Measuring Learning Outcomes

Joakim Caspersen, Centre for the Study of Professions, Oslo and Akershus University College for Applied Sciences, PO Box 4 St. Olavs Plass, NO-0130 Oslo, Norway. Joakim.caspersen@hioa.no

Jens-Christian Smeby, Centre for the Study of Professions, Oslo and Akershus University College for Applied Sciences, PO Box 4 St. Olavs Plass, NO-0130 Oslo, Norway. Jens-Christian.Smeby@hioa.no

Per Olaf Aamodt, Nordic Institute for Studies of Innovation, Research and Education, P.O. Box 2815 Tøyen, N-0608 Oslo. Per.Aamodt@nifu.no

Abstract: The growing interest for measurement of learning outcomes relates to long lines of development in higher education, the request for accountability, intensified through international reforms and movements such as the development and implementation of qualifications frameworks. In this article, we discuss relevant literature on different approaches to measurement and how learning outcomes are measured, what kinds of learning outcomes are measured, and why learning outcomes are measured. Three dimensions are used to structure the literature: Whether the approaches have an emphasis on generic or disciplinary skills and competence; whether they emphasise self-assessment or more objective test based measures (including grades); and how the issue of the contribution from the education program or institution (the value-added) is handled. It is pointed out that large scales initiatives aimed at comparison between institutions and even nations seem to fall short because of the implicit and explicit differences in context, while small-scale approaches are burdened with a lack of relevance outside of local contexts. In addition, competence (actual level of performance) is often confused with learning (gain and development) in many approaches, laying the ground for false assumptions about institutional process-quality in higher education.

Keywords: Learning outcomes, measurement, knowledge, learning, assessment, Higher education

Introduction: Learning Outcomes and the Drive for Measurement

With growing attention being given to intended learning outcomes in higher education – as described in, for example, the European Qualifications Framework (EQF) – it has become clear that the development of adequate measurements of achieved learning outcomes (Caspersen, Frølich and Muller, this issue, pp.) is needed (Stensaker & Sweetman, 2014). The focus on achieved learning outcomes has been reinforced with the Bologna Process in Europe and the EQF, as well as with trends towards greater accountability and outcome focus on other continents[1]. The assessment of what students learn and know may be used as an indicator of the quality of institutions and study programmes. Douglass *et al*. (2012) argued that the precise measurement of learning outcomes was globally seen as a means of assessing the quality and effectiveness of higher education institutions, stating, 'Government ministries, along with accrediting agencies, the media, and critics of higher education, desire a universal tool to measure learning outcomes at the campus level, and that can be compared across institutions, regions, and perhaps even countries' (p. 318). Our point of departure is not limited to measures of learning outcomes in this specific use of the term. Rather, the purpose is to provide more knowledge on how learning outcomes are measured by identifying various approaches in the current literature and using examples from recent and ongoing projects. We use the term *measurement* in a broad sense, i.e. as a systematic assessment and reporting of students' learning.

Research Questions

A typical approach to the measurement of outcome is to ask students to assess their learning outcomes on a range of different variables which can often be subsumed under different dimensions (e.g. knowledge, general competence and skills). Such self-reported outcomes are often aggregated to indicate the quality of programmes. However, as greater attention has been given to outcomes and greater emphasis has been placed on the governance of higher education institutions in most Western countries (Frølich & Caspersen, 2015), large-scale initiatives have been launched nationally and internationally. The need for systematic assessment of quality in higher education prompted the

[1]In the US, the increased focus on measurement has been particularly pushed through in the aftermath of the Spellings Commission (Commission on the Future of Higher Education), which focused on accountability in higher education through the measurement and systematisation of results.

Organisation for Economic Co-operation and Development (OECD) to undertake a feasibility study on the Assessment of Higher Education Learning Outcomes (AHELO). This international testing programme in higher education is an ambitious project which builds on a number of existing assessment experiences in different countries (e.g. the Collegiate Learning Assessment [CLA], a US-based test, and a multiple choice instrument developed by ACER in Australia) and student surveys. Together, they represent measurement models that move across the different dimensions presented earlier and aim to respond to a broad set of purposes. To provide a basis for our discussion on measurements of learning outcomes, we used a limited review of the research literature on measurement as a background to point out and discuss three different dimensions that are emphasised as important to understand learning outcomes They relate to whether they emphasise self-assessment or more objective test-based measures, including grades (how); whether they emphasise generic skills or disciplinary skills and competence (what); and, finally, how the contribution of the education programme or institution (the value added) is handled. This final dimension also relates to an important distinction between measurements of knowledge and measurements of learning and refers to the purpose (why) of measurement.

In short, we aim to shed light on and point out dilemmas regarding:

- How learning outcomes are measured.
- The kinds of learning outcomes that are measured. and
- Why learning outcomes are measured.

