Chapter 12
Adapting and Validating the Collegiate Learning Assessment to Measure Generic Academic Skills of Students in Germany: Implications for International Assessment Studies in Higher Education

Olga Zlatkin-Troitschanskaia, Miriam Toepper, Dimitri Molerov, Ramona Buske, Sebastian Brückner, Hans Anand Pant, Sascha Hofmann, and Silvia Hansen-Schirra

Abstract Starting in 2015, a German research team from the program Modeling and Measuring Competencies in Higher Education (KoKoHs), in collaboration with the US Council for Aid to Education (CAE), adapted and validated the Collegiate Learning Assessment (CLA+) for the German language and cultural context to measure generic higher-order cognitive skills of university students and graduates in Germany. In this chapter, the conceptual and methodological background, the framework of the adaptation and validation study, as well as preliminary results are presented. Finally, findings are discussed critically, and future challenges and perspectives are explored.

12.1 Relevance and Background

Globalization, digitalization, and demographic change are current challenges in the societies, labor markets, and educational systems in most member countries of the Organization for Economic Co-operation and Development (OECD). Policy-driven

O. Zlatkin-Troitschanskaia (✉) · M. Toepper · R. Buske · S. Brückner · S. Hofmann
S. Hansen-Schirra
Johannes Gutenberg University, Mainz, Germany
e-mail: lstroitschanskaia@uni-mainz.de; miriam.toepper@uni-mainz.de; buske@uni-mainz.de; brueckner@uni-mainz.de; s.hofmann@uni-mainz.de; hansenss@uni-mainz.de

D. Molerov · H. A. Pant
Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: molerov@hu-berlin.de; hansanand.pant@hu-berlin.de

O. Zlatkin-Troitschanskaia et al. (eds.), *Assessment of Learning Outcomes in Higher Education*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-319-74338-7_12

reform strategies aimed at narrowing down the existing skill gaps between labor market demands and skill levels of students and graduates. The OECD skills strategy and the survey of adult skills in the OECD Program for the International Assessment of Adult Competencies (PIAAC) have gained international attention (OECD 2016). In higher education, prominent reform strategies such as the Bologna reform in Europe have raised questions regarding the individual and societal returns on higher education. There is a growing need for valid performance-based assessments of higher-order skills that can be used with different groups of students from different countries (see also Shavelson et al., Chap. 10 in this volume). One reason for this can be seen in the current internationalization and harmonization trends in higher education systems with regard to the bachelor-master study model, which have resulted in students becoming increasingly mobile between universities in different countries.

Student learning outcomes (SLOs) have been defined in national and international frameworks in order to manage the accreditation of degree courses and institutions in higher education (e.g., European Qualifications Framework (EQR), European Commission 2015, and the German Qualifications Framework (DQR)). At the institutional level, SLOs as the output of higher education have been defined in study program regulations and module descriptions. However, neither the certificates of academic achievement based on SLO specifications that have been established nationally or internationally nor various existing institutional ranking models have been based on suitable, psychometrically sound methods of assessment. On the contrary, grades and certificates are hardly comparable between higher education institutions even at the national or local level (Zlatkin-Troitschanskaia et al. 2017). Hence, national and international comparative assessment studies are becoming more relevant. These developments over the last decade have emphasized the importance of SLO assessments and the demand to measure SLOs in higher education in a valid, reliable, and fair manner (Zlatkin-Troitschanskaia et al. 2016b; see also Coates 2014, Coates, Chap. 1 in this volume).

Challenges specific to higher education such as high international and national diversity of degree courses, study programs, and institutions make developing and implementing SLO assessments in higher education and in particular assessments of students' generic higher-order cognitive skills a highly complex and multidimensional task. In most OECD countries, the importance of twenty-first century generic skills such as critical thinking, problem solving, quantitative and qualitative reasoning, analytical reasoning, information literacy, and digital literacy are recognized (cf. Alexander, Chap. 3 in this volume). Nonetheless, the increasing importance of such skills is undisputed in international educational practice and research. They are supposed to be a high priority for succeeding in knowledge-based economies, addressing judgments, decisions, and challenges in everyday life, and being an engaged citizen of a globalized world and are therefore necessary for individuals' lifelong learning (e.g., OECD 2014; Shavelson et al., Chap. 10 in this volume).

In order to provide a comprehensive overview of the research and developments in the field of competency assessment in higher education, the KoKoHs research team (Zlatkin-Troitschanskaia et al. 2016b) conducted a broad and detailed document analysis, which included systematic literature and database searches and qualitative content analyses from 2010 to 2016. This review presented that grades across institutions are incomparable. The existing assessments are, for the most part, only suitable as higher education admission tests, for gathering data on individual learning opportunities and as subjective measures (Zlatkin-Troitschanskaia et al. 2015, 2017). Overall, the review's results suggested that the relevance of SLO assessments in higher education is continuously increasing, thanks to their potential to be used for multiple purposes and to provide multi-perspective, evidence-based information for diverse stakeholders (see, e.g., Spiel and Schober, Chap. 4 in this volume).

In Germany, the Federal Ministry of Education and Research (BMBF) has established a national research program on "Modeling and Measuring Competencies in Higher Education" (KoKoHs). The first funding phase (2011–2015) involved 24 collaborative projects comprising approximately 70 individual projects conducted by almost 220 researchers, focusing on modeling and measuring domain-specific and generic competencies in higher education.[1] In the next funding phase, which runs between 2016 and 2020, the new KoKoHs program focuses on "Validations and Methodological Innovations." The KoKoHs researchers build on the newly developed models, instruments, and findings, and validate assessments in greater depth according to the Standards of Educational and Psychological Testing ("the Standards," American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) 2014) and expand existing models and assessment instruments to be used in different study domains or for measurement over time. International connectivity and compatibility of assessments have been an important aim of the KoKoHs program as well (e.g., Brückner et al. 2014). Many KoKoHs project teams are eager to discover international best practice models and to adapt and validate more innovative international approaches for use in German higher education (e.g., the WiWiKom project, Zlatkin-Troitschanskaia et al. 2014, Brückner and Zlatkin-Troitschanskaia, Chap. 6 in this volume).

