The international state of research on measurement of competency in higher education

Olga Zlatkin-Troitschanskaia\textsuperscript{a}, Richard J. Shavelson\textsuperscript{b} & Christiane Kuhn\textsuperscript{a}

\textsuperscript{a} University of Mainz, Business Education, Mainz, Germany
\textsuperscript{b} Stanford University & SK Partners, Graduate School of Education, Stanford, CA, USA

Published online: 10 Mar 2015.
The international state of research on measurement of competency in higher education

Olga Zlatkin-Troitschanskaia\textsuperscript{a*}, Richard J. Shavelson\textsuperscript{b} and Christiane Kuhn\textsuperscript{a}

\textsuperscript{a}University of Mainz, Business Education, Mainz, Germany; \textsuperscript{b}Stanford University & SK Partners, Graduate School of Education, Stanford, CA, USA

With the Program for International Student Assessment and Trends in International Mathematics and Science Study surveys, competency assessment became an important policy instrument in the school sector; only recently has international competency measurement gained attention in higher education with the Assessment of Higher Education Learning Outcomes (AHELO) feasibility study. As AHELO showed, measurement of competencies in higher education is a complex and multidimensional task, which poses great methodological challenges. These challenges arise out of the high diversity of degree courses, study programs, and institutions of higher education. Nevertheless, we need to address these challenges immediately if we aim to create evidence-based high-quality educational systems. This paper overviews the field of international research on competency measurement in higher education. Our analyses revealed a substantial lack of research in this area. Nonetheless, existing studies and assessment practices in various countries provide orientation on how to model competencies based on curricular and professional requirements, design assessments following the assessment triangle, and validate them comprehensively.

Keywords: learning outcome; competency; measurement; assessment; higher education; quality

1. Relevance and objectives

After the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and other assessments of school education received widespread international attention and stimulated changes in schooling and teaching, we might have hoped to see similar empirical research in higher education providing a scientific basis for customized outcome-oriented reforms. However, higher education is still underrepresented in international empirical educational research, and mainly policy-driven changes in higher education, such as the new orientation toward competencies, run far ahead of existing research. Discussions about and early approaches to competency in teaching and learning in higher education, such as the Bologna reform in Europe,\textsuperscript{1} have led to definitions of learning outcomes in national and international frameworks that were mainly meant to guide accreditation of degree courses and institutions in higher education (e.g. Dublin Descriptors, Tuning Educational Structures, degree qualifications frameworks or profiles). Qualifications frameworks for university degrees, such as the European Qualifications Framework
(EQF), offer an approach to combining competencies that students should acquire, that is, knowledge, comprehension, and skills, with other evidence, such as entry requirements and study duration for different degrees. This qualification framework is a step toward internationally comparative competency assessment, but it lacks an empirical foundation and domain-specific definitions. Evidence-based competency models in higher education are still an exception. Competencies have been formally defined in module descriptions, and study and examination regulations. Yet, the actual acquisition of competencies in higher education remains a ‘black box.’ In an analysis of the international state of research on measurement of competency in higher education in 2011, Kuhn and Zlatkin-Troitschanskaia detected a huge lack of empirical studies in competency research in higher education. Previous approaches had mostly been limited to determining prerequisites for admissions tests, gathering data on learning opportunities, and using self-report measures of outcomes.

Challenges specific to higher education such as high international and intra-national diversity of degree courses, study programs, and institutions make research on the assessment of learning outcomes in higher education and especially on the assessment of university students’ domain-specific and generic competencies a highly complex and multidimensional task. Such assessment calls for elaborate assessment methods. Internationally, Assessment of Higher Education Learning Outcomes (AHELO) was an initial approach to structural stabilization in higher education research (e.g. OECD 2010, 2012, 2013a, 2013b; Tremblay 2013). Experiences from AHELO show that the modeling and measurement of competencies in higher education must be carried out according to scientific standards from the fields of psychology, teaching methodology, and statistics. Moreover, assessment research needs to consider, for example, different types of institutions, manifold disciplines, diverse course structures, or different social and economic conditions (Secolsky and Denison 2012).

In this overview, we systematically review key research on assessment of competencies in higher education in order to point out research gaps and challenges, but also to highlight possibilities and further perspectives. Our analysis includes a broad range of studies in line with a holistic understanding of competencies. Accordingly, competency comprises both (latent) cognitive (knowledge and skills) and metacognitive (e.g. self-regulation), as well as non-cognitive (motivational, volitional, affective, and social dispositions) components and observed performance on criterion tasks in varying contexts and situations (see Kulikowich and Alexander 2003). Only the mix and interaction of these components are considered to produce competency in a domain (see Ewell 2005; Rychen 2004; Weinert 2001; Winch and Foreman-Peck 2004).