Methods

The background for our discussion is a literature search through EBSCOhost and the ERIC, SocINDEX and Academic Search Premier databases. The keywords were *skills*, *learning outcomes* or *competencies* (in the singular and plural) in combination with *higher education* and *measure* or *measurement* (in the singular and plural). The search covered articles in peer-reviewed academic journals during the period 2010-2015. The keywords *skills* and *competencies* were included in addition to *learning outcomes* because much of the literature on measurement of outcomes is not explicitly directed towards outcomes *per se*, but learning and growth in a broad sense. In addition, only

academic journals that focus on higher education in general were included.[2] We also excluded results that focused solely on learning outcomes as pedagogical devices. This limited the results to 246 articles. The abstracts and titles were then reviewed for relevance and the results were narrowed down to 94 articles that were included in a final screening. For this round, two of the authors evaluated the articles for relevance, including the abstract, title and full text when in doubt. This further reduced the number of articles to 46, which were all included in the review. All three authors then read the articles and provided short summaries, which were then compared and adjusted. As the search procedure did not cover all relevant literature and was limited to the period up to 2015, we also included articles from higher education journals and books that we knew from our work in the field, even though these were not part of the results from the review. These articles were also summarised and included. All summaries and full-text articles were used as the basis for this article.

Although the search strategy can be described as broad, it is evident that not all relevant literature is included on our search. For instance, much of the literature on measurement is published in reports by international associations such as the OECD, CEDEFOP, UNESCO and their contractors, (i.e. IEA) and is not included in scientific databases. Furthermore, limiting the last year in the review to 2015 (as the search was undertaken in spring 2016), there could be newer contributions that were not included. However, our aim is not to produce a comprehensive synthesis of the literature, but rather to use the literature to point out important themes for further discussion and exploration.

How: Grades, Self-Reported and Test-Based Measures

One of the main debates in the field of learning outcome measurement is between scholars who advocate the use of surveys based on self-reported measures and those who support the use of more 'objective' measures and test-based measurement (Douglass *et al*., 2012). However, we should not overlook the traditional way of measuring outcomes in higher education: grades. Even though grades are generally not included in the debate on how to measure learning outcomes in a restricted sense,

[2] The journals were Quality & Quantity, Professional Development in Education, British Educational Research Journal, Scandinavian Journal of Educational Research, Education, Research in Higher Education, British Journal of Educational Psychology, Teaching in Higher Education, Quality Assurance in Education: An International Perspective, Assessment Update, Educational Sciences: Theory and Practice, Educational Psychology, Electronic Journal of Research, Educational Psychology, Educational Research and Reviews, Studies in Higher Education, Assessment & Evaluation in Higher Education.

they are used as a means of communicating assessments of students' competencies and indirectly their learning outcomes. They may be based on exams and assignments, as well as on students' portfolios and performance in class. The literature on grading is relatively limited (Bloxham & Boyd, 2012) and it is argued that grading is flawed and affected by a number of socially driven factors that are not well understood (Yorke, 2011). Grades are also criticised for their lack of reliability. Studies have shown that assessors disagree significantly in their grading (Sadler, 2010) and it has been argued that the lack of common standards is a key characteristic of grades (Yorke, 2009). Despite these weaknesses, grading is still used and accepted as an assessment of individual students' performance. It has, however, great weaknesses as an indicator of the quality of study programmes and institutions since criteria and scaling vary across educational programmes, institutions and nations. One reason is that, even though attempts to set up common criteria developed in qualifications frameworks and the use of external examiners aimed to contribute to national standards in the different fields of study, a certain distribution of grades – for example, a Gaussian normal distribution over time – is also called for. It may be argued that the lack of standardisation is a problem not only for those who use the data for assessment and comparative purposes, but also for employers and graduates.

Procedures may be developed to make grades valid indicators for comparing course levels (Rexwinkel *et al.*, 2013). Such procedures imply that descriptions of learning outcomes become highly-specified and therefore demand significant standardisation, at least at the national level. The VALUE rubrics developed in the US are a kind of highly-specified learning outcome descriptions that aim to guide grading and provide better student feedback on various types of assignments. They incorporate diverse outcomes and include explanation of issues, evidence, influence of context and assumption, and ethical self-awareness. The rubrics were created by teams of faculty across the nation and represent something close to a set of national student learning outcomes (Pike, 2014). The rubric for scientific writing consists of concrete criteria to be addressed in an assignment. It also differentiates between what should be expected from novice, intermediate and proficient students. The use of the rubric in biology laboratory courses showed an increase in the substance and consistency of grading among teaching assistants, particularly with respect to the assessment of student achievement in scientific reasoning and writing. Students also reported that the use of rubrics facilitated their

learning (Timmerman *et al.*, 2011). The use of rubrics implies a standardisation of student assessment criteria. Depending on how they are used, they may also imply a standardisation of grading.