With focus on assessing generic skills, an increase in research efforts can also be observed at the international level (Zlatkin-Troitschanskaia et al. 2016b). The OECD's feasibility study Assessment of Higher Education Learning Outcomes (AHELO) was an initial approach to internationally assess SLOs in higher education (OECD 2013; Tremblay 2013). In addition to measuring domain-specific competencies in engineering and economics, AHELO employed the Collegiate Learning Assessment (CLA) to assess generic skills. Based on the experiences from AHELO, the US Council for Aid to Education (CAE) developed a new test, the CLA+, as a performance assessment which measures students' generic skills at the level of

[1] The outcomes of the KoKoHs research initiative, which also gave the basis for this study, included 40 competency models and more than 100 measuring instruments. The assessments were carried out with altogether more than 50,000 students at more than 220 higher education institutions throughout Germany to gather evidence of their psychometric quality (Zlatkin-Troitschanskaia et al. 2016a, b).

higher education in the United States (CAE 2013). So far, the CLA+ has been adapted for use in Italy and the United Kingdom. Currently, the CAE in cooperation with the OECD has launched a new program, CLA+ International, to further develop and expand the work on an international level (CAE 2015, Zahner and Ciolfi, Chap. 11 in this volume).

A meta-analysis by Zlatkin-Troitschanskaia et al. (2016b) showed that no German-language instruments for assessing performance exist that meet academic requirements for measuring university students' generic higher-order cognitive skills. Therefore, starting in 2015, a German research team from the KoKoHs program collaborated with the CAE to adapt and validate the CLA+ for the German language and cultural context to measure such skills of higher education students and graduates in Germany.

12.2 Aims and Framework of the German Adaptation and Validation Study

12.2.1 Goals

The goal of the German study was to enable the assessment of generic higher-order cognitive skills in Germany by adapting and validating the CLA+ for a German context while also aiming to ensure the international compatibility and comparability of the adapted assessments and results. In this sense, this study seeks to contribute substantially to the development of assessments of and research on university students' generic skills in Germany. The additional research challenge was to carry out the adaptation and validation in a way that the underlying concept and assessment framework of generic higher-order cognitive skills would be aligned with those established in other countries using the CLA+ (so far, the United States, Italy, and the United Kingdom). In all interpretations in the adaptation and validation process, the team aimed for functional equivalence between the German and the US versions (on functional equivalence, see Braun 2006).

To achieve these goals, the German study comprised four major milestones:

1. Translating the US test instrument into German and adapting it to the German culture to obtain a localized German instrument
2. Validating the German instrument comprehensively for use in higher education in Germany according to the Standards (AERA, APA, and NCME 2014)
3. Based on the validation results, exploring the need for further development and adaptation
4. In collaboration with the CAE team and possibly partners in Italy and the United Kingdom, conducting international comparability analyses

Adapting and validating an educational assessment is a complex and multifaceted task. In the German study, in order to ensure that the adapted instrument is of high quality, the translation, adaptation, and validation processes were based on the Test Adaptation Guidelines (TAG) by the International Test Commission (ITC 2016; Coyne 2000; Hambleton 2001) and the Standards (AERA, APA, and NCME 2014). The TAG provide a rough orientation on appropriate framework conditions for adaptations and were specified for this project. The Standards provide general guidance on the validation of (1) test content, (2) response processes, (3) internal test structure, and (4) relations of the assessed construct to other variables (AERA, APA, and NCME 2014). To meet the validity criteria related to the (1) test content of the Standards, which correspond with the content criteria in the TAG, the German team had to ensure that the constructs of generic higher-order cognitive skills were conceptualized and understood in a similar way in Germany and the United States.

To this end, the theoretical concepts and models underlying the CLA+ tests by the CAE in order to validate it for Germany have been explored (Sect. 2.2). The test instruments were then translated and adapted (Sect. 3.1). The validation analyses so far have included curricular analyses, expert panels, and lecturers' online ratings (see Sect. 3.2). In addition, the (2) cognitive requirements and response processes were analyzed using cognitive interviews with students (Sect. 3.3).[2] Overall, in the German study, a systematic adaptation and validation framework were employed to determine whether the adapted assessment enables a valid measurement of generic higher-order cognitive skills among students and graduates in higher education in Germany. The next step will include preliminary comparability analyses with data from other countries (see milestone 4).

12.2.2 Study Framework

The term higher-order cognitive skills is not defined in a uniform way, and diverse conceptualizations and conceptual frameworks can be found in the research literature (e.g., an overview in Liu et al. 2014; Pellegrino and Hilton 2012). For example, Wheeler and Haertel (1993) conceptualized higher-order skills by determining two contexts in which these skills are employed: (a) situations where thought processes are needed for solving problems and making decisions in everyday life and (b) contexts where mental processes can be applied that have to be developed by formal instruction, including processes such as comparing, evaluating, and justifying. For both contexts, being able to employ higher-order skills is perceived as crucial in a knowledge-based society and digital world (see also Alexander, Chap. 3 in this volume). This kind of conceptualization is commonly accepted in international research, and the first context has served as a starting point for international assessment programs (Forster 2004). While the term higher-order skills refers to a very broad range of domains, the CLA+ aims to measure specific aspects (CAE 2013). The CLA+ assessments rubrics and the constructs have been developed to

[2] Further analyses of (3) the internal test structure and (4) relations to other variables will be conducted after the first administration of the test in the field.

holistically assess analytical reasoning and problem solving (Zahner and Ciolfi, Chap. 11 in this volume).

There are many approaches in measuring SLOs in higher education, such as self-report surveys of learning, multiple-choice tests, or short-answer tests (Zlatkin-Troitschanskaia et al. 2016b). However, the underlying concept of higher-order skills refers to real-life decision making and judgment, which should be reflected as closely as possible in the assessment format (Shavelson 2013; Shavelson et al., Chap. 10 in this volume). According to the literature on international studies on cognitive dispositions, such skills should be assessed mainly via complex item formats that present authentic cases with an adequate and meaningful action-oriented situational context from real life (e.g., Shavelson et al. 2015). Various studies recommend the use of different item formats for the assessment of different aspects of higher-order cognitive skills (e.g., Herl et al. 1996; Ruiz-Primo and Shavelson 1996; Snow 1993).

The CLA+ includes different case-based task formats and both complex performance tasks (PT) and selected-response questions (SRQs) administered on a computer. The PT consists of a short frame scenario and an additional document library where further information of varying relevance is presented. To respond, test takers are prompted to use the information and write a text (e.g., a report). The PT is designed to measure three dimensions: problem solving and analysis, writing effectiveness, and writing mechanics. The second task format, the SRQs, also present a situational context and prompt test takers to choose one correct answer from a selection of four to five options. The SRQs items are designed to assess three additional dimensions: scientific and quantitative reasoning, critical reading and evaluation, and the ability to criticize an argument. The length of the test is limited to 60 min for the PT and 30 min for the SRQs (see also Zahner and Ciolfi, Chap. 11 in this volume).