This overview is based on a broad literature and database analysis. More specifically, we review studies on competency modeling with the goal of providing a source of knowledge for building elaborate assessments of competencies. We take into account two types of approaches: (1) direct assessments of competency in higher education. We focus in particular on large-scale assessments. However, since they have been carried out in only few countries so far, we also consider standardized higher education entrance tests and end-of-studies tests at the bachelor’s and master’s level. (2) Indirect assessments using self-report, such as students’ self-evaluation.

We have sorted assessments according to these two categories and according to geographical origin. We have included internationally visible large-scale projects as well as smaller, local studies that assess competencies in higher education or that have contributed seminal modeling or measurement approaches. We caution that, given the global scope of this overview, there may be great variation in the quality of the presented
We do not aim to make statements or comparisons of the quality of assessments or approaches. This being said, we paint a picture of alternatives to give an overview of the variety of assessments in higher education and relevant research in different countries. Nevertheless, we have tried to include at least some best practices and standards of competency measurement, for example from the USA, in order to highlight potential points of departure for assessments in other countries. We hope this overview will inspire further rigorous research to substantially improve the quality of measurements and instruments or to generate new promising assessment approaches and, ultimately, to improve teaching and learning in colleges and universities.

2. Measurement of competencies in higher education

2.1. Direct assessments

2.1.1. International studies

Although higher education competencies have been assessed in numerous studies in different countries in different ways, there is still little research on large-scale assessment of higher education students’ domain-specific and generic competencies. Apart from AHELO, there are no other international approaches to assessment in higher education that meet the quality criteria of national and international comparative studies from the school sector (e.g. PISA or TIMSS; see e.g. Blömeke et al. 2013; Coates 2014).

The main objective of the OECD’s AHELO study was to design a cross-national and cross-language ‘proof of concept’ for assessing students’ generic and domain-specific learning outcomes from different degree courses, study programs, and institutions of higher education in a valid and internationally comparable way (OECD 2012). The AHELO study is of great interest, as it comprises summative assessment of learning outcomes, with a future goal of estimating the actual added value of university education. The AHELO feasibility study narrowed down the object of study to online tests of generic analytical cogitation and domain-specific skills of economics and engineering sciences. At the same time, AHELO surveyed 23,000 students from 17 OECD countries from all continents on general conditions of student life, such as financial support from parents, structural conditions of studies, cultural characteristics, and interpersonal attitudes, such as satisfaction. All participating countries contributed to the conceptual and methodological design of this comprehensive feasibility study. Such international initiatives are important for current research and have great potential for generating new assessment concepts and methods (see Wolf, Zahner, and Benjamin 2015).

Another international comparative assessment of competencies, the Teacher Education and Development Study in Mathematics (TEDS-M), initiated by the International Association for the Evaluation of Educational Achievement, assessed pre-service mathematics teachers for primary schools and lower secondary schools from 17 countries with regard to their content knowledge, pedagogical content knowledge, pedagogical knowledge, and their beliefs about mathematics teaching and learning (Blömeke, Suhl, and Kaiser 2011). The test was structured according to a theoretical model of action-oriented professional competency based on Shulman (1987) and Weinert (2001). While TEDS-M succeeded in many ways, it was limited to teacher education in mathematics. This international study showed that, as a precondition for
valid assessment of competencies, we need theoretical models of the structure and levels of domain-specific competencies (see also Haertel et al. 2012; Kane 2013; Pellegrino, Chudowsky, and Glaser 2001).

2.1.2. Europe

For Europe, we found few studies of large-scale assessment directly measuring students’ and graduates’ domain-specific and generic competencies. This being said, there are some preliminary studies on certain aspects of competency. For example, Carrillo de la Pena et al. (2009) assessed the academic achievement of students in medicine, psychology, and biology at four Spanish institutions of higher education longitudinally. The pivotal question of the study was to what extent students’ participation in formative assessments, for example, using new feedback instruments, affected the summative outcome, that is, the students’ grades at the end of the term. The study showed that participation in a formative assessment led to better final grades, regardless of how students performed in the summative assessment. Such studies laid the groundwork for further detailed analyses of student learning and competency acquisition in higher education.

During the last decade, a large number of methods for direct assessment have also been developed in the field of teacher education, for example, in the Teacher Education and Development Study: Learning to Teach in Germany (Blömeke et al. 2013). This study focused on pre-service German, English, and mathematics teachers. This quasi-longitudinal study aimed to objectively assess the target group’s content knowledge, pedagogical content knowledge, and pedagogical knowledge on a regional and national level, based on a theoretical competency model (see Shulman 1987; Weinert 2001).