The CLA has been very influential in setting the standards for the test-based measurement of learning outcomes in the US. Developed by the Council for Aid to Education (CAE), the CLA is a test with open-ended questions to measure higher-order skills, such as analytic reasoning and evaluation, problem-solving and written communication (Wolf *et al.*, 2015). An alternative test-based approach based on multiple-choice questions in combination with open-ended questions has been developed by the Australian Council for Educational Research (ACER). These tests also differed in their emphasis on generic versus discipline-specific skills. A more detailed description of these tests with illustrations is presented in a study by Tremblay *et al.* (2012). Zlatkin-Troitschanskaia *et al.* (2015) argue that competencies should be measured by tests and that passing only a multiple-choice questionnaire cannot demonstrate the level of competencies. Others have emphasised that it was challenging to motivate students to take these tests, which are often time-consuming (Douglass *et al.*, 2012).

The primary aim of grades and tests is to measure individual students' learning outcomes – or, more precisely, students' level of acquired knowledge and skills. Second, tests can be used as aggregates to assess the performance of a programme or institution. The aim of surveys and self-reported outcome measures is, however, mainly to assess the quality of educational programmes and their contribution to students' learning through quality assessment, quality assurance and various types of ranking of programmes and institutions. There are examples of large-scale self-report surveys conducted at national (e.g. NSSE) and international levels (e.g. REFLEX), as well as a large number of small-scale surveys carried out at institutional, programme and course levels. Self-report surveys are relatively easy and resource-effective to conduct. The large-scale surveys are also used as the basis for research that examines the impact of different factors on student learning. Based on the NSSE data alone, more than 150 research articles have reportedly been published.[3] Various analyses of the validity of self-reported measurements of outcomes conclude that it is reasonable. For instance, Douglass *et al.* (2012) showed that differences between ethnic and demographic groups in self-reported outcomes

[3] http://nsse.iub.edu/html/pubs.cfm?action=&viewwhat=Journal%20Article,Book%20Chapter,Report,Research%20Paper&pubFlag=yes

corresponded with what was known about such differences based on other measures, thereby supporting the validity of the self-reported learning outcome measures. However, low-achieving students tend to overestimate their achievement and high-achieving students tend to underestimate their achievement (Boud & Falchikov, 1989; Dochy *et al.*, 1999; Mowl & Pain, 1995; Orsmond *et al.*, 1997). Furthermore, Humburg and van der Velden (2015) found that self-reported results did not correlate well with test-based measures of similar constructs in cross-country comparisons. They concluded, however, that self-reported measures were adequate as long as they were restricted to within-country differences and were useful to predict differences across fields of study. Other studies have concluded that variances existed in some dimensions of self-reported learning outcomes between disciplines and professions and that these differences reflected different knowledge structures rather than 'real' outcome differences (Caspersen *et al*., 2014; Sweetman *et al*., 2014). Although those who advocate self-reported measures admit that these are somewhat biased, they claim that they are complementary supplements to test-based methods (Douglass *et al.*, 2012; Gonyea & Miller, 2011). It is recognised that self-reported measures do not measure actual outcomes, but rather students' perceptions and attitudes (Gonyea & Miller, 2011). Some evidence shows that students are fairly accurate in their estimation of personal performance in a multiple-choice exam. In a study of undergraduate psychology and teacher training students, 90% were able to predict a correct level, but were far less able to acknowledge their metacognitive competence to predict their performance (Händel & Fritzsche, 2015). The ability of students to predict personal performance in an exam is not the same as being able to assess their learning outcomes in general, and students' ability to understand the assessment criteria and their performance increases as they become more experienced. Others emphasise that students have a limited capacity to assess their cognitive outcome and that a close correlation exists between self-reported outcomes and students' overall satisfaction with college (Bowman, 2014). A study concludes that there is a significant positive relationship between students' self-rated competencies and later vocational success. However, students' self-rated competencies accounted for achievement of individual goals in their current occupation to a higher degree than objective success by annual income (Braun *et al*., 2011). This illustrates the subjective and attitudinal aspects of self-reported measures.

It is important to keep in mind that the aim of self-reported measures was to measure important aspects of the quality of educational programmes in terms of the factors contributing to students' learning rather than to develop measures in terms of students' acquired competence (Gonyea & Miller, 2011). Although test-based and self-reported approaches are often considered to be opposite approaches, they may, of course, be combined. The AHELO Feasibility Study combined testing of field-specific competencies in engineering and economics and generic competencies and included a combination of constructed-response tasks (CRTs) and multiple-choice questions (MCQs). The aim of these tests is to focus on students' skills in the application of concepts and problem-solving, not on their factual knowledge (Tremblay *et al.*, 2012).

What: Generic and Disciplinary Outcomes

In the literature, a key distinction is made between generic and disciplinary learning outcome measures. Emphasis on generic outcomes can be seen as a response to an inherent limitation with grades: they mainly measure students' acquired knowledge in specific disciplinary domains. In contrast, learning outcomes are defined as a broader set of competencies, i.e. 'what a learner is expected to know, understand and/or be able to do at the end of a period of learning' (European Commission, 2012, p. 12). Such a broad perspective is also emphasised in the literature on employability. In their review on the conceptualisation of employability, Williams *et al.* (2015) argued that a core dimension was 'anything an individual possesses that can be seen as leading to an increased probability of positive economic outcomes, or other personal outcomes relating to the area of work' (p. 11).

