we do not discuss the findings in detail, we present the results as a reference

[c] After finishing the adaptation study LINCA developed and implemented tasks that were similar to ALUSIM but structured for Swiss conditions and embedded in Swiss work-life situations, e.g., 'processing a customer order'. Nevertheless, we can only compare some items and have to be careful with the interpretation. Chi^2 tests for item 1, 3 and 4 between t_1 and t_2 showed no statistically significant differences. Although we do not discuss the findings in detail, we present the results as a reference

(step 5), and with data collected in the first and third round of the main longitudinal study.[32]

In general, the piloting results showed that the anchor tasks were suitable for first-year apprentices. In particular, the apprentices were able to perform the items of the anchor task (1.1) 'processing a customer order' with a range from 25.9% (item 5, Customer specific price per item) to 85.2% (item 3, Amount) correctly. The first-year students performed very well on some items, even better than expected. The 'specialized item price' (item 5) was correctly answered by 25.9% in the pilot study. In the first data collection of the main study, approximately 52.7% of the apprentices answered item 5 correctly. For comparison, 59.8% of 214 German industrial clerk apprentices in their third year answered item 5 correctly (Winther and Achtenhagen 2009b, p. 549).

However, only a selection of three ALUSIM anchor tasks was applicable to first-year apprentices in Switzerland and then adapted. Thus, it appeared unreasonable to compare these data with the original data from ALUSIM. Nevertheless, after finishing the adaptation study, LINCA developed and implemented tasks such as 'processing an order', which contains content similar to ALUSIM but is structured specifically for Swiss conditions. Although the task's items differ, it revealed that this task provides a reliable measure (coefficient alpha. 80).[33]

The feedback on the three closed questions (i.e., task difficulty, explicitness, authenticity) was rather neutral. For anchor task (1.1) 'processing a customer order', the mean 'task difficulty' was 4.3 on a 7-point scale (SD = 1.4), the mean 'explicitness' was 3.5 (SD = 2.0) and the mean 'authenticity' was 3.0 (SD = 1.7). In their open feedback, the apprentices stressed that it was an exciting simulation. However, they also said that they had never performed such tasks before and considered them more branch-specific for industrial companies and unlike a typical duty, for example, in a bank. The apprentices liked the test format, particularly the video prompts. In their opinion, the simulation's content and visualization were rather appealing and authentic. However, they reported that some tasks were too complex for the beginning stages of apprenticeship training. Furthermore, they indicated that they had little or no understanding of some of the terms (e.g., 'CW'[34] and invoice number).

Although the computer-based test had a standardized introduction, the apprentices still required some time to understand what to do. Thus, they lost valuable time, potentially resulting in a failure to complete all the anchor tasks. Other reasons for time loss might have been that they were not used to this type of test format or that we miscalculated the time required because the original test had no time limit.

Results regarding IT components Despite intensive technical preparation, technical problems arose during the piloting. Problems with non-functioning school computers

[32] The adapted computer-based test was implemented in 85 classes in the E-profile and M-profile the first time from November to December 2012 and the third time from January to March 2015.

[33] Two of LINCA computer-based test tasks are similar to ALUSIM's tasks structure but differ regarding setting, number of items ('writing an email to a customer') as well as knowledge tested ('processing a customer order'). Therefore statistical comparisons are not possible.

[34] CW is an abbreviation for calendar week.

and the loss of one class were noted above. Additionally, the data-saving process did not work for some of the apprentices, certain buttons in the program did not work, and the program was not compatible with different browsers. These difficulties led to Internet browsers crashing, poor video quality and slow-running computers and data loss. Thus, intensive efforts were necessary to ensure the technical operating ability of the computer-based test for the main study.

The quantitative piloting was the last step in the adaptation study. The results of each step are summarized and discussed in the following section.

Summary of the results and discussion

The first step in the adaptation study showed that there are indeed similarities between Germany and Switzerland but that an adaptation may also entail difficulties as a result of different commercial work conditions (e.g., job definitions, commercial activities and different branches) and cultural characteristics (e.g., language, interpersonal behavior, and company structures). Thus, a closer analysis of the simulation's operationalization and synchronization was required, especially with respect to its contextualization, construct representativeness, and IT components. Unfortunately, previous studies (e.g., Fitzgerald 2005; Geisinger 1994; Lokan and Fleming 2003) proved to be of limited assistance as they revealed a research and methodological gap concerning the adaptation of computer-based tests, particularly in VET. This gap required the development of a suitable procedure for adapting computer-based tests in VET.