In Germany, a national research program on ‘Modeling and Measuring Competencies in Higher Education’ (KoKoHs) was initiated in 2011. The program comprises 70 projects from Germany and Austria that model and assess domain-specific and generic competencies in various disciplines, including educational sciences, teacher training, psychology, social sciences, engineering, and business and economics. The KoKoHs projects take into account curricular and job-related requirements, transform theoretical competency models into measuring instruments, and validate test score interpretations (the KoKoHs assessment design followed Pellegrino, Chudowsky, and Glaser [2001]; for test validation the project followed the American Educational Research Association’s, American Psychological Association’s and National Council on Measurement in Education’s [AERA, APA and NCME 2014] guidelines). The four KoKoHs projects in this special issue (see Brückner et al. 2015; Bender et al. 2015; Dresel et al. 2015; Musekamp and Pearce 2015) provide examples of competency models and assessment methods, and identify challenges for further assessment research in higher education (Kuhn, Toepper, and Zlatkin-Troitschanskaia 2014; Zlatkin-Troitschanskaia, Kuhn, and Toepper 2014).

In Europe, we find further forms of competency assessment for diagnosing prospective students’ qualifications and predicting their study success. One example is the Swedish Scholastic Aptitude Test. It is a quality assurance program of the higher education system, and it influences the allocation of students to university places (Swedish National Agency for Higher Education 2014).

Another example is the aptitude test for medical studies in Austria (EMS), which is based on the test for medical studies in Germany (Mallinger et al. 2008). The EMS revealed a problem with respect to the composition of aptitude tests and performance
tests in general. The EMS had a gender bias as women performed worse on the test than men, even though more women than men completed their medical studies successfully (Spiel, Schober, and Litzenberger 2008). This highlights that, when developing competency tests, as with all tests, we should minimize construct-irrelevant variance in test items, for example, due to gender bias. This research points to the importance of validation work and minimizing bias in competency assessment.

2.1.3. The USA

In the USA, competency research, although not under the label ‘competency,’ has been mainly undertaken by non-governmental testing organizations. They offer a variety of tests for higher education, which enable standardized, direct assessment of students’ achievement; they fulfill test-quality criteria, such as those in the Standards for Educational and Psychological Testing (AERA, APA and NCME 2014). The three major assessments of higher education achievement outcomes are the American College Testing Program’s (ACT) Collegiate Assessment of Academic Proficiency (CAAP), Educational Testing Service’s (ETS) Proficiency Profile (ETS-PP), and CAE’s CLA (CAE 2014).7 The instruments variously include tests for assessing generic and domain-specific skills and knowledge in a variety of fields. Such tests are used mainly for cross-sectional, but sometimes also for longitudinal, comparisons of student achievement (e.g. Arum and Roska 2010). The ACT offers tests for assessing students at various points in education, for example, when they enter higher education (ACT Compass) or at the end of undergraduate studies (CAAP). They assess generic skills in post-secondary education, such as writing, reading, calculating, but also critical thinking and scientific understanding (see ACT 2015a). The ETS’s Major Field Tests (ETS 2014b) assess domain-specific skills and knowledge, such as factual knowledge, analytical skills, and problem-solving abilities in various fields. Subject-specific tests are available for a number of disciplines.8 There are further tests for assessing generic abilities, such as writing ability, and context-related abilities, such as critical thinking in the humanities (ETS-PP and Critical Thinking Assessment Test, see ETS and the College Board 2014). The testing formats vary among the tests, including multiple-choice formats in the ETS-PP, open-ended formats in the CLA, and combinations of both in the CAAP.

The USA also produces assessment instruments for specific fields. For teachers, there is the Test of Teaching Knowledge, which assesses content knowledge and pedagogical content knowledge among early career teachers of various subjects, including mathematics and sciences (see Council of Chief State School Officers 2010; for biology, see Brownell and Kloer 2015). This criterion-referenced assessment is based on core standards of states in the Interstate New Teacher Assessment and Support Consortium. Other field-specific tests are developed, for example, by the Council for Economic Education, such as the Test of Understanding in College Economics (TUCE), which assesses economics students’ knowledge of microeconomics and macroeconomics cross-sectionally and longitudinally (see Walstad and Rebeck 2008; for a Japanese TUCE Version, see Yamaoka and Asano 2015).

2.1.4. Other countries

In Mexico, institutions of higher education can participate voluntarily in a national assessment. These assessments are mainly based on three types of multiple-choice
tests: (1) The Examen General para el Egreso de la Licenciatura (EGEL) assesses domain-specific competencies of bachelor-level graduates from 39 disciplines. (2) The Examen Nacional de Ingreso al Posgrado (EXANI-III) assesses generic competencies among students who apply for postgraduate courses. (3) The Exámenes Generales para el Egreso del Técnico Superior Universitario (EGETSU) assesses domain-specific and generic competencies as well as general content knowledge among students in 36 disciplines at technical universities (see CENEVAL 2014a, 2014b, 2014c).