The importance of generic skills and abilities, irrespective of subject area, is often highlighted in the research literature reviewed for this study. A benefit of this focus is that the same scales may be used to compare students across fields and institutions. Outcomes are seen as the development of general competencies and skills, not in direct relation to the specific subject. In large-scale studies of graduates' self-reported skills (e.g. REFLEX, HEGSCO, NSSE, NSE), students are asked to assess their educational outcomes in terms of their mastery of their field or discipline; analytical and critical thinking; various types of skills and abilities such as numeracy, literacy, oral presentation and

problem-solving skills; and ability to coordinate activities and work productively with others. Other studies have a more limited scope, focusing on, for example, social engagement (van den Wijngaard *et al.*, 2015). Even though these studies cover many of the same dimensions, and a number of similar and even identical items exist, no validated instruments measuring these broad sets of learning outcomes have been established yet (Caspersen *et al.*, 2011; Karlsen, 2011).

Another trend in studies using self-reported measurements of learning outcomes is the performance of subject- or profession-specific assessments that focus on core competencies, including knowledge, skills and problem-solving. Some examples are medical graduates' standards of outcome (Dimoliatis *et al.*, 2014), pre-service teachers' perceptions of their competencies and attitudes (Köksal, 2013; Struyf *et al.*, 2011) and students' self-confidence in science, mathematics and engineering courses (Litzler *et al.*, 2014). There are also many assessments that focus on even more specific skills, such as listening skills among students of psychological counselling and guidance (Cihangir-Çankaya, 2012), sales and customer orientation in a personal selling class (Totten, 2014) and media literacy competencies among teachers (Recepoglu & Ergun, 2013). Some examples are also found in rubrics that address readiness for doctoral-level work (Maher & Barnes, 2010) and undergraduate scientific-reasoning skills (Timmerman *et al.*, 2011). In our literature search, we excluded all profession- and discipline-specific journals. Nevertheless, such studies may have obvious qualities when the aim is not to conduct assessments across fields of study, but rather to address specific outcomes in a course or programme.

Many studies have also been conducted on test-based measures of domain-specific as well as on generic skills. The AHELO study, which addresses generic and analytical cognition and domain-specific competencies in economics and engineering, used an international adoption of the CLA (Zlatkin-Troitschanskaia *et al.*, 2015) to assess critical and analytical thinking. Critical thinking is considered to be one of the core competencies and outcomes of higher education and several similar measures have been developed, including some in specific educational programmes such as medicine (Macpherson & Owen, 2010). In their review of these measures, Liu *et al.* (2014) concluded that the following dimensions were essential and recommended their inclusion in the next generation of assessments: analytical skills (the evaluation of evidence and its use and the analysis and evaluation of

arguments), synthetic skills (the understanding of implications and consequences and the development of sound and valid arguments) and the understanding of causation and explanations that are relevant to the analytical and synthetic dimensions.

Test-based alternatives to self-reported assessments of profession-specific knowledge and skills have also been developed. Unlike traditional exams, these tests tend to focus on core competencies within a specified professional field or domain. An example is the Teacher Education and Development Study in Mathematics (TEDS-M) which assesses pre-service mathematics teachers with regard to their content knowledge, pedagogical content knowledge and pedagogical knowledge, as well as their beliefs about mathematics teaching and learning (Blömeke *et al.*, 2011). Another example is the Modelling and Measuring Competencies in Higher Education – Validation and Methodological Innovations (KoKoHS) programme in Germany and Austria which includes 70 projects that assess domain-specific and generic competencies in various disciplines, taking into account curricular and job-related requirements (Zlatkin-Troitschanskaia *et al.*, 2015).

Whereas the CLA was designed for institutional assessment, the CLA+ was further developed to identify individual students and can be used as a competence certification measure to document their critical thinking and problem-solving abilities. In addition to being presented with realistic problems, students are asked to respond to 25 selected-response questions. Ten measure scientific and quantitative reasoning, ten assess critical reading and evaluation, and five evaluate a student's ability to critique an argument. Institutions may include these scores on students' transcripts to document their critical thinking and problem-solving abilities (Pike, 2015). Therefore, the CLA+ serves as an illustration of how measurement approaches can combine disciplinary (scientific and quantitative reasoning) and generic skills (critical reading and evaluation). The distinction between generic and disciplinary approaches is analytical. Many of the surveys and tests developed to assess learning outcomes aim to measure these in terms of subject and discipline-specific knowledge, as well as a broader set of competencies and skills (even though the balance varies). Grades, however, tend mainly to measure acquired knowledge.. The CLA+ illustrates that generic skills may also be included in student grading.