Based on our findings, we were forced to start almost from scratch. We employed a combination of several methods to analyze the context, construct, and IT components of the to-be-adapted computer-simulated test for a new situation. Furthermore, the findings answering the research questions presented below provide further insight into adapting computer-based tests in VET.

1. What considerations must be taken into account when attempting to adapt a computer-based test in VET cross-nationally to increase the likelihood of producing a valid measure of the construct 'commercial competence'?

   We assumed that in commercial VET the competence construct could be authentically represented and operationalized within computer-based tests with high fidelity. This said, research indicates that ensuring the appropriateness of the original test's context, content, and IT requirements are of primary concern when performing an adaptation. The adaptation study, then, had to address the following three criteria context equivalence, construct equivalence and IT feasibility and reliability. Additionally, the adaptation study revealed that a fourth criterion, personnel, time and financial resources, must also be considered.

   In general, if a test has previously been developed in a country for a specific target group, it will mainly apply to the intended construct measurement of that country. When developing an adaptation, it is therefore necessary to compare the constructs to ensure that the construct in the target country is fairly represented. Comparability will also be influenced by *context*—the VET system in which the construct (commercial competence) to be tested is embedded and measured at the individual level. Therefore, clearly described VET systems and precisely defined school and job curricula/job requirements from both countries are required. Accordingly, when assessing *construct* equivalence in VET, it is advisable to provide a sufficient definition of the commercial construct in German and Swiss VET. In this case, a curriculum analysis alone would be insufficient. We also suggest conducting expert interviews, think-alouds with apprentices, and pilot studies with apprentices. Furthermore, observational studies of job performance might be conducted. These additional methods are required because each step contributes to a broad basis for describing and analyzing construct equivalence.

From an *IT perspective*, a well-developed computer-based test in VET should include typical work situations in which apprentices perform typical tasks from their everyday work lives. These tasks could be simulated with, authentic videos, means of communication (e.g., telephone calls, emails), and work documents, to name few possibilities. Consequently, the stimulation and simulation materials in a computer-based test must be detailed and extensive. Therefore, when developing an adaptation of technical components, not only the test tasks but also the entire content and technical setting of the test must be adapted. IT specialists and job experts must collaborate in the development of content-driven test adaptations. Additionally, adaptors must decide whether to provide technical equipment or use the equipment available on-site. The technical implementation may be more complex if on-site equipment is used, e.g., at schools, because of varying software and hardware conditions that must be met.

Personnel, time and financial *resources* required for the adaptation also must be addressed. It is difficult to calculate the extent of the adjustment effort in advance because details concerning the adaptation do not fully emerge until the actual adaptation is in progress. For our adaptation we guesstimated roughly one full-time senior-level position and one full-time research assistant for half a year to conduct the adaptation study and one full-time senior-level position for IT support for 2 months. It would also be helpful to interview test developers for the required test development resources. If we had taken only one task from ALUSIM and conducted a pilot study including appropriate IT components, we would have gained important insights into the adaptation process and details, found pitfalls in advance, and perhaps we would have decided to develop (parts) of the simulation from the beginning. Thus, in cases of uncertainty, it is highly advisable to implement preliminary adaptations in the field before adapting an entire computer-based test.

Table 8 summarizes the questions that should be answered for the intended and the to-be-adapted measurement before performing the adaptation. Moreover, it provides criteria and guiding questions to facilitate the development and validation of an adaptation of a computer-based instrument. The statements that are made provide a useful basis for comparing the context, the construct, the technology-based implementation, and the personnel, time and financial resources of a to-be-adapted test.