There are standardized assessment projects also in South America, Colombia, and Chile that have model systems that assess competencies throughout higher education and, thus, also measure the value added by colleges and universities. For example, Colombia tests for such generic competencies as quantitative reasoning and critical reading at both the beginning and the end of undergraduate studies, thus permitting the estimation of value added (Domingue 2014; Marino 2014; Wiley and Shavelson 2014). Brazil has institutionalized national large-scale assessments of students’ academically acquired skills and knowledge. The assessment is compulsory for students of all institutions of higher education. The Exame Nacional de Cursos assesses domain-specific competencies of all students in 24 disciplines. Furthermore, the Exame Nacional de Desempenho dos Estudantes assesses students’ general content knowledge about social diversity and biodiversity, law and order, and so on, as well as generic skills, such as logical thinking, at the beginning and the end of their studies (see Fundação Cesgranrio 2014; Fundação Cesgranrio and Cespe 2014).

In Australia, institutions of higher education can participate voluntarily in the national Graduate Skills Assessment survey, which is conducted twice a year. The instrument encompasses closed-ended and open-ended items and assesses students’ generic skills, including writing and problem-solving ability, critical thinking, and social skills at the beginning and end of their studies (see GCA 2014a). Further research initiatives are underway to develop new tests (see Coates 2014; Coates and Mahat 2014). The national Tertiary Engineering Capabilities Assessment targets the engineering sciences and was also used in the AHELO study (see Coates and Radloff 2008). The Work Readiness Assessment Package assesses students’ professionalism, including basic competencies, career planning, and work-related reasoning (see Coates and Edwards 2008).

2.1.5. Undergraduate admissions tests

Overall, the majority of large-scale assessments at the undergraduate level focus less on measuring higher education students’ and graduates’ skills and knowledge than on predicting higher education applicants’ prospective college achievement and study success. An instrument that is exemplary for many admissions tests at higher education entry level is the ACT, which covers the four areas of language usage skills in English, reading comprehension, mathematics, and science reasoning. It is used nationally in the USA and also in other countries such as Chile (see ACT 2015a). The Graduate Record Examination (GRE) is used in the USA, Canada, and other countries for assessing generic skills of applicants for graduate schools and studies at the master’s level (see ETS 2014a). The test includes sections on analytical writing, verbal skills, and quantitative skills. The Graduate Management Admission Test is a similar test specifically for applicants of business schools. Furthermore, the Australian uniTEST is an aptitude test that measures secondary education graduates’ generic reasoning and thinking skills (see ACER 2014).
2.2. **Indirect assessment of competencies**

2.2.1. **International and non-European studies**

A substantial number of studies develop or are based on indirect methods of competency assessment. Indirect methods typically rely on respondents’ self-reports of self-perceptions and self-evaluations. Indirect assessments tend to predict academic success less well than direct assessments due to inherent problems. Participants can always overestimate or underestimate their own competencies or can respond untruthfully if they think the true answer would not be socially desirable. Compared to direct assessments, indirect assessments are only to a limited extent suitable for assessing generic and domain-specific knowledge and skills, the cognitive components of competency, which are the main focus of academic teaching and learning. However, indirect methods are typically used for assessing the equally important non-cognitive components of competency, including motivational, volitional, and social dispositions. Self-reports from indirect assessment, then, can complement test results from direct assessment of competency.

Much like direct assessment methods, indirect, mostly questionnaire-based assessment methods are popular internationally. The National Survey of Student Engagement (NSSE) (see NCHEMS 2014) is the largest survey in the USA and is also used in Canada. NSSE surveys first- and final-year students about their attitudes, expectations, and activities with respect to their studies. In Canada, a domain-specific survey (Ministry of Advanced Education, British Columbia 2008) of English as a Second Language gathered data on students’ experiences and learning progress in their courses of study as well as data on further aspects of performance, including their motivation to participate and their financial situation. In Australia, the Course Experience Questionnaire gathers similar data on the output of specific study programs (see GCA 2014b).

In the USA, the Valid Assessment of Learning in Undergraduate Education (VALUE) (see AACU 2009) defined competencies theoretically, taking into account Essential Learning Outcomes for the twenty-first century (LEAP). Competencies were then assessed through computer-based instruments, including electronic portfolios, and external and self-evaluations using multiple expert opinions and students’ self-reflections. More specifically, VALUE assesses competencies with 16 rubrics in the areas of ‘intellectual and practical skills,’ ‘individual and social responsibility,’ and ‘integrative learning.’ These rubrics were operationalized into 16 instruments based on core criteria, distinguishing 4 different levels of competencies. VALUE is intended to help institutions of higher education achieve both individual and systemic goals by enhancing learning processes through formative assessment. This is intended to work by stimulating self-reflection, and by striving for comparability among different higher education systems. VALUE has been used at 3300 institutions of higher education throughout the USA and in other countries.