Why: Quality Assessment and Value Added

An indicator on a study programme of high quality is how it contributes to students' learning and development during the course of studies, or what value has been added. Therefore, if measuring students' acquisition of knowledge is meant to serve as an assessment of the quality of study programmes of institutions, one must take into consideration that the performance in a given student group is strongly related to what it brought to the study programme in terms of previous school achievement, family background, etc. Many different measurement approaches have been developed in order to facilitate this (Coates, 2009; Liu, 2011). One approach is to control for grades in upper secondary school in the assessment of higher education outcomes. However, it is hardly an adequate measure of the level of acquisition of competencies, as social background and parents' educational background also have a significant independent effect on students' grades in higher education (Karabel & Halsey, 1977; Kuh *et al.*, 2008; Shavit & Blossfeld, 1993). Sorting out all other factors that impact students' learning outcomes is also difficult. The use of students' theses and dissertations as indicators of the educational quality of graduate degree programmes suffers from the same weaknesses as those associated with the use of grades as indicators (Hamilton *et al.*, 2010).

A simple way to address the value-added approach is to ask students to assess the extent to which they have acquired various types of knowledge, skills and competencies during their studies. These kinds of scales are included in several large-scale studies (e.g. REFLEX, HEGSCO, NSSE, NSE) and small-scale surveys at the programme level. However, grades and self-reported outcomes correlate weakly (Carini *et al.*, 2006; Caspersen *et al.*, 2014; Sweetman *et al.*, 2014). One reason may be that self-reported outcomes include consideration of the value added of higher education, whereas grades upon graduation (supposedly) refer to a set level of proficiency. A more elaborate way to address the value added of higher education is a longitudinal design. In the Student Experience in the Research University Survey (SERU-S), which was initially developed at the University of California, Berkeley to analyse the student experience among undergraduate students, students are asked to estimate retrospectively their knowledge and skills at enrolment and their current level using the same dimensions, thereby allowing students' knowledge gain to be calculated. Although the inclusion of a

retrospective scale is argued to provide an adequate diagnostic tool for learning outcomes (Douglass *et al.*, 2012), this design raises some of the same questions concerning students' ability to estimate their acquired learning outcomes as those raised in the other previously discussed self-reported studies.

In the test-based approach, instruments are developed to test the aptitude of students applying to college (SAT) and their academic proficiency and progress. The pre-test–post-test design was formulated to measure how higher education contributed to students' development of generic skills. Based on comprehensive data on undergraduate students who took the CLA at various points before and during their college education, Arum and Roksa (2011) concluded in their book that 45% of the students showed no significant improvement in learning, including critical thinking, complex reasoning and writing, during their first two years of college and that 36% did not show any improvement during their four years of college. Their book is highly controversial, one objection being that the outcome is incredibly small at most institutions owing to the high correlation between the SAT and CLA scores (Douglass *et al.*, 2012). The development of CLA+ should be seen as a response to this, with an explicit emphasis on individual growth and multiple measurements within individual educational careers.

Conclusion: The Pitfalls of Comparison, Conflation and Standardisation

We argue that the main problem with using grades as indicators of learning outcomes at the aggregate level is the lack of standardisation. As also discussed by Caspersen, Frølich and Muller (this issue, pp.), the intent of the standardisation of education with the Bologna Process was to establish some commonality in a diverse European higher education landscape (Elken, 2015). The Tuning Project began in 2000 with the aim to follow up these processes by changes in education structures and converging the different national systems in Europe. When the European Qualifications Framework was established in 2008, the aim had shifted to benchmarking, or 'referencing' the qualifications of the proliferating national QFs to a common framework for Europe, out of which the AHELO project developed. Standardisation not only involves ways of developing common standards, but also relates to general methodological problems in comparative research. How can one be certain that what is being compared has comparable entities? What does comparison really highlight? The claim that

differences in self-reported learning outcomes between disciplines and professions reflect differences in 'knowledge structures' serves as an example (Caspersen *et al.*, 2014). The standardisation of grading is obviously challenging. However, the development of measurement instruments as a means to assess learning outcomes globally is still in its infancy (Wolf *et al.*, 2015).

In the Introduction, we raised three research questions: How are learning outcomes measured? What kinds of learning outcomes are measured? Why are learning outcomes measured? Our review highlighted different approaches, but all have their shortcomings. We will use our three research questions to highlight the inherent problems with comparison and with the conflation of learning and competence (and quality) and the challenges of standardisation.

The Problems with Comparison

The most important controversy in the debate on the measurement of learning or learning gains is between those who advocate a test-based approach and those who defend the appropriateness of self-reported measures. The test-based approach has some clear methodological advantages, but its implementation is also demanding and challenging. The experiences from the AHELO project demonstrate some of the challenges when test-based approaches are applied in large-scale international comparisons. It is challenging to come to an agreement about instrument development across different countries and to adapt instruments to national, cultural and linguistic settings. Furthermore, there is little international consensus about generic skills and their connection to professional, cultural and disciplinary contexts. For any future AHELO project, the development of completely new tailor-made instruments should be considered (Tremblay *et al*, 2012).