An overview of the project steps is provided in Table 12.1.

Table 12.1 Overview of the German study

Spring 2015	Selection of tasks (PT 1 and 25 SRQs) for the German study
Summer 2015	Workshop with CAE's developers of the CLA+, including scorer training
Summer 2015	Meeting with colleagues from Italian National Agency for the Evaluation of the University and Research Systems (ANVUR)
Summer/autumn 2015	Agreement on translation guidelines between CAE, German team and translation agency cApStAn
Autumn 2015	Translation by cApStAn (PT 1, 25 SRQs, test instructions, scoring guidelines)
Autumn 2015	Review and revisions of translation by German team and first adaptation round for PT1
Autumn 2015	Curricular analyses
Winter 2015/16	Expert workshop I: Group discussion with 10 national experts from different fields of studies
Winter 2015/16	Second adaptation round by German team for PT 1
Winter 2015/16	Expert workshop II: Group discussion with 10 national experts from different fields of studies

(continued)

Table 12.1 (continued)

Winter 2015/16	Third adaptation round by German team for PT1
Winter 2015/16	Translation of PT 2 by cApStAn
Winter 2015/16	Review and revisions of translation by German team and first adaptation round for PT2
Winter 2015/16	Expert workshop III: Group discussions with 3 translation experts
Spring 2016	Second adaptation round for PT 2 by German team
Spring 2016	Meeting with colleagues from UK's Learning Gain Program (representatives from the Centre for Excellence in Learning and Teaching (CELT))
Spring/summer 2016	Ten cognitive interviews with students (PT 1 and 2)
Spring/summer 2016	Localization of PT 2 by the German team
Autumn 2016	Back translation of localized PT 2 and review by CAE
Winter 2016	Online rating by 12 lectures (25 SRQs)
Winter2016/ spring 2017	Ten cognitive interviews with students with localized PT 2
Spring/summer 2017	Further analyses and documentation of results
Summer/autumn 2017	Comparison of the original version from the U.S. and the adapted test versions from the UK, Germany, and Italy
	Exchange of data and further cross-cultural comparative analyses

12.3 Project Overview and Preliminary Results of the Validation

12.3.1 Translation and Adaptation

In addition to the TAG, the Translation, Review, Adjudication, Pretesting, and Documentation (TRAPD) process then followed (Harkness 2003) – a standard process used when adapting international assessments and surveys. TRAPD is a process approach used to ensure that the test is reviewed, revised, and appraised by a variety of experts on its content, methodology, and translation (Harkness 2003, for a discussion of each step, see also the Cross-Cultural Survey Guidelines by Mohler et al. 2016; see also Behr and Shishido 2016).

The CLA+ was translated into German by cApStAn; a translation service provider specialized in the translation and adaptation of international educational and psychological tests.[3] Linguistic supervision, translation reviewing, and quality assurance were provided by team members from the Faculty of Translation Studies, Linguistics, and Cultural Studies at Johannes Gutenberg University Mainz. The

[3] The company had also been involved in the adaptation and linguistic verification of the previous version of the test, the CLA, for various countries in the Assessment of Higher Education Learning Outcomes feasibility study (Tremblay et al. 2012, p. 198; on the general approach, see also Ferrari et al. 2013).

German adaptation and test validation team also had experience in the translation of tests, including over 5 years' worth of prior projects in the areas of business and economics.[4] Thus, team expertise was deemed adequate for the adaptation of assessments (e.g., Arffman 2013; Behr 2012). The steps of the TRAPD process were carried out under time constraints due to practical reasons of research (see Table 12.1). Given the complexity and novelty of the CLA+ assessment, the adaptability and suitability of the test for Germany had to be critically evaluated following each validation step (see Table 12.1). The decision to adapt a second PT came as a result of the expert panels (see Sect. 3.2.1). CApStAn provided the double translation and reconciliation of the assessment, which were subsequently reviewed by the German team in order to ensure a high level of quality of the German test version. The translated materials included two open-ended PTs on topics of health and sports and the 25 SRQs as well as the detailed item scoring guidelines for the CLA+. CApStAn translators based their work on experience and general guidelines from previous projects, for instance, from the adaptation of the AHELO study (AHELO 2011) or the Programme for International Student Assessment by the Organisation for Economic Co-operation and Development (OECD) (PISA 2010). Specific problem-oriented translation guidelines for the CLA+, such as documenting all linguistic and cultural translation problems sentence by sentence, sometimes requiring adaptation, were drafted and agreed upon by cApStAn, CAE, and the German team. They were based on guidelines drafted previously for the Italian adaptation of the same CLA+ tasks, which were designed to facilitate cross-national comparisons between Italy and Germany later on.

The translation process itself varied due to the complexity of the items. In addition to 25 SRQs, 2 PTs were selected that were deemed generally adaptable to a German context. The translatability evaluation was supported by the item-specific translation guidelines. The SRQs only presented minor adaptation challenges. For the 1st PT on health, no major cultural differences were identified, which is why it was first to be adapted. In turn, both the analysis of translatability and expert panels (see Sect. 3.2.1) indicated major cultural differences for the 2nd PT on sports. Various aspects of the baseball scenario would have been unfamiliar to students in Germany or implausible in a German context. However, since the experts had judged the test, in particular the 2nd PT (see Sect. 3.2.1), to be generally relevant for higher education in Germany, the German team decided to explore various adaptation strategies. First, as was the case with the other parts of the CLA+,[5] the 2nd PT was translated by two translators independently, and the preliminary versions were reconciled by a senior translator at cApStAn. Necessary cultural adaptations were documented beforehand and discussed between test validators and translators. The

[4] For example, on the adaptation of the Test of Understanding in College Economics (TUCE) and the Examen General de Egreso de la Lícenciatura (EGEL) in the WiWiKom project, see Brückner et al. (2014).

[5] The scoring guidelines were translated by one translator only, as they would be rephrased by the test validators in Germany in line with the German conceptualization of the construct, as advised by CAE.

initial assignment for the 2nd PT was to preserve the original baseball context and adapt it as little as possible. This translation strategy, discussed in survey translation under the term ask-the-same-question approach (Mohler et al. 2016), aims to alter the original item composition as little as possible in order to preserve psychometric properties (across several languages), but also bears the risk that students might consider the item "foreign" or difficult to understand. The interviewed experts (see Sect. 3.2.1) concerned that German students might have difficulty picturing themselves as part of a group of decision makers in the United States and suggested rather to prompt them to assume the role of foreign advisors as opposed to decision makers in the United States. This adaptation would have affected only a small part of the text, but the consequences for test performance and cross-national comparability would have been difficult to foresee. Instead of selecting one alternative, the German team ultimately decided to test the effects of a nonadapted version against a localized version.