Table 8 Description of target and source measurement before conducting an adaptation study. Own development in accordance with ITC (2005b); Tanzer and Sim (1999)

Criteria	Source measurement			Target measurement		
	Main question	Purpose/questions	Methods	Main question	Purpose/questions	Methods
Context						
To what extent is the context equivalent between the source (group) and target (group)?	What is the systemic framework of the intended measurement?	Who is/are the target group(s)? What is context of the source test? How is the VET/educational system organized?	Describing the to-be-adapted test	What is the systemic framework of the intended measurement?	Who is/are the target group(s)? What is context of the source test? How is the VET/educational system organized?	Describing the context of the intended measurement
Construct						
To what extent is the construct equivalent between the source (group) and target (group)?	What do we know about the measured construct of the source test?	What is the aim of the measurement? What construct is measured? How is the construct operationalized? Under what assumptions and with what methods was the computer-based test developed? Which empirical results of the computer-based test are assessable?	Studying publications with regard to test development, results, interpretation Interviews with the test developers Document and tool analyses	What is the construct of the to-be-adapted test?	What is the aim of the measurement? What construct should be measured? How should the construct be operationalized? How should the measurement be implemented, e.g., cross-sectional or longitudinal?	Defining the construct Studying theoretical and empirical results regarding the construct Curricula analysis Interviews with experts and target group Field tests (piloting)

Table 8 continued

Criteria	Source measurement			Target measurement		
	Main question	Purpose/questions	Methods	Main question	Purpose/questions	Methods
IT requirements						
To what extent are IT requirements similar between the countries?	What do we know about the source test, and what documents/tools are available?	To what extent is the technical background of the test known and/or documented? Are the test documents, manuals and tools available in an authentic format, e.g., online? How does the computer-based test work, e.g., in terms of Internet connections, firewalls, storage of data, programming, software updates and upgrades?	Studying publications with regard to test development, results, interpretation Interviews with the test developers Document and tool analyses	What should our test look like/which requirements must be met?	Which constraints must be met, e.g., technical requirements, operating system, software, browsers for the target group?	Documenting details of the planned survey Asking target group for information regarding administrative conditions, e.g., IT conditions at schools Interviews with experts and the target group Field tests (piloting)
Resources						
What personnel, time and financial resources are needed and available?	What do we know about the personnel, time, and financial resources that were used to develop the test?	To what extent are the resources needed documented?	Interviews with the test developers	What personnel, time and financial resources are available for the test adaptation?	Which competences are necessary for adaptation, e.g., with respect to content or technical features? How many people provide their know-how for analyzing adaptation needs, e.g., construct equivalence, IT requirements? Who knows the vocational curricula as well as jargon and company-specific vocabulary and culture? What is the time plan for the adaptation? What is the budget?	Interviews with test developers Estimation with the aid of exemplary adaptations

One might apply the criteria context, construct, IT components, and resources very carefully and thereby increase the likelihood of producing a valid measurement. However, unfortunately, there is no guarantee for it; validation is an ongoing process.

2. To what extent is 'commercial competence' comparable across the German and Swiss VET context?

   The commercial VET system analysis revealed that the contexts for the two countries partly overlap due to the similarity of the VET systems in Germany and Switzerland. For example, there exist overlaps regarding the duration of apprenticeship, learning venues, and major commercial activities. Finally, although the adaptation study revealed surface similarities at the system level, there are cultural and linguistic differences (which might be also true within a single nation) leading to different school objectives and commercial job requirements. Indeed, regional conventions and practices may influence work-life, companies and business processes. Therefore, the constructs vary cross-nationally. In particular, the tasks of the German computer-based test do not precisely cover the same construct of 'commercial competence' in Switzerland. On the one hand, the German construct of 'commercial competence' focuses on commercial activities in industrial companies. For example, one of the computer-based tasks—(3) production—is embedded in a production department, which is not representative of Swiss commercial activities. On the other hand, the Swiss construct of 'commercial competence' covers more commercial activities (e.g., advising customers, performing administrative and organizational tasks) than are implemented in the computer-based test. The aforementioned differences concerning the construct also occur in IT components. For example, videos show typical job situations in a German industrial company, which is less representative of Switzerland.

3. Can a German computer-based test be adapted to measure 'commercial competence' in Switzerland, and if so, how?

   The construct 'commercial competence' was equivalent to only some extent in Germany and Switzerland. There are two possible strategies to cope with this situation that depend on the aim of test implementation. First, if a country focuses on adapting an already existing test from another country to save resources, then an alternative—test development—should be considered. In the end, it might be less costly to build the test from scratch. Second, if the focus is on measuring competence in a cross-national setting, then the commonalities of constructs must be identified. Sometimes, commonalities come down to trivial tasks, and if that is the case, the utility of the original comparative study might be of concern.