Previous research also indicates that making comparisons across disciplines based on self-reported measures is problematic (Caspersen *et al.*, 2014). One of the main problems with using grades as indicators of learning outcomes is the lack of standardisation. The methodological problems in comparative research pose challenges to the standardisation of grading and measures. Previous comparisons of grades, self-reported measures and the test-based approach have revealed the strengths and weaknesses of each. Identifying these is not only relevant to understand the differences between them, but also may pave the way for further development of learning outcome measures. Different

approaches may be combined in new ways, and new measures may be developed to serve various purposes. The development of the test-based approach has highlighted the need for further development of self-reported instruments and for further analysis of what the different measures are actually measuring. Moreover, grading has tended to measure discipline-specific knowledge acquisition, whereas test-based and self-reported measures have also emphasised generic skills. The CLA+ is an example of how generic skills may be included in students' transcripts and as a basis of grading. A further development in this direction would be in line with the EQF, which emphasises not only knowledge, but also skills and competence.

The Conflation of Learning and Competence (and Quality)

It is evident that grades, knowledge and skills tests and self-reported measures are developed for different purposes and that they measure somewhat different aspects of student learning outcomes. Grades are developed to communicate students' acquired knowledge, to give feedback to students and to function as indicators of students' competence to employers. Also, tests are developed to assess their acquired knowledge. But grades and tests do not directly tell us anything about the quality of student learning, even though the acquisition of knowledge is, of course, a result *of* learning. On the other hand, since one cannot expect students to assess their knowledge in relation to the demands in the study programme or to their fellow students, self-report instruments are not well suited to assess acquired knowledge. However, students may well assess the experience they have gained. This shows that it is important to distinguish between measurement of student learning and measurement of students' level of competence. Self-assessment mainly states the level of individual growth or learning and indicates little about the absolute level of knowledge. Students starting from a low level of knowledge can experience a great deal of growth and learning without attaining a high level of knowledge or proficiency. Students with a high level of knowledge initially can experience little growth or learning, but retain a great deal of knowledge upon graduation - the higher their grade point average (GPA) in upper secondary education, the lower their self-reported learning outcomes in higher education.

Such findings highlight the importance of separating the different purposes of measuring learning outcomes and, in particular, of refraining from mixing the measurement of learning (growth) with the measurement of knowledge at a given point in time and of being clear whether one wants to measure quality, competence or learning at any given point. Measurements of knowledge (in this context, knowledge upon graduation) do not indicate anything about learning – unless some kind of value-added design is implemented. However, asking students about their growth in relation to their starting point does not reveal anything about knowledge or proficiency, only about 'the degree of learning'. Hence, we argue that measures of knowledge can be used to assess learning (given a sound value-added design), but measures of learning cannot be used as indicators of knowledge. Thus, if the main purpose is to assess individual students' acquired competencies, then grades and tests may be well suited. If, however, the aim is to assess the quality of programme delivery, one must focus on the learning process, either by testing at the start and at the end of the studies or by asking students about what they have gained during their studies. Furthermore, the goal of the assessment has a bearing on the type of LO to be measured and the measurement approach. The AHELO feasibility study has shown that MCQ-based tests are more reliable than open-ended questions-based tests such as the CLA. Therefore, wherever the test is seen as a formative assessment, a focus on generic skills/open ended questions' instruments is adequate. If the goal is more accountability driven or high stakes, then instruments using MCQs provide more reliable measurements.

The Way Forward

In a policy-laden field such as the measurement of learning outcomes in higher education, there are strong demands for policy-relevant conclusions and recommendations. The measurement of learning and knowledge has received a great deal of attention and its development has occurred in many different disciplinary traditions, with sometimes overlapping, and other times contradictory approaches. Moreover, as measurement has been ascribed increasing importance, a quasi-commercial market has developed where proponents of different approaches to measurement advocate their own approach. This has been described as a 'learning outcomes race' (Douglass *et al.*, 2012) where the aim is to assess the 'value added' of colleges and universities. What should be the recommendations for

future research on the measurement of learning outcomes? Many attempts to measure these may be useful for course providers, but they may be of limited value beyond the local context. The AHELO project represents the other extreme as an attempt to make international comparisons.

Although standardisation (of grades and other measurements) seems like a way forward, it is difficult to discern any true practical results in the near future. Large-scale coordinated efforts such as the AHELO have proven to be difficult in terms of implementing existing test instruments in addition to being extremely resource-demanding. However, this does not mean that small-scale efforts with little relevance for other contexts should dominate the agenda. Perhaps the middle road is the most productive: the systematic development of different indicators, with systematic comparisons of results, will probably provide the best way forward in terms of costs and benefits. Also, systematic meta-reviews of the results and experiences from a wide range of assessment projects may be useful.

REFERENCES

ARUM, R. & ROKSA, J. (2011) *Academically Adrift: limited learning on college campuses* (Chicago, University of Chicago Press).

BLOXHAM, S. & BOYD, P. (2012) Accountability in grading student work: securing academic standards in a twenty-first century quality assurance context, *British Educational Research Journal*, 38, pp. 615–634.

BLÖMEKE, S., SUHL, U. & KAISER, G. (2011) Teacher education effectiveness: quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge, *Journal of Teacher Education*, 62, pp. 154–171.

BOUD, D. & FALCHIKOV, N. (1989) Quantitative studies of student self-assessment in higher education: a critical analysis of findings, *Higher Education*, 18, pp. 529–549.