As a consequence, an ask-a-different-question (Mohler et al. 2016) approach was applied to produce a second, fully localized version of the same PT. To this end, previous work materials including the first translation and translation guidelines were used as input, and the entire TRAPD process was reapplied from the start. To control and better document the production conditions of this localized version for subsequent research, this second version was translated and localized entirely at the Faculty of Translation Studies of Mainz University. Starting with the translation of the scoring guidelines, the team identified major lines of reasoning and the supporting dimensions of meaning in the scenario context. Then, an assessment of translatability was carried out, which identified general translation problems, also reflected in cApStAn's specific scoring guidelines. The localization of realistic micro case studies in the PT was particularly challenging and required in-depth research in order to find German equivalents. In this, the scoring guidelines were helpful for preserving the most relevant item aspects. Various alternatives that covered the same dimensions of the domain of sports in Germany and the United States were discussed. Based on the decision to place the popular sport of soccer at the center of the German scenario, the rest of the task was localized, while the overall structure of the item and approximate amount of distractor information were maintained. In addition to the adaptation of the item text, the localization of graphics was also recommended, both for cultural reasons and for matching the information in the text. This work will require further testing in cross-national comparability analyses (e.g., of effects of cross-cultural differences in illustrations, see Solano-Flores et al. 2016).

The localized version is currently being validated for future use in assessment in Germany (see Table 12.1). The localization illustrates the generally interpretative nature of the translation and adaptation process and the need for close cooperation between test developers and translators. Correspondingly, an additional review based on a back translation is being carried out by CAE for further quality assurance. Other quality assurance measures included, for example, terminology management to ensure consistency within and across tasks and proofreading by two professional translators to ensure linguistic quality. Overall, the translation process complied with the highest academic quality standards.

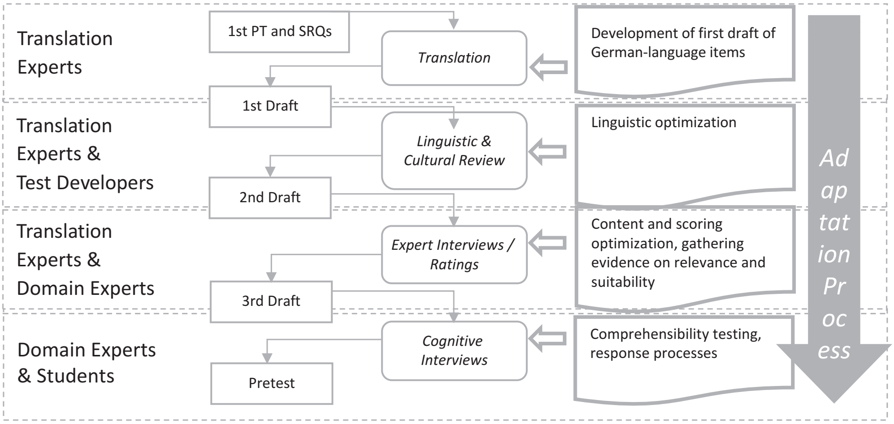
Translation Experts
1st PT and SRQs
Translation
Development of first draft of German-language items
Translation Experts & Test Developers
1st Draft
Linguistic & Cultural Review
Linguistic optimization
2nd Draft
Translation Experts & Domain Experts
Expert Interviews / Ratings
Content and scoring optimization, gathering evidence on relevance and suitability
3rd Draft
Domain Experts & Students
Cognitive Interviews
Comprehensibility testing, response processes
Pretest
Adaptation Process

Fig. 12.1 Translation, adaptation, and validation process for the 1st PT in the German study

The extensive validation procedures (see Sects. 3.2 and 3.3) served to continue to systematically enhance certain aspects of the items. Specific translation problems were discussed in a workshop with experts from various areas. In several feedback sessions, experts reviewed the test and reported shortcomings. The translated versions were revised in further workshops with experts in translation studies (corresponding to the step of Adjudication). Cognitive interviews with students offered indications on whether the items contained any remaining passages that were difficult to understand or unintentionally misleading (see Fig. 12.1). A first version of the CLA+, including 2 PTs and the 25 SRQs, was successfully adapted for use in Germany. Pending successful validation, the localized 2nd PT will be examined further to enhance the quality of adaptations.

12.3.2 Preliminary Findings from the Test Validation

12.3.2.1 Expert Panels

In order to validate the construct underlying the two PTs and SRQs with the validation criterion (1) of the Standards (AERA et al. 2014), three expert workshops were carried out between December 2015 and February 2016 in Mainz and Berlin. The first two workshops aimed to evaluate the partially adapted instrument in terms of its general suitability, content validity, and curricular relevance for use in higher education in Germany. The first panel (December 2015, Mainz) focused on the content-related validation and relevance to the curriculum of the 1st PT *life expectancy* (PT1) and the SRQs. In the second panel (January 2016, Berlin), the construct definition and its operationalization in the two PTs and SRQs were critically discussed with experts on psychometrics and experts on the assessment of higher-order cognitive skills, such as problem solving. In the first and second workshops, respectively, ten experts from various subject areas (including lecturers in biology, business and economics, chemistry, English linguistics and translation studies, higher education research, mathematics, medicine, physics, psychology, and social and political sciences) from different German universities discussed item quality, domain-specificity and generality, challenges of transdisciplinary, cross-institutional, and cross-national testing and comparative analyses, scoring problems, necessary additions, and optional modifications.

The third panel (February 2016, Mainz) focused on the content validation of the 2nd PT *stadium building* (PT2) as well as the evaluation of the translation of both PTs and SRQs. Together with three experts from the field of linguistics and translation studies, the items were discussed with regard to their acceptability and need for further adaptation for Germany to achieve the project aims.

All three expert panels took place in the form of topic-focused, structured group discussions, which were recorded and examined through content analysis. The results of each panel formed the basis for further adaptation and validation (see Table 12.1). For example, findings on the SRQs fed into the subsequent online rating by experts (see below).