In the field of teacher education, further national and cross-national studies have been conducted with indirect measures, which all formatively assess the development of pre-service teachers’ key competencies with respect to certain teaching and learning methods and professionalization measures (for the subject of chemistry, see the study on Israel by Mamlok-Naaman et al. 2006 and the study on Turkey by Pitton 2005; for the subjects of foreign languages, see the study on Germany and the USA by Müller-Hartmann 2005).
2.2.2. European studies

The UK is conducting much of the research on indirect assessment methods. For example, the Teaching and Learning Research Program (TLRP) sponsored by the British Economic and Social Research Council includes the Social and Organizational Mediation of University Learning (SOMUL) project (see Brennan and Jary 2005). In the SOMUL project students and graduates of biology, business studies, and sociology from five study programs were interviewed with regard to their study habits and personal and professional characteristics. The comparison of results of the interviews provided evidence of the relation between output of the higher education system and students’ socially and organizationally mediated conceptions of self and of learning (for data on pre-service teachers in two other TLRP projects, see Creech et al. 2008; Smith et al. 2008). Such information could be useful for comparing results among institutions and disciplines on a large scale on the one hand, and for highlighting suitable concepts and methods for the implementation of assessments on the other hand. Apart from these TLRP projects, other projects have been conducted in the UK at individual institutions (e.g. Pokorny and Pokorny 2005) within the discipline of economics.

In Europe, further studies have been conducted on indirect competency assessment in teacher training including studies from Sweden and the Netherlands based on domain-specific competency models that assessed how pre-service science teachers develop competencies of teaching methods (Henze and van Driel 2006; Nilsson 2006). Further country-specific studies have focused on formative assessment of pre-service and in-service teachers’ common teaching and classroom competencies (e.g. for England, see Boyle, Lamprianou, and Boyle 2005; for Switzerland, see Beck et al. 2008). Such studies have also been conducted from a cross-national comparative perspective (for Germany and Switzerland, see Baer, Dörr, and Fraefel 2006: for Germany, Switzerland, Austria, Italy, and Poland, see Frey 2008). All these studies emphasize the crucial importance of formative assessment in teacher education (Carillo de la Pena et al. 2009). Large-scale summative assessments of students’ skills and knowledge are also needed, in order to provide the necessary empirical basis for continuous improvement of educational policy and institutional practice in higher education.

In view of a constantly advancing globalized knowledge society, further studies have focused on learning outcomes as a central quality criterion of higher education. One broad comparative interview study assessed higher education graduates with regard to competencies they had acquired and competencies they were required to have (see Joumady and Ris 2005). In this study, graduates from 209 institutions of higher education in 8 European countries were interviewed three years after graduation with regard to how well university had prepared them for working life. The results indicated that competency assessments in higher education should also take into account interpersonal and professional individual development to gather empirical evidence on the assumed trade-off between general democratic societal and situation-specific requirements. Because the object of study is multidimensional and complex, it seems justifiable to first narrow it down to key cognitive components of competency, such as job-related skills and knowledge.

The large-scale international study, ‘The Flexible Professional in the Knowledge Society: New Demands on Higher Education in Europe’ (REFLEX), is particularly important for the modern knowledge-based society. It assessed flexible generic competencies that graduates had acquired or were required to have (Allen and van der Velden...
A total of 36,000 university graduates from 15, mostly European, countries were interviewed about their self-evaluation of their competencies and how this related to structural factors of higher education and the labor market. The cross-continent EU simulation study (EuroSim) conducted by the Trans-Atlantic Consortium for European Union Studies and Simulations (see Jones 2008) investigated the effectiveness of using simulation methods in teaching and learning. To this end, the study interviewed students from 13 American and 9 European EuroSim partner universities about their learning behavior before and after the introduction of the simulation method. The study identified positive effects on learning behavior.

3. Conclusions and recommendations for further research

3.1. Conclusions

Overall, research on competency measurement in higher education has increased in most OECD countries. Increasing internationalization and globalization in higher education teaching and learning has led to calls for more transparency about what competencies are being taught or acquired. The AHELO is a prime example of increased concern about higher education learning outcomes.

Internationally, then, competency assessment in higher education is becoming increasingly important. This is true for the beginning of studies, where there are university admission tests and assessments of prior knowledge; for the course of studies, where there are formative diagnostic assessments, which sometimes also provide students’ with individual feedback; and for the end of studies or of study phases, where there are summative assessments of learning outcomes.

The broad concept of competency invites an equally broad range of assessment approaches, focusing on students’ or graduates’ knowledge, skills, or motivational, volitional, or social dispositions. We need conceptualization and measurement instruments that are consistent with our broad definition of competency including performance assessments like the CLA along with more traditional measures of latent traits. The approaches have different strengths and limitations and vary in their ability to meet testing standards. Nevertheless, progress is being made in defining and measuring competency, and measures need to be revised or new ones created to keep pace with research findings.