BOWMAN, N. A. (2014) The meaning and interpretation of college student self-reported gains, *New Directions for Institutional Research*, 2014 (161), pp. 59–68.

BRAUN, E. M. P., SHEIKH, H. & HANNOVER, B. (2011) Self-rated competences and future vocational success: a longitudinal study, *Assessment & Evaluation in Higher Education*, 36, pp. 417–427.

CARINI, R. M., KUH, G. D. & KLEIN, S. P. (2006) Student engagement and student learning: testing the linkages, *Research in Higher Education*, 47, pp. 1–32.

CASPERSEN, J., DE LANGE, T., PRØITZ, T. S., SOLBREKKE, T. D. & STENSAKER, B. (2011) *Learning about Quality – perspectives on learning outcomes and their operationalisations and measurement* (University of Oslo, Oslo).

CASPERSEN, J., FRØLICH, N., KARLSEN, H. & AAMODT, P. O. (2014) Learning outcomes across disciplines and professions: measurement and interpretation, *Quality in Higher Education*, 20, pp. 195–215.

CASPERSEN, J., FRØLICH, N. & MULLER, J. (2017) Higher education learning outcomes – ambiguity and change in higher education. *European Journal of Education*.

CIHANGIR-ÇANKAYA, Z. (2012) Reconsideration of the Listening Skill Scale: comparison of the listening skills of the students of psychological counseling and guidance in accordance with various variables, *Educational Sciences: Theory & Practice*, 12, pp. 2370–2376.

COATES, H. (2009) What's the difference? A model for measuring the value added by higher education in Australia, *Higher Education Management and Policy*, 21 (1), 1–20.

DIMOLIATIS, I. D. K., LYRAKOS, G. N., TSERETOPOULOU, X., TZAMALIS, T., BENOS, A., GOGOS, C., MALIZOS, K., PNEUMATIKOS, I., THERMOS, K., KALDOUDI, E., TZAPHLIDOU, M., PAPADOPOULOS, I. N. & JELASTOPULUJ, E. (2014) Development and validation of the 'iCAN!' – a self-administered questionnaire measuring outcomes/competences and professionalism of medical graduates, *Universal Journal of Educational Research*, 2 (1), pp. 19–36.

DOCHY, F., SEGERS, M. & SLUIJSMANS, D. (1999) The use of self-, peer and co-assessment in higher education: a review, *Studies in Higher Education*, 24, pp. 331–350.

DOUGLASS, J. A., THOMSON, G. & ZHAO, C.-M. (2012) The learning outcomes race: the value of self-reported gains in large research universities, *Higher Education*, 64, pp. 317–335.

ELKEN, M. (2015). "New EU instruments for education: vertical, horizontal and internal tensions in the European Qualifications Framework." *Journal of Contemporary European Research*, 11(1).

EUROPEAN COMMISSION (2012). Using Learning Outcomes. *European Qualifications Framework Series: Note 4*. Luxembourg: Publications Office of the European Union

FRØLICH, N. & CASPERSEN, J. (2015) Institutional governance structures, in: J. HUISMAN, H. DE BOER, D. D. DILL & M. SOUTO-OTEROS (Eds) *The Palgrave Handbook of Higher Education Policy and Governance* (London/New York, Palgrave MacMillan), pp. 379–397.

GONYEA, R. M. & MILLER, A. (2011) Clearing the AIR about the use of self-reported gains in institutional research, *New Directions for Institutional Research*, 2011 (150), pp. 99–111.

HAMILTON, P., JOHNSON, R. & POUDRIER, C. (2010) Measuring educational quality by appraising theses and dissertations: pitfalls and remedies, *Teaching in Higher Education*, 15, pp. 567–577.

HUMBURG, M. & VAN DER VELDEN, R. (2015) Self-assessments or tests? Comparing cross-national differences in patterns and outcomes of graduates' skills based on international large-scale surveys, *Studies in Higher Education*, 40, pp. 482–504.

HÄNDEL, M. & FRITZSCHE, E. S. (2015) Students' confidence in their performance judgements: a comparison of different response scales, *Educational Psychology*, 35, pp. 377–395.

KARABEL, J. & HALSEY, A. H. (1977) Educational research: a review and interpretation, in: J. KARABEL & A. H. HALSEY (Eds) *Power and Ideology in Education* (New York, Oxford University Press), pp. 1–85.

KARLSEN, H. (2011) *Klare for arbeidslivet? En drøfting av metodiske utfordringer for måling av læringsutbytte i høyere utdanning* [Ready for work-life? A discussion of methodological challenges in the measurement of learning outcomes in higher education] Report 42/2011 (Oslo, NIFU).

KUH, G. D., CRUCE, T. M., SHOUP, R., KINZIE, J. & GONYEA, R. M. (2008) Unmasking the effects of student engagement on first-year college grades and persistence, *Journal of Higher Education*, 79, pp. 540–563.

KÖKSAL, N. (2013) Competencies in teacher education: preservice teachers' perceptions about competencies and their attitudes, *Educational Research and Reviews*, 8, pp. 270–276.