First Results

1. *Construct definition.* In all workshops, experts recommended that the construct to be assessed should be defined more clearly for Germany and linked to theory and empirical data. The individual dimensions of the construct should be substantiated a posteriori. During the discussion, culture-specific particularities and cross-national differences in central aspects of the construct and terminology, such as critical thinking and problem solving, became evident; they were attributed to different scientific traditions and a different understanding of academia in the United States compared to Europe or Germany. For instance, German experts were concerned that the two PTs would assess different dimensions of critical thinking. The 1st PT would assess the ability to “deal with (scientific) evidence,” “evidence-based argumentation,” or the “competency to evaluate information,” whereas the 2nd PT would rather assess “problem solving.” For an additional specification of the construct and test definition, it was suggested to link the dimensions examined in the construct to categories of scientific theory or philosophy (e.g., analytical-logical argumentation), in order to specify hypotheses and differentiate the scoring more precisely on this basis.

2. *Further development of the adapted test instrument.* All experts suggested that in order to further develop the instrument and assess important facets of critical thinking, further questions should be added to the tasks. These questions should ask students to evaluate whether they need additional information to solve the task or whether some of the information given was unnecessary and to rate the quality and credibility of the information sources and evidence. All experts pointed out that it would be indispensable to critically examine the extent to which the assessed skills in fact correspond with academically taught competencies. With regard to potential construct-relevant influence factors, the experts identified prior knowledge and skills such as the ability to read diagrams that can determine performance on the

item. Such skills are mainly acquired at school in Germany; hence, greater attention should be given to assessing students' preconditions. Therefore, all experts recommended assessing and controlling for additional individual student characteristics and influence factors in subsequent validation analyses, including controlling for reading comprehension, intelligence, language skills, ability of abstraction, and attitudes or epistemological beliefs.

3. *Relevance to everyday life and sensitivity to study domain*. All experts pointed out the practical relevance of the PTs as particular strength of this assessment. There were, however, critical discussions about the extent to which the instrument assesses generic abilities or rather subject-specific skills acquired in higher education. The experts unanimously pointed out that the instrument might not be suitable for comparisons across disciplines due to the subject-sensitivity of German higher education. It was criticized that students from certain disciplines in Germany, such as "degree courses without an empirical focus" or "arts degree courses," would have a disadvantage in the test, whereas, for example, medical students or students of life sciences were expected to achieve better results on the 1st PT.

4. *Cultural and linguistic comparability of the adapted instrument*. The experts discussed whether the original CLA+ instrument and its adaptations (so far in England, Germany, and Italy, Zahner and Ciolfi, Chap. 11 in this volume) could be generally suitable for an international comparative study. Challenges of linguistic and cultural comparability were identified in particular for the PTs and their scoring. The experts questioned whether the scoring criteria *writing effectiveness* and *writing mechanics* could be compared across countries. While the scenario of the 1st PT was judged to be understandable for German addressees without major adaptations and therefore cross-culturally comparable, the baseball context of the 2nd PT was judged to be much less typical for the German culture. Comprehension and response processes for the 2nd PT were therefore judged to be more difficult than in the US original. Thus, adaptations proved to be inevitable, even though they might negatively impact measurement equivalence across countries. The adapted task would need to be examined more thoroughly regarding its suitability for an international study (see Sect. 3.1).

Micro adaptations of individual aspects or macro adjustments to the entire text were discussed as possible solutions. On the one hand, the original US context could be maintained with only minor changes; however, in order to make the scenario plausible to students in Germany, they would be prompted to assume the role of external, international consultants for another country rather than local decision makers. On the other hand, a localized alternative was deemed suitable for higher education in Germany; this would, however, require a comprehensive change of the scenario context, for instance, from baseball to soccer. As pointed out by the experts, this option would involve risks of altering the psychometric properties of the item, affecting subsequent international comparisons.

5. *Equivalence of different performance tasks and scoring*. With regard to the potential parallel use of both PTs in a field study, the experts compared underlying construct definitions and relations to domains and culture. As noted above, the PTs

were judged to measure different facets of critical thinking and could therefore not be used for comparisons without further analyses. Furthermore, the experts estimated that participants' individual motivation and interests, such as attitudes toward healthy living or particular interests in sports, could confound performance on the items. Possible cultural or gender effects were also expected due to the scenario contexts. According to the experts, the two different PTs allow assessment of different construct facets in higher education, such as a critical approach to sources and evidence, argumentation, or problem solving. Similar questions were raised for the scoring, which was judged problematic when used as a uniform scoring across PTs. Suggestions were made such as giving up the holistic coding scheme, designing more differentiated scoring categories, optimizing the fit between scoring and item instructions in the German version, and using experimental responses from the validation studies to further develop the scoring. The categories could also be defined based on the facets of the German construct definition. In this case, however, international comparability of the scoring might be problematic.

Overall, the expert panels indicated that the CLA+ is an innovative approach to performance assessment that is relevant for higher education practice; the assessment format was judged an interesting and useful addition to current examination practice in Germany. However, experts recognized various challenges to be addressed before the instrument could be used in Germany as well as in an international study, including appropriately adapting the instrument for the higher education context in Germany. This concerns questions of domain-specificity of scenarios and dependence of student performance on prior subject knowledge, which would make it more difficult to use the instrument across disciplines and institutes. It also refers to the equivalence of the construct, dimensions, and facets assessed by the two PTs. Moreover, experts critically discussed the extent to which the test assesses skills acquired in higher education rather than preconditions acquired in upper secondary education in Germany. In accordance with the construct, which needs further differentiation, revisions should be made to the scoring, which should be more closely aligned to the facets of the construct definition and could be developed on the basis of the experimental responses. Further insights to guide necessary modifications were expected from the cognitive interviews, which, according to the experts, were a suitable approach for validating comprehension of and mental response processes to the two adapted PTs.

12.3.2.2 Curricular Analysis and SRQ Rating

In a preliminary curricular analysis, examining whether generic skills in general and the test content of the CLA+ in particular represents part of the curriculum in various fields of studies in higher education in Germany, curricula and module descriptions from 32 different degree courses were analyzed. Overall, the curricular analyses suggested that the adapted item content of the CLA+ is part of curricula in higher education in Germany. In addition, curricular relevance and content validity were supported by the experts' evaluations during the expert workshops and online expert rating, which indicated that these types of skills assessed are being taught at higher education institutions in Germany.

The SRQs were rated by 12 professors and lecturers at higher education institutions in Germany. This expert rating served to cross-validate the curricular analyses and to evaluate additional aspects that were relevant to content validation. The experts rated the curricular relevance and the difficulty of the items and gave a general evaluation of each item. To keep the experts' work within acceptable limits, each of them was asked to rate no more than four items. The questionnaire included closed-ended rating items on a seven-point Likert scale as well as open questions and feedback areas for general concluding remarks.[6] All experts rated both the difficulty and the complexity of the test tasks as appropriate for undergraduate students across the different fields of studies. Additional, the question of whether the test tasks capture central facets of generic skills relevant to the higher education has also been judged as appropriate by the experts. In particular, the experts regarded the relevance of the test facets for the transition to the job market as strong. Overall, content validity was confirmed for all adapted SRQ items from the CLA+. The findings also suggested that the constructs of generic skills were understood in a similar way in different study domains at various universities (for more details, see Kaufmann 2017).