   In our case, the latter strategy was applied. The curricula analysis, expert interviews, think-alouds with the apprentices, and piloting with the apprentices revealed commonalities shared by the German and Swiss constructs. The commonalities are represented by only three subtasks of (1) sales and (2) purchase that are suitable for both given constructs. Consequently, those three anchor tasks measure only a part of the commercial competence in both countries. To measure the entire construct in each country, the computer-based test must be complemented by tasks that adequately represent the entire construct of 'commercial competence' in each country. Therefore, to obtain an instrument that adequately represents the commercial competence construct, it would be necessary to develop further tasks for Switzerland that more validly represent and assess ‘commercial competence’. These tasks should cover, for example, advising customers and performing administrative and organizational activities. In contrast, the ‘production’ task, which is not a Swiss job requirement, should be omitted in the computer-based test.

Moreover, the curricula analysis, expert interviews, think-alouds with apprentices, and piloting with apprentices revealed that the anchor tasks must be ‘transcreated’ to the needs of the target country, which presents a substantial challenge. The adaptations of computer-based tasks include modification of terms (e.g., ‘structured convertible bonds’), format (e.g., layout of letters, emails and faxes) and details (e.g., currency EURO-CHF). From a technical perspective, the more complex a test instrument is, the more complex and multifaceted the need for adaptation becomes, particularly for technology-based tests. Every part of the computer-based test, such as videos, stimulation material, and program code, must be adjusted.

Even if a computer-based test was adaptable for the anchor tasks, the administrative conditions in the schools, such as the time slot provided for the computer-based test, still had to be fulfilled. These restrictions may influence the adaptation so strongly that the test results can become incomparable with one another.

Conclusions

The purpose of this paper was to present criteria, guiding questions and steps for performing a cross-national computer-based test adaptation in VET. The ASCOT research initiative, for instance, focused on valid vocational competences’ and skills’ assessment applying innovative technology-based methods.[35] Computer-based instruments developed in ASCOT are destined for adaptation in other countries’ VET. Our study demonstrated that adapting a cross-national computer-based test in commercial VET is possible for a common core of tasks. However, to adequately reflect the diversity of context, construct, and IT components, it is likely that certain tasks will be valid only for a subset of countries. This implies that the participating countries come together to develop the tasks in cooperation.

However, as implied by our feasibility study, adaptations entail challenges regarding the contexts and constructs to be measured that must be addressed. For example, every content and technical detail of the test must simulate the original context as closely as possible to ensure reliable and valid test results. Therefore, technology-based testing in VET and the adaptation of such tests is not only contextually and technically complex but also extremely time-consuming. This is because the same steps must be taken that were taken when the original test was developed to evaluate construct equivalence and ensure that the technical requirements are comparable.

Moreover, the efforts and resources required for an adaptation are easily underestimated. Therefore, it is important to examine the construct equivalence and technical requirements for selected anchor tasks before adapting the entire computer-based test. The decision to perform an adaptation predominantly depends on whether the focus is on international comparisons. If the main reason is to save resources, adaptation may

[35] For more information, see: http://ascot-vet.net/_media/ascot__MASTER_Broschuere_Projekte_EN_V06.pdf. Accessed 25 June 2016.

not prove to be effective because issues regarding construct, technical components and resources do not emerge until the field study. Finally, it may prove less time-consuming and less costly to develop test instruments for core areas in cross-national collaboration from the bottom-up to achieve comparable and reliable results.

The use of similar languages in Switzerland and Germany is peculiar to this adaptation study. However, the adaptation study revealed that although the 'commercial competence' constructs in Germany and Switzerland overlap, they differ in more ways. Although the scope of the adaptation study focused on the 'commercial competence' of apprentices in Germany and Switzerland, the procedures and the consequences of performing an adaptation study are highly relevant for further cross-national adaptations of computer-based tests in VET and/or work life in other countries. This adaptation study raises feasibility questions regarding how a VET-LSA can be conducted and provides important findings for the adaptation of technology-based tests in general.

Tanzer and Sim (1999, p. 268) suggested an enhancement of the ITC guidelines and the development of complementary material, such as a 'compendium of pitfalls and remedies' for test adaptations. Our results regarding context and construct equivalence, IT components, and resources add a practical perspective to the AERA standards and ITC guidelines—and particularly to the context guidelines.