From the presented assessments, we can draw some conclusions about the generalizability and transferability of competency measurements. The internationally most widespread practice in competency assessment in higher education is local assessment of achievement. This includes direct assessments with tests developed specifically for one study program, degree course, or institution. Usually, the test items cannot be transferred without limitations and results are not comparable across degree courses or institutions. While these assessments are not generalizable beyond the local context, they may be useful for locally evaluating the effects of teaching and institutional interventions.

Generalizable achievement assessments have been established in only few countries so far, for example in the USA. Such assessments are developed for administration at many different study programs, degree courses, and institutions and are often intended to enable comparisons, such as CAAP (ACT 2015b), ETS-PP (ETS 2014c), CLA (CAE 2014), and GRE (ETS 2014c). Some, such as the CLA, have been adapted for other countries, as well, for example, in the AHELO study. Other countries in the Americas,
such as Colombia, Brazil, and Mexico, assess all higher education students (voluntary or mandatory) using standardized national assessments.

Research on competency assessment in higher education has focused mainly on cognitive outcomes, that is, on knowledge and cognitive skills, such as problem-solving, verbal, quantitative and analytic reasoning, critical thinking, and so on. This is understandable since they are considered key components of competency that are to be acquired in higher education. However, the modeling and measurement of non-cognitive outcomes is equally complex and still under-researched. Currently, one of the major challenges is to connect results of cognitive learning outcomes to valid measurement of motivational orientations and beliefs (Pascarella and Terenzini 2005). There is little research on motivations and beliefs based on directly observed behavioral data. Usually, non-cognitive outcomes are assessed indirectly through surveys and interviews, which are based on students’ self-reports or evaluations by others, for example, employers. The drawback of such indirect measures is that they are based on individual subjective perceptions, or others’ limited perceptions instead of direct indicators.

Among all reviewed studies and assessments, we found very few with complex designs for collecting and analyzing data, such as longitudinal modeling, multilevel modeling, experimental, or quasi-experimental designs. Hence, internationally, there are only a few empirically sound findings on how competency is acquired and how it develops over time. The reviewed studies and assessments used mainly common types of instruments, such as questionnaires or multiple-choice paper-pencil tests. Internationally, there is a lack of innovative measurement methods, such as adaptive or computer-based testing.

Overall, we found that two kinds of studies and assessments predominated: evaluations of teaching interventions (e.g. of the use of e-learning) and assessments at the transition from secondary to higher education. The latter usually focused on students’ general readiness for undergraduate studies, their readiness for specific courses of study, or influence factors that might predict their study success. However, many of these assessments are not based on objective methods. For example, in many countries, content knowledge is not assessed directly using tests, but only indirectly using grades, degrees, or students’ self-evaluations. We need a broader body of objective, reliable, and valid research and assessments that specifically measure competencies in higher education.

3.2. Recommendations for further research on competency measurement

On the basis of our review and the gaps in both the theory and practice of competency measurement, we make the following recommendations for further research. The recommendations are sorted according to criteria for valid competency assessment (AERA, APA and NCME 2014; Clauser and Hambleton 2012; Haertel et al. 2012; Hambleton, Merenda, and Spielberger 2005; Pellegrino, Chudowsky, and Glaser 2001; Shavelson 2013). Generally, the design of competency assessments should follow the ‘assessment triangle’ (Pellegrino, Chudowsky, and Glaser 2001). Competency assessments, then, should include ‘a defining of model of student cognition and learning in the domain, a set of beliefs about the kinds of observations that will provide evidence of students’ competencies, and an interpretation process for making sense of the evidence’ (44). The three vertices of the triangle – cognition, observation, and interpretation – are central to all assessments and marked three major
activities in competency measurement: (1) defining the constructs to be assessed, (2) developing suitable models and measuring items/tasks for eliciting competencies, and (3) drawing valid inferences from the response data.

3.2.1. Theoretical conceptualization and development of competency models

Theoretical models are indispensable for valid assessment. First, the objects of study of higher education, students’ and graduates’ domain-specific and generic competencies, need to be specified and defined clearly. In the beginning, it may be useful to limit the assessment of competencies to the cognitive components of procedural skills and conceptual knowledge. This limitation makes sense if we assume that universities are generally meant to teach domain-specific skills and knowledge that form academically taught professional competencies. While international research projects show that generic competencies are indeed highly relevant (see the CAAP or CLA tests from the USA), they are not significant enough on their own to constitute professional competencies. These assessments need to be complemented by assessments of key domain-specific skills and knowledge.