Content validation was interlinked partly with the adaptation (see Sect. 3.1) and was followed by cognitive interviews.

12.3.3 Cognitive Interviews As a Validation Measure

For the validation of the translated, linguistically and culturally adapted PT1, as well as of the translated and linguistically adapted PT2,[7] cognitive interviews were conducted with ten students, drawn by a purposeful sampling (Miles and Huberman 1994) to explore their understanding of the items as well as to identify and analyze mental processes occurring during the response process. The sample included beginner and advanced students, students from different study domains and from different performance levels in order to allow for the observation of possible effects of different domain-specific contexts and learning experiences versus generic skills and attitudes[8] when solving the PT.

[6] For example, "Does the item represent a higher education curriculum or a higher education domain?" "In what ways are constructs likely to differ across German higher education institutions?"

[7] Because of the specific content and context, the cultural adaptation of the PT2 was initially forgone. A cultural adaptation of the PT2 was conducted at Faculty 06 of Mainz University in the summer semester of 2016. Further coglabs have been conducted on both the culturally adapted and the nonculturally adapted version of PT2.

[8] For example, the sample included one student from the domain of medicine who was particularly interested in a healthy lifestyle.

12.3.3.1 Aim of the Cognitive Interviews

Cognitive interviews are used in a multitude of areas in test development and validation. They assess not only formal aspects such as comprehensibility and correctness in the phrasing of tasks but also more complex aspects of a process-related analysis of the mental processes during task-solving in order to derive significant insights about the assessed construct, especially with regard to cognitive validation (Brückner and Pellegrino 2016; Leighton 2013). Another field of application is the linguistic and cultural adaptation of test instruments as well as translation research (e.g., Willis 2005; Fitzgerald et al. 2011; Goerman 2006; on the cognitive validations of CLA tasks in the context of the AHELO study, see Hyytinen et al. 2014).

A cognitive interview study preceded the field application as a pretest, aiming to create functionally equivalent tasks for multiple languages in which CLA+ is used. In cooperation with researchers from different domains (e.g., economists, translation experts, and psychologists), the tasks on life expectancy and the building of a stadium were adapted from the US – American context for the German linguistic and cultural background (see Sect. 3.1). Then, the tasks were assessed in cognitive interviews with regard to their alignment with the understanding of test developers: "These techniques are used to examine whether respondents' interpretations of [self-report] items are consistent with researchers' assumptions and intended meanings given the constructs the items are designed to measure" (Karabenick et al. 2007, p. 139).

The intention to analyze the equivalence between the two tasks and the related mental processes justified by the fact that the tasks were developed from different linguistic and cultural contexts which potentially have a divergent understanding of certain concepts and can therefore present culture-specific peculiarities which may need to be adapted (see Sect. 3.1). An excellent example is the original PT2 from the American context, which is about the building of a baseball stadium. In Germany, however, baseball is not a popular sport; therefore, German students might have more difficulties solving this task, as they can hardly comprehend the cultural and contextual significance of building such a stadium in Germany. Here, the question ensues whether the task should be adapted for the German context in building a new, for example, soccer stadium.

The benefits of the think-aloud methods have been "rediscovered" over the last few years (e.g., Leighton 2013) in order to enable a comparison of measuring instruments from different linguistic and cultural contexts based on mental processes (Goerman 2006). The German study also used this method and embedded it in an assessment design in order to evaluate comparability and to be compatible with the pretest procedures with the CLA+ from previous adaptation processes in other countries (see also Zahner and Ciofi, Chap. 11 in this volume, Solano-Flores et al. n.d.).

12.3.3.2 Preparation and Conduction of the Interviews

Overall, ten students from different degree courses (economics, education, medicine, cultural studies, sociology, politics) were interviewed, six of whom were given the PT on life expectancy, and four were given the PT on the building of the stadium. The interviews were conducted according to a standardized procedure (Solano-Flores et al. n.D.).[9]

Before the beginning of each interview, students were told that the aim was not to test them but to assess the adapted German test versions. As the task documents include a lot of graphs and tables, an intelligence test (IST, Liepmann et al. 2007) with visual tasks was conducted with each student. Then, they were subjected to a short training on thinking aloud. Student could voice potential reservations to receive clarification. After giving a method description, training the students with simple “warm-up” exercises, and asking them to confirm their understanding and ability to think aloud, test coordinators conducted the actual thinking aloud interviews. At the end of each interview, some socio-biographical data were gathered, as well (e.g., degree course, gender, study progress).

Before both the *concurrent* and the *retrospective* interview phase, students were once again explained the purpose of the interview.[10] The characteristic feature of the concurrent phase was that the students worked on the tasks autonomously and without interacting with the test coordinator; the only interaction were reminders to keep talking when they forgot to say their thoughts aloud for a longer period of time (approx. 10 s). During this phase, the interviewer took notes about, for example, how often the student read a certain sentence or passage repeated or underlined words had difficulties with certain terms. In the second phase, the *retrospective phase*, the test coordinator was allowed to ask the students further questions. In addition, similar to cognitive interviews in ANVUR (Solano-Flores et al. n.d.), in this final phase, a standardized interview guideline was used by the test coordinator to ask 10 questions on different aspects of the tasks and the solving process (see Table 12.2).

The data from both phases will then be discussed with the test developers of the US tasks and compared to the data generated from cognitive interviews with the original English instrument. The comparison will allow for a first insight into the response processes in both countries and indicate need for adaptation.

[9] Test coordinators avoided creating a testing atmosphere by seating themselves inclined to the assesse, positioning video recording devices out of sight, and maintaining a disturbance-free environment. In addition, data privacy was observed by filming only the respondents’ hands and multiple test documents.

[10] The note they were read said: “With this interview, we want to investigate how students handle information that they come across in everyday life. For this purpose, we developed a test and we now want to find out whether the tasks that we developed are suitable for use in higher education. It is therefore not the aim of this experiment to measure your expertise; the results will have no influence on your grades whatsoever. We are interested in how students handle the task, how they solve it and what thoughts cross their minds in the process. We would therefore like to ask you to say everything you are thinking out loud while working on the task, even when you have an idea and then end up dismissing it or when you seem to not understand a word! Everything you would say silently to yourself, you should please say out loud. Just imagine you are alone in the room.”