Authors' contributions

DH had the lead in authorship, drafted and developed the manuscript. She carried out and led the adaptation study. SM participated in the development of the adaptation study, the instruments and contributed substantially to implementing the adaptation study. SM participated in the design and structure of the adaptation study. SM formulated parts of the manuscript. EW formulated parts of the methodological sections and translated parts from German into English. DH, SM, EW, FE and RS reviewed the manuscript critically and developed and improved it. DH and RS developed the frame and scope of the paper. All authors read and approved the final manuscript.

Author details

[1] Institute of Education, University of Zurich, Kantonsschulstrasse 3, 8001 Zurich, Switzerland. [2] Stanford University, Stanford, USA.

Acknowledgements

This study was supported and subsidized by SERI, the State Secretariat for Education, Research and Innovation as part of the Leading House project 'Learning and Instruction for Commercial Apprentices' (LINCA). We would also like to thank the experts who participated in our interviews, the apprentices who agreed to participate in our think-aloud interviews and the apprentices who participated in the tests. Special thanks go to Marion Lenggenhager and Sarah Olnhoff, who conducted the interview study and contributed to writing the adaptation reports as well as Sören Vogel, who helped to develop the IT implementation of the simulation. We thank Matthias Baer, Christel Brüggenbrock, Sarah Forster-Heinzer, and Fabio Sticca for their inputs during the writing process of the research paper. We thank the anonymous reviewers for their helpful comments and constructive feedback. Last but not least, we thank Frank Achtenhagen and Esther Winther for their support with the adaptation study.

Competing interests

The authors declare that they have no competing interests.

Received: 4 February 2016 Accepted: 22 November 2016
Published online: 12 December 2016

References

Achtenhagen F, Baethge M (2007) Kompetenzdiagnostik als large-scale-assessment im Bereich der beruflichen Aus-und Weiterbildung. In: Prenzel M, Gogolin I, Krüger HH (eds) Kompetenzdiagnostik, 8th edn. Zeitschrift für Erziehungswissenschaft, Berlin, pp 51–70. doi:10.1007/978-3-531-90865-6_4

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. Georg-August-Universität Göttingen, Seminar für Wirtschaftspädagogik, Göttingen

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

Baethge M, Arends L (2009) Feasibility Study VET-LSA. A comparative analysis of occupational profiles and VET programmes in 8 European countries—International report, vol 8. Vocational Training Research. The Federal Ministry of Education and Research (BMBF) Division for Vocational Training Policy Issues, Bonn

Baethge M, Achtenhagen F, Arends L, Babic E, Baethge-Kinsky V, Weber S (2006a) Berufsbildungs-PISA: Machbarkeitsstudie. Franz Steiner Verlag, Stuttgart

Baethge M, Achtenhagen F, Arends L, Babic E, Baethge-Kinsky V, Weber S (2006b) PISA-VET: a feasibility-study. Franz Steiner Verlag, Stuttgart

Baethge M, Arends L, Winther E (2009) International large-scale assessment on vocational and occupational education and training. In: Oser FK, Renold U, John EG, Winther E, Weber S (eds) VET boost: towards a theory of professional competencies. Essays in Honor of Frank Achtenhagen. Sense Publishers, Rotterdam, pp 3–24

Beaton AE, Mullis IVS, Martin MO, Gonzalez EJ, Kelly DL, Smith TA (1996) Mathematics achievement in the middle school years. IEA's Third International Mathematics and Science Study (TIMSS). Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, Chestnut Hill

Blömeke S, Gustafsson JE, Shavelson RJ (2015) Beyond dichotomies: competence viewed as a continuum. Zeitschrift für Psychologie 223(1):3–13. doi:10.1027/2151-2604/a000194

Bortz J, Döring N (2006) Forschungsmethoden und evaluation—für Human- und Sozialwissenschaftler. Springer, Berlin. doi:10.1007/978-3-540-33306-7

Bundesamt für Berufsbildung und Technologie (BBT) (2011a) Berufsbildung—Ein Schweizer Standort- und Wettbewerbsfaktor. Studie bei multinationalen Unternehmen sowie Expertinnen und Experten in der Schweiz, in Deutschland und Grossbritannien. 31. August 2011. Eidgenössisches Volkswirtschaftsdepartement (EVD) und Bundesamt für Berufsbildung und Technologie (BBT), Bern