Further research should focus on the theoretical modeling of academically acquired generic skills and knowledge, for example, analytical reasoning, critical thinking and communication, and domain-specific skills and knowledge. Domain-specific competency models should be structured and specified in a clear way, including clear definitions of the type and number of different competency dimensions in the domain and clear specifications of the situations that students and graduates should be able to manage successfully at each competency level or for each competency profile. Definitions should reflect not only institutional requirements, as described in degree qualifications frameworks like the EQF, but should also contain subject-specific aspects, such as curricular aspects, content areas, and situational contexts, as well as cognitive levels. For the development of adequate models, it is indispensable to cooperate systematically with experts of teaching methodology in higher education. Since content and teaching methods in each discipline may vary considerably between states and institutions of higher education, it is essential to reach a nationwide consensus with experts of subject-specific teaching methodology with regard to what core contents and methods should be included in an assessment, for example, as in economics in the USA (Walstad and Rebeck 2008). For the development of the cognitive component of competency assessment, psychologists should be consulted.

Domain-specific and generic competency models for higher education that are operationalized through adequate methods lay the groundwork for measuring individual students’ study performance both longitudinally over the course of studies and cross-sectionally in comparison to other students from the same degree course or university. These are necessary for evaluating the effectiveness of learning opportunities and facilitating the continuous improvement of higher education systems. Since the organization of higher education institutions may differ considerably, we recommend first analyzing competencies to be acquired separately according to domains.

3.2.2. Development of assessments and comprehensive validation

Theoretical models of the structure, levels, and development of generic and domain-specific competencies need to be operationalized through suitable measuring instruments. Such instruments should meet strict psychometric criteria (e.g. AERA, AEA
The nature of the psychometric methods (e.g. appropriate item-response models) should evolve from the conceptual model of competency (construct definition) and not vice versa. Put another way, the psychometric model should not dictate the nature of the construct to be measured.

Test development could be specified in line with Pellegrino, Chudowsky, and Glaser’s (2001, 44) assessment triangle or alternatively and compatibly with Mislevy and Haertel’s (2006, 8) evidence-based design. Following Pellegrino et al., the first step consists in defining the construct (‘cognition’) by analyzing and defining the domain (domain analysis) and modeling the domain-specific construct(s) to be assessed (cognition vertex of triangle) (see Section 3.2.1). The second step is to construct items consistent with the construct definition and likely to elicit behavior that provides evidence of an individual’s standing on the construct (observation vertex). The third step consists of testing the measuring instrument empirically, converting item responses into test scores, and drawing and validating inferences from scores to the construct (interpretation vertex). These analyses always include evaluations of fit of the data to the theoretically modeled constructs and to the corresponding test score interpretations. After the empirical analyses, further decisions can be made, for example, as to whether and to what extent models and instruments need further development and whether they are suitable for assessing students’ competency in higher education (e.g. Briggs 2011). Alternatively, following Mislevy and Haertel, test development begins by stating a claim to be made on the basis of test scores (cf. construct/cognition). From the claim, developers determine the evidence (test-item responses) that would be needed to support or refute the claim and then create items or tasks that would elicit the needed evidence (cf. observation). And finally from there analyses examine the extent to which the observations/evidence support the proposed claim.

Current research shows that item formats must be selected with particular care because they should be consistent with the construct to be measured. Selected-response formats are often preferred for large-scale assessment not because they fully measure the construct but because they are inexpensive, time efficient, can be objectively scored, and fit well within current psychometric scaling methods (Schmeiser and Welch 2006). However, the selection of item formats should depend on the competencies to be measured and not on the convenience of a popular psychometric method. For example, driver examinations across nations include both a selected-response portion and a performance-test portion. No one would believe that passing only a multiple-choice test would demonstrate driving competency. Indeed, constructed-response formats are both valued by educators and widely used in international studies (e.g. CLA). Alternatives to standardized constructed- and selected-response tests include portfolio assessment (e.g. the VALUE project). However, for large-scale assessment, especially if stakes are high, portfolios are susceptible to plagiarism, multiple authorship, and lack of standardization, and they are difficult and time-consuming to evaluate (e.g. Shavelson, Klein, and Benjamin 2014).

International standards, such as the Standards for Educational and Psychological Testing (AERA, APA and NCME 2014), provide criteria for evaluating the interpretability (validity) of test scores: ‘test content,’ ‘internal structure,’ ‘response processes,’ and ‘relations to other variables.’ Test content is analyzed to determine how accurately test items represent the theoretical constructs. The internal structure of a test is analyzed to determine whether the relations between single items or different parts of a test represent the construct. Analysis of response processes determines item clarity and provides evidence that has a bearing on test takers’ mental processes. Finally, relations
between test scores and other, external variables (in a nomological network) provide evidence of interpretability (for example, the relation between student knowledge and the courses students have attended at university).