Table 12.2 Standardized questions of the retrospective phase

Coglab questionnaire
Please summarize how you arrived at your solution.
What information did you find to be especially helpful in responding to the item?
Under which circumstances would you have perhaps argued differently?
What did you find especially difficult about the task?
Which materials or information would you have needed in order to solve the task in a satisfactory way?
How did you decide which information is especially relevant for you to solve the task?
Which strategy did you use to respond to the task?
Did you find the task motivating? If yes, why? If no, why not?
How realistic do you consider the situation described in the task?
To what extent do you think the tasks could help you prepare for a professional career?

12.3.3.3 Preliminary Results

With a range from 18 to 29 years the average age of the participants was 23.2 years. Two thirds of the students were female and one third was male students. While the sample showed differences in the family background of the participants – 22.2% indicated that at least one parent originates from another country than Germany and the educational qualification of the parents ranged from a high school diploma to a doctoral degree – all participants stated that the most commonly spoken language in their family environment was German. The sample did also vary regarding the grade on the higher education entrance qualification: a variation from 1.6 to 3.3 could be determined with an average of 2.5.

All students were asked to fill in self-evaluations, which contained four questions. The first two questions concerned the possible disruptions through thinking aloud and the presence of the interviewer. The disruption trough thinking aloud was experienced differently by the students – with a mean score of 2.56 on a scale from 1 (not at all) to 5 (a lot). In comparison, all of the students stated that they were "not at all" (1) or "a little" (2) disrupted by the presence of the test coordinator. The third question asked about the interviewer's expertise regarding critical thinking, which was answered with an average of 4.0 on a scale of 1 (very low) to 5 (very high). Through the last question, concerning the willingness of the students to participate in the study, it was shown that the participants were highly motivated (average 4.22).

The results of the figural and verbal analogies IQ tests conducted with each student revealed large differences between the students – figural test: min. 4 and max. 19 right answers out of 20 tasks; test about analogies: min. 2 and max. 16 right answers out of 20 tasks. While male participants performed better on both IQ tests – figural test: male average 14.67 and female average 10.5; analogy test: male average 13.33 and female average 8.67 – test results also showed correlations with parents' origin and the grade of the higher education entrance qualification.

Further findings indicate that the time of item responding varies between students. Some students needed merely 40 min to solve a PT, while others needed nearly twice as much time. A large part of solving time was spent for studying the provided documents. Typically, a student who solved the tasks in 60 min initially spent nearly 30 min reading and understanding the documents, 2 min for reading the task description, 13 min for rereading the documents and selecting and noting down the most important pieces of information and arguments, and 15 min for finally writing down the answer. Generally, however, all students believed that the target solving time of 60 min should be increased by approximately 20 min.

In terms of content, we observed that many students perceived the topics of the tasks as interesting but were not necessarily motivated to process and solve them. This overlaps with the experiences made by the test developers in the United States, who also reported motivational limitations in item responding. The problem situation described in both tasks was perceived as realistic by many students, even though in the life expectancy task they would have liked to have had more information on the topics of exercise and sleep instead of nutrition and diet. Such information seemed helpful to them as a multifactorial construct. The relations to everyday real life also became evident, as many students perceived the tasks to be useful in preparation for their future professional life. For example, it was pointed out that solving the task helped to use information presented through various media more critically. Furthermore, it was noted that in one's life, both professional and private, one is repeatedly confronted with decisions and that it is therefore helpful to learn to weigh different arguments against one another. However, in order to create an even higher relevance to future professional activities, the students would have liked different, more (domain) specific contents so that the task would specifically prepare them for their professional life.

12.4 Conclusion and Outlook

In this study, we adapted and validated the internationally proven performance assessment CLA+ for Germany, taking into account the underlying conceptual model and assessment framework. For this purpose, we took a multi-perspective and multi-method qualitative approach in examining, among others, the content validity and curricular relevance of the assessment for higher education in Germany as well as the underlying response processes and mental operations. By further in-depth analyses of the think-aloud protocols, we will be able to explore whether item responses of different groups of students were based on different mental processes and representations or different test-taking strategies.

The preliminary results from our validation study showed that this performance assessment enables measuring higher-order cognitive skills at the academic level in higher education. This kind of assessment is innovative for higher education practice in Germany and has significant potential for enhancing curricula and instruction to promote students' interdisciplinary skills. Yet, further research and development are needed in particular with a focus on the concept and test definition. The question as to which concrete skills are assessed with the PTs and SRQs remains unclear and requires further theoretical and empirical research. Another issue lies with the further examination of domain-specificity and the extent to which generic skills can be assessed through specific situational contents and contexts which make reference to certain domains. In other words, the question is whether the same skills can be assessed despite different contents of the tasks.

When implementing this kind of assessment, a number of practical issues arise, such as the question of ensuring test security and test motivation. Our preliminary results show that test motivation is very strongly dependent on the students' interest in the item context, for example, in a healthy lifestyle or a certain sport. Overall, many of the interviewed students would have liked to see a stronger connection to their respective study domains to find the tasks more interesting, which would be problematic with regard to domain-specificity. Should it be possible to assess the same skills using different contents and contexts, it would be possible to let students choose from a pool of tasks. To this end, however, further analyses of the internal test structure are necessary for Germany to empirically prove that all tasks within the item pool assess the same skills and that the test results are comparable. The expert interviews and discussions with professors and lecturers indicated that the implementation of such assessments in higher education practice should be accompanied by corresponding teaching and learning tools. For the United States, CAE has already developed such a tool and reported positive experiences.

To what extent this assessment is suitable for intra- or cross-institutional comparisons remains to be explored in further research. This also holds true for comparisons with other countries. To this end, different adapted versions shall be examined with regard to their measurement equivalence in order to ensure that the adapted tasks measure the same skills and to determine which further adaptations are necessary. Conducting cognitive labs on all adapted versions would be desirable in order to explore whether the same cognitive thought operations are used for responding to adapted versions. Another useful complementation would be to conduct eye-tracking studies in order to control, for example, the effects of general reading abilities, such as reading speed.

References

AHELO Consortium. (2011). *Translation and adaptation manual/guidelines*.

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

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educational Measurement: Issues & Practice, 32*(2), 2–14.

Behr, D. (2012). The team translation approach in questionnaire translation: a special form of expert collaboration. In *Proceedings of the 2nd international specialist conference of the German Federal Association of Interpreters and Translators (BDÜ) 28–30 September 2012* (pp. 644–651). BDÜ, 32. Berlin: BDÜ.

Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 269–287). London: Sage.

Braun, M. (2006). *Funktionale Äquivalenz in interkulturell vergleichenden Umfragen. Mythos und Realität* [Functional equivalence in comparative intercultural surveys: myth and reality.] Mannheim: ZUMA.

Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement, 53*(3), 293–312.

Brückner, S., Zlatkin-Troitschanskaia, O., & Förster, M. (2014). Relevance of adaptation and validation for international comparative research on competencies in higher education – A methodological overview and example from an international comparative project within the KoKoHs research program. In F. Musekamp & G. Spöttl (Eds.), *Competence in higher education and the working environment. National and international approaches for assessing engineering competence*, *Vocational education and training: Research and practice* (Vol. 12, pp. 133–152). Frankfurt am Main: Lang.

Coates, H. (Ed.). (2014). *Higher education learning outcomes assessment – International perspectives*. Frankfurt/Main: Peter Lang.

Council for Aid to Education (CAE). (2013). *Introducing CLA+ Fostering great critical thinkers*. New York: CAE. http://cae.org/images/uploads/pdf/Introduction_to_CLA_Plus.pdf

Council for Aid to Education (CAE). (2015). *The case for generic skills and performance assessment in the United States and international settings*. New York: Council for Aid to Education. http://cae.org/images/uploads/pdf/The_Case_for_Generic_Skills_and_Performance_Assessment.pdf

Coyne, I. (Ed.). (2000). *International test commission test adaptation guidelines*. Accessed 11 December from: www.intestcom.org/test_adaptation

European Commission (EC). (2015). *European qualifications framework*. https://ec.europa.eu/ploteus/search/site?f%5B0%5D=im_field_entity_type%3A97

Ferrari, A., Wayrynen, L., Behr, D., & Zabal, A. (2013). Translation, adaptation, and verification of test and survey materials. In OECD *Technical report of the survey of adult skills (PIAAC) 2013* (pp. 1–28, section 1, chapter 4). http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf. Accessed 14 Jan 2017

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics, 27*(4), 569–599.

Forster, M. (2004). Higher order thinking skills. *Research Developments, 11*, article 1. http://research.acer.edu.au/resdev/vol11/iss11/1

Förster, M., Happ, R., & Molerov, D. (2017). Using the U.S. test of financial literacy in Germany – Adaptation and validation. *The Journal of Economic Education, 48*(2), 123.

Goerman, P. L. (2006). An examination of pretesting methods for multicultural, multilingual surveys: The use of cognitive interviews to test Spanish instruments. In J Harkness (Ed.), *GESIS-ZUMA (Ed.): Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the international workshop on Comparative Survey Design and Implementation (CSDI)*. Mannheim, 2006 (ZUMA-12).

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164–172.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: Wiley.

Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., Dennis, R. A., Klein, D. C. D., Schacter, J., & Baker, E. L. (1996). *Measurement of learning across five areas of cognitive competency: Design of an integrated simulation approach to measurement. Year 1 report*. Los Angeles: University of California.

Hyytinen, H., Holma, K., Toom, A., Shavelson, R. J., & Lindblom-Ylänne, S. (2014). The complex relationship between students' critical thinking and epistemological beliefs in the context of problem solving. *Frontline Learning Research, 2*(5), 1–25.

International Test Commission (ITC). (2016). *The ITC guidelines for translating and adapting tests* (2nd ed.). www.InTestCom.org. Accessed 14 Jan 2017.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, V. B., Blazevksi, J., Bonney, C. R., et al. (2007). Cognitive processing of self report items in educational research: Do they think what we mean? *Educational Psychologist, 42*(3), 139–151.

Kaufmann, F. (2017). *Validierung des Testinhalts eines Kompetenzerfassungsinstruments anhand von Expertenratings*. Unveröffentlichte Masterarbeit.

Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education, 26*(2), 136–157.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* [Intelligence Structure Test] (2., erweiterte und überarbeitete Aufl.). Göttingen: Hogrefe & Huber Publishers.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment* (Research Report). Princeton: ETS.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Mohler, P., Dorer, B., de Jong, J., & Hu, M. (2016). *Translation: Overview. Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Organisation for Economic Co-operation and Development (OECD). (2013). *Assessment of higher education learning outcomes. AHELO feasibility study report – Volume 2. Data analysis and national experiences*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2014). *Education at a glance 2014: OECD indicators*. Paris: OECD Publishing. https://doi.org/10.1787/eag-2014-en

Organisation for Economic Co-operation and Development (OECD). (2016). *Getting skills right: Sweden*. Paris: OECD Publishing. https://doi.org/10.1787/9789264265479-en

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.

PISA Consortium. (2010). *Translation and adaptation guidelines for PISA 2012*. National Project Managers' meeting, Budapest 2010. https://www.oecd.org/pisa/pisaproducts/49273486.pdf. Accessed 14 Jan 2017.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569–600.

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*(2), 73–86.

Shavelson, R. J., Davey, T., Ferrara, S., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton, NJ: Educational Testing Service.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45–60). Hillsdale, NJ: Erlbaum.

Solano-Flores, G., Wang, C., & Shade, C. (2016). International semiotics: Item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. *International Journal of Testing, 16*(3), 205.

Solano-Flores, G., Chia, M., Shavelson, R. J., & Kurpius, A., (n.d.). *CAE cognitive labs guidelines*. Unpublished document by the Council for Aid to Education. New York

Tremblay, K. (2013). OECD assessment of higher education learning outcomes (AHELO): Rationale, challenges and initial insights from the feasibility study. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 113–116). Rotterdam: Sense Publishers.

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes. Feasibility study report. Volume 1 – Design and implementation*. OECD. http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf

Wheeler, P., & Haertel, G. (1993). *Resource handbook on performance assessment and measurement*. Berkeley, CA: The Owl Press.

Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher education learning outcomes assessment – International perspectives* (pp. 175–197). Frankfurt am Main: Peter Lang.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*(3), 393–411.

Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Lautenbach, C., & Toepper, M. (2016a). Assessment practices in higher education and results of the German research program modeling and measuring competencies in higher education (KoKoHs). *Journal Research & Practice in Assessment, 11*, 46–54.

Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M., & Lautenbach, C. (2016b). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen. Ein Überblick zum nationalen und internationalen Forschungsstand* [Assessment of academic competencies of students and graduates – An overview of the national and international state of research]. Wiesbaden: Springer

Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., & Brückner, S. (2017). *Modeling and measuring competencies in higher education. Approaches to challenges in higher education policy and practice*. Wiesbaden: Springer.