Bundesamt für Berufsbildung und Technologie (BBT) (2011b) Bildungsplan Kauffrau/Kaufmann EFZ vom 26. September 2011 für die betrieblich organisierte Grundbildung und Leistungszielkataloge für die Branchen und für die Schulen. BBT, Bern

Bundesamt für Statistik (2015) Berufliche Grundbildung: Basistabellen. http://www.bfs.admin.ch/bfs/portal/de/index/themen/15/04/00/blank/allgemein-_oder_berufsbildung.html. Accessed 29 June 2016

Eberle F, Holtsch D, Mentele S, Lenggenhager M (2012a) Adaptionsprüfung der Testsimulation ALUSIM (D) für die Schweiz ALUSIM (CH). Pilotierungsergebnisse. Unveröffentlicht. Universität Zürich Institut für Erziehungswissenschaften Abteilung Lehrerinnen- und Lehrerbildung Maturitätsschulen, Zürich

Eberle F, Holtsch D, Olnhoff S, Lenggenhager M (2012b) Adaptionsprüfung der Testsimulation ALUSIM (D) für die Schweiz (CH). Unveröffentlicht. Universität Zürich Institut für Erziehungswissenschaften Abteilung Lehrerinnen- und Lehrerbildung Maturitätsschulen, Zürich

Fitzgerald CT (2005) Test adaptation in a large-scale certification program. In: Hambleton RK, Merenda PF, Spielberger CD (eds) Adapting educational and psychological tests for cross-cultural assessment. Lawrence Erlbaum Associates, Mahwah, pp 195–212

Franke A (2014) Arbeitsmarktkompetenzen im sozialen Wandel—berufsspezifische Anforderungen am Beispiel von vier Megatrends. In: Rohlfs C, Harring M, Palentien C (eds) Kompetenz-Bildung. Soziale, emotionale und kommunikative Kompetenzen von Kindern und Jugendlichen. Springer, Wiesbaden, pp 195–221. doi:10.1007/978-3-658-03441-2

Geisinger KF (1994) Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychol Assess 6(4):304–312

Gelman R, Greeno JG (1989) On the nature of competence: principles for understanding in a domain. In: Resnick LB (ed) Knowing, learning and instruction: Essays in honor of Robert Glaser. Lawrence Erlbaum Associates, Hillsdale, pp 125–186

Häder M (2015) Empirische Sozialforschung. Eine Einführung. VS Verlag für Sozialwissenschaften, Wiesbaden. doi:10.1007/978-3-531-19675-6

Hambleton RK (2001) The next generation of the ITC test translation and adaptation guidelines. Eur J Psychol Assess 17(3):164–172

Hambleton RK (2005) Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In: Hambleton RK, Merenda PF, Spielberger CD (eds) Adapting educational and psychological tests for cross-cultural assessment. Lawrence Erlbaum Associates, Mahwah, pp 3–38

Hartig J, Frey A, Jude N (2012) Validität. In: Moosbrugger H, Kelava A (eds) Testtheorie und Fragebogenkonstruktion. Springer, Berlin, pp 143–171. doi:10.1007/978-3-642-20072-4_7

Hoeckel K, Schwartz R (2010) Learning for jobs: OECD Reviews of vocational education and training. Germany. Organisation for Economic Co-operation and Development (OECD), Paris

Hoeckel K, Field S, Grubb WN (2009) Learning for jobs: OECD reviews of vocational education and training. Switzerland. Organisation for Economic Co-operation and Development (OECD), Paris

Holtsch D, Eberle F (2016) Learners' economic competence in Switzerland: conceptual foundations and considerations for measurement. In: Wuttke E, Seifried J, Schumann S (eds) Economic competence and financial literacy of young adults in European countries: status and challenges. Verlag Barbara Budrich, Opladen, pp 101–119. doi:10.3224/978384740602

Kultusministerkonferenz der Länder (KMK) (2002) Rahmenlehrplan für den Ausbildungsberuf Industriekaufmann/Industriekauffrau (Beschluss der Kultusministerkonferenz vom 14.06.2002). KMK, Bonn