LITZLER, E., SAMUELSON, C. C. & LORAH, J. A. (2014) Breaking it down: engineering student STEM confidence at the intersection of race/ethnicity and gender, *Research in Higher Education*, 55, pp. 810–832.

LIU, O. L. (2011) Value-added assessment in higher education: a comparison of two methods, *Higher Education*, 61, pp. 445–461.

LIU, O. L., FRANKEL, L. & ROOHR, K. C. (2014) *Assessing Critical Thinking in Higher Education: current state and directions for next-generation assessment* (Princeton, Educational Testing Service).

MACPHERSON, K. & OWEN, C. (2010) Assessment of critical thinking ability in medical students, *Assessment & Evaluation in Higher Education*, 35, pp. 41–54.

MAHER, M. A. & BARNES, B. J. (2010) Assessing doctoral applicants' readiness for doctoral-level work, *Assessment Update*, 22 (5), pp. 8–10.

MOWL, G. & PAIN, R. (1995) Using self and peer assessment to improve students' essay writing: a case study from geography, *Innovations in Education & Training International*, 32, pp. 324–335.

MUSEKAMP, F. & PEARCE, J. (2015) Assessing engineering competencies: the conditions for educational improvement, *Studies in Higher Education*, 40, pp. 505–524.

ORSMOND, P., MERRY, S. & REILING, K. (1997) A study in self-assessment: tutor and students' perceptions of performance criteria, *Assessment & Evaluation in Higher Education*, 22, pp. 357–368.

PIKE, G. R. (2014) Assessment measures: developing surveys of student engagement, *Assessment Update: Progress, Trends, and Practices in Higher Education*, 26 (1), pp. 9–11.

PIKE, G. R. (2015) Assessment measures: the CLA+, *Assessment Update: Progress, Trends, and Practices in Higher Education*, 27 (4), pp. 8–9.

RECEPOGLU, E. & ERGUN, M. (2013) Analyzing perceptions of prospective teachers about their media literacy competencies, *Education*, 134 (1), pp. 62–73.

REXWINKEL, T., HAENEN, J. & PILOT, A. (2013) Quality assurance in higher education: analysis of grades for reviewing course levels, *Quality & Quantity*, 47, pp. 581–598.

SADLER, D. R. (2010) Fidelity as a precondition for integrity in grading academic achievement, *Assessment & Evaluation in Higher Education*, 35, pp. 727–743.

SHAVIT, Y. & BLOSSFELD, H.-P. (1993) *Persistent Inequality: changing educational attainment in thirteen countries* (Boulder, Westview Press).

STENSAKER, B. & SWEETMAN, R. (2014) Impact of assesment initiatives on quality assurance, in: H. COATES (Ed) *Higher Education Learning Outcomes Assessment* (Frankfurt, Peter Lang), pp. 237–262.

STRUYF, E., ADRIAENSENS, S. & MEYNEN, K. (2011) Are beginning teachers ready for the job? The development and validation of an instrument to measure the basic skills of beginning secondary teachers, *Assessment & Evaluation in Higher Education*, 36, pp. 429–449.

SWEETMAN, R., HOVDHAUGEN, E. & KARLSEN, H. (2014) Learning outcomes across disciplinary divides and contrasting national higher education traditions, *Tertiary Education and Management*, 20, pp. 179–192.

TIMMERMAN, B. E. C., STRICKLAND, D. C., JOHNSON, R. L. & PAYNE, J. R. (2011) Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing, *Assessment & Evaluation in Higher Education*, 36, pp. 509–547.

TOTTEN, J. W. (2014) Appplication of the shortened version of the SOCO scale in a personal selling class, *Journal of Learning in Higher Education*, 10 (2), pp. 25–30.

TREMBLAY, K., LALANCETTE, D., & ROSEVEARE, D. (2012). "Assessment of higher education learning outcomes." *Feasibility study report*, 1.

VAN DEN WIJNGAARD, O., BEAUSAERT, S., SEGERS, M. & GIJSELAERS, W. (2015) The development and validation of an instrument to measure conditions for social engagement of students in higher education, *Studies in Higher Education*, 40, pp. 704–720.

WILLIAMS, S., DODD, L. J., STEELE, C. & RANDALL, R. (2015) A systematic review of current understandings of employability, *Journal of Education and Work*, pp. 1–25. doi: 10.1080/13639080.2015.1102210

WOLF, R., ZAHNER, D. & BENJAMIN, R. (2015) Methodological challenges in international comparative post-secondary assessment programs: lessons learned and the road ahead, *Studies in Higher Education*, 40, pp. 471–481.

YORKE, M. (2009) Grading student achievement in higher education: measuring or judging? in: M. TIGHT, K. H. MOK, J. HUISMAN & C. MORPHEW (Eds) *The Routledge International Handbook of Higher Education* (New York, Routledge), pp. 211–223.

YORKE, M. (2011) Summative assessment: dealing with the 'measurement fallacy', *Studies in Higher Education*, 36, pp. 251–273.

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, pp. 393–411.