The above considerations highlight essential activities in building assessments of higher education students’ and graduates’ domain-specific skills and knowledge. They are key prerequisites for conducting large-scale, output-oriented assessments that aim to provide evidence on the quality of degree courses and institutions at the level of the educational system. The lack of research on domain-specific competency models may account for the equal lack of research on measurement models for higher education (see e.g. Osterlind and Ze Wang 2012a, 2012b; Webb, Shavelson, and Steedle 2012).

Further research can draw on assessments in the school sector, such as PISA. Furthermore, projects can draw on experiences from related kinds of assessment, such as indirect measures of competency. In some cases, existing instruments might be adopted or adapted if they tap portions of the competency to be measured. Internationally, instruments from current research programs, such as the CLA in the AHELO study, might serve as a point of departure.12

We cannot interpret assessment results without taking into account the context of their origin. Therefore, direct large-scale competency assessment from different degree courses and universities also needs to analyze contextual conditions in order to generate results that can inform educational policy. Research so far suggests that the following additional factors are relevant for assessing competency: (1) non-cognitive competencies, such as study-related and job-related attitudes, interests, and motivation; (2) socio-biographic data, including key parameters like age or sex, but also students’ and graduates’ educational history, study path, and previous professional careers; (3) data about the degree course and institution of higher education, such as curriculum, teaching performance, learning opportunities, infrastructure, and regional characteristics; (4) generic and domain-specific competencies required in professional life, as described by global employers. Research projects could be dedicated to systematically mapping out such a conceptualization with its different facets and causal relations at the national level (e.g. for a mapping on the European level, see the Tuning project).

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Notes**

1. The European Union’s (EU) Bologna process created the European higher education area, which represents heterogeneous structures that were unified mainly with regard to university degrees. Yet, numerous studies show that, at present, the system has no homogeneous logic for different system levels (e.g. on the scope and speed of implementation of the new bachelor and master degree courses or on the composition of courses and student bodies; see Brennan, Patel, and Tang 2009).

2. Two other OECD studies have included international large-scale assessment of competencies: the Programme for the International Assessment of Adult Competencies and International Adult Literacy Survey, but they did not specifically target higher education students and graduates, only adults in general (OECD 2004; see Humburg & van der Velden 2015).
3. To assess the acquisition of analytical thinking, the study uses an international adaptation of the Collegiate Learning Assessment (CLA) developed in the USA (CAE 2014).

4. The participating countries assessed generic competencies, engineering competencies, economic competencies, or a combination of these: generic competencies were assessed by Colombia, Egypt, Finland, Korea, Kuwait, Mexico, Norway, Slovakia, and the USA (three states). Engineering competencies were assessed by Abu Dhabi, Australia, Canada, Colombia, Egypt, Japan, and Mexico. Economic competencies were assessed by Belgium, Egypt, Italy, Mexico, the Netherlands, Russia, and Slovakia.

5. For the pre-study of TEDS-M, the Mathematics Teaching in the twenty-first century study assessed pre-service mathematics teachers of lower secondary education in Bulgaria, Germany, Mexico, South Korea, Taiwan, and the USA.

6. The 17 participating countries were Botswana, Canada, Chile, Germany, Georgia, Malaysia, Norway, Oman, the Philippines, Poland, Russia, Singapore, Spain, Switzerland, Taiwan, Thailand, and the USA.

7. On content and on measuring methodology of CLA, see Klein et al. (2007) and Shavelson (2010).

8. Including economics, biology, chemistry, computer sciences, criminal law, history, literature, mathematics, music, physics, politics, psychology, and sociology.

9. The rubrics describe learning outcomes very generally and need to be further specified for institutions, disciplines, and courses.

10. The following countries participated in the REFLEX study: Austria, Belgium, the Czech Republic, Estonia, Finland, France, Germany, Italy, Japan, the Netherlands, Norway, Portugal, Spain, Switzerland, and the UK. A parallel project was conducted in Russia and Latin America.

11. For another large-scale survey not primarily in higher education, assessing the linguistic and numerical skills of adults from Bermuda, Canada, Italy, Norway, Switzerland, and the USA, see IES National Center for Educational Statistics (2014).

12. Many domain-specific and generic tests are available and in wide use, but must nevertheless be selected carefully, since sufficient validity has not always been established.

References


ETS (Educational Testing Service). 2014a. “Graduate Record Examination (GRE) General Test.” [http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/](http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/)


ETS (Educational Testing Service). 2014c. “Measure of Academic Proficiency and Progress (MAPP).” [http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=ff3aaf5e44df4010VgnVCM10000022f95190RCRD&vgnextchannel=f98546f1674f4010VgnVCM10000022f95190RCRD](http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=ff3aaf5e44df4010VgnVCM10000022f95190RCRD&vgnextchannel=f98546f1674f4010VgnVCM10000022f95190RCRD)