Lokan J, Fleming M (2003) Issues in adapting a computer-assisted career guidance system for use in another country. Lang Test 20(2):167–177. doi:10.1191/0265532203lt250oa

Moosbrugger H, Höfling V (2012) Standards für psychologisches Testen. In: Moosbrugger H, Kelava A (eds) Testtheorie und Fragebogenkonstruktion. Springer, Berlin, pp 203–224. doi:10.1007/978-3-642-20072-4_9

Organisation for Economic Co-operation and Development (OECD) (2013) Skilled for life? Key findings from the survey of adult skills. OECD, Paris

Organisation for Economic Co-operation and Development (OECD) (2014) PISA 2012 results in focus. What 15-year-olds know and what they can do with what they know. OECD, Paris

Rothe G (2001) Die Systeme beruflicher Qualifizierung Deutschlands, Österreichs und der Schweiz im Vergleich. Kompendium zur Aus- und Weiterbildung unter Einschluß der Problematik lebensbegleitendes Lernen. Neckar-Verlag, Villingen-Schwenningen

Schapfel-Kaiser F, Brötz R (2010) Gemeinsamkeiten in kaufmännischen Ausbildungsberufen ermitteln. Zwischenergebnisse einer computergestützten Dokumentenanalyse. Berufsbildung in Wissenschaft und Praxis (BWP) 39(4):26–30

Seeber S, Nickolaus R, Winther E, Achtenhagen F, Breuer K, Frank I, Lehmann R, Spöttl G, Straka GA, Walden G, Weiß R, Zöller A (2010) Kompetenzdiagnostik in der Berufsbildung. Begründung und Ausgestaltung eines Forschungsprogramms. Berufsbildung in Wissenschaft und Praxis (BWP) Beilage 39(1):1–15

Shavelson RJ (1991) Evaluating the Quality of performance measures: content representativeness. In: Wigdor AK, Green BF Jr. (eds) Performance assessment for the workplace, vol I. National Academy Press, Washington, D.C., pp 128–140

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

Shavelson RJ (2012) Assessing business-planning competence using the Collegiate Learning Assessment as a prototype. Empir Res Vocat Educ Train 4(1):77–90

Sireci SG, Patsula L, Hambleton RK (2005) Statistical methods for identifying flaws in the test adaptation process. In: Hambleton RK, Merenda PF, Spielberger CD (eds) Adapting educational and psychological tests for cross-cultural assessment. Lawrence Erlbaum Associates, Mahwah, pp 93–115

Statistisches Bundesamt (destatis) (2015) Häufigster Ausbildungsberuf 2014: Kaufmann/Kauffrau im Einzelhandel. Pressemitteilung vom 21. Juli 2015—264/15. Statistisches Bundesamt, Wiesbaden

Tanzer NK (2005) Developing tests for use in multiple languages and cultures: a plea for simultaneous development. In: Hambleton RK, Merenda PF, Spielberger CD (eds) Adapting educational and psychological tests for cross-cultural assessment. Lawrence Erlbaum Associates, Mahwah, pp 235–263

Tanzer NK, Sim CQE (1999) Adapting Instruments for use in multiple languages and cultures: a review of the ITC guidelines for test adaptations. Eur J Psychol Assess 15(3):258–269. doi:10.1027//1015-5759.15.3.258

The International Test Commission (ITC) (2005a) International guidelines on computer-based and internet delivered testing. 1st July, 2005, Version 1.0. http://www.intestcom.org. Accessed 30 June 2016

The International Test Commission (ITC) (2005b) International guidelines on test adaptation. 15th July, 2005, Version 1.0. http://www.intestcom.org. Accessed 30 June 2016

The International Test Commission (ITC) (2006) International guidelines on computer-based and internet-delivered testing. Int J Test 6(2):143–171. doi:10.1207/s15327574ijt0602_4

Winther E (2010) Kompetenzmessung in der beruflichen Bildung. W. Bertelsmann Verlag, Bielefeld

Winther E, Achtenhagen F (2009a) Measurement of vocational competencies—a contribution to an international large-scale assessment on vocational education and training. Empir Res Vocat Educ Train 1:85–108

Winther E, Achtenhagen F (2009b) Skalen und Stufen kaufmännischer Kompetenz. Zeitschrift für Berufs-und Wirtschaftspädagogik 105(4):521–556