Pearce *Empirical Research in Vocational Education and Training* (2015) 7:1
DOI 10.1186/s40461-015-0016-6

Empirical Research in Vocational Education and Training
a SpringerOpen Journal

RESEARCH Open Access

Assessing vocational competencies in civil engineering: lessons from AHELO for future practice

Jacob Pearce

Correspondence:
jacob.pearce@acer.edu.au
Assessment and Reporting (Mathematics and Science), Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell 3124, Australia

Abstract

Background: There has been much interest in the notion of vocational competencies of late. The pressing question, however, is how to measure vocational competencies. In order to contribute to this 'how', the results of the Civil Engineering strand of the recent Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study are discussed. The focus of the assessment instrument was explicitly on "above content" areas of the domain, and the assessment framework provides an example of a construct which articulates certain domain-specific competence components. This paper draws specific attention the Constructed Response Tasks (CRTs) in the instrument and discusses the results of the study with a focus on the lessons to be learnt from AHELO for future practice.

Methods: The assessment framework for the Civil Engineering strand was developed in line with the worldwide move to increase focus onto graduate competencies. The CRTs aimed to engage students with interesting, real-world problems. The tasks were geared towards finding out whether students could think like an engineer and display the non-technical competencies required of practicing engineers. Some items and the scoring rubrics are examined.

Results: In the Civil Engineering strand, 9 countries participated comprising 92 universities and 6,078 students. Content and construct validity were arguably achieved, as was reasonable reliability and good inter-rater scorer reliability. Overall, the cohort of engineering students worldwide found the tasks too difficult. Surprisingly, there was a large proportion of "zero" scores for the CRTs. It seems that either many final year engineering students are lacking in many of the vocational competencies required of them as they enter the workforce, or the experts and test developers had unrealistic expectations of them. It is difficult to ascertain whether the items were truly too difficult or if there were motivational and/or test targeting issues at play, especially due to the low-stakes nature of the implementation.

Conclusions: As the AHELO study shows, this type of undertaking is significantly complex. There is much room for improvement for future large-scale competency based assessments of this kind. In designing future work for measuring vocational competencies, lessons learnt from AHELO should be duly considered.

Keywords: AHELO; Civil Engineeringp; Engineering; Competence; Assessment; Measurement; Validity; Reliability

Background

The measurement of vocational competencies

There has been a great deal of interest in the notion of measuring vocational competencies of late. This trend has been driven mainly by the distinct lack of reliable data on the quality of graduates in Higher Education (HE), in contrast to the well-established large-scale international assessment programs at the school level (such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMMS)). The gap in an evidence base at this level has prevailed partly due to the fact that the higher education landscape is replete with a diversity of programs and contexts. This makes the assessment and indeed, measurement of competencies in this sector challenging, and the likely success of undertakings to do so, questionable.

The practice of assessing (and conceptualizing) vocational competencies has garnered further support thanks to the momentum gathered by an initiative funded by the German Federal Ministry for Educational and Research (BMBF). The research program "Modelling and measuring competences in Higher Education" (KoKoHS) was launched in late 2010 and directs much needed attention to this field of research (Blömeke et al. 2013). Although KoKoHS is concerned with the assessment of academic and scientific competencies, its momentum has directed attention to issues concerning vocational assessment approaches.

The pressing question, however, is how to go about measuring vocational competencies. At the school level, many international projects are designed and developed according to tried and tested procedures, which are followed to ensure validity and reliability in assessment. And yet, a recent study by Winther and Klotz (2013) showed that the reliability of current assessment practices in Vocational Education and Training (VET) economic domains were seriously lacking. Applying Item Response Theory (IRT) Scaling to 1,768 final examinations, they found that most often,

> the assessment entails not the intended, process-oriented structure but rather a fractured, subject-specific, content structure. This content-related structure model reflects a previous, officially abolished teaching structure and curriculum, which makes it quite surprising that this conceptualization still dominates the test. The instrument may be partially valid for assessing subject-specific content—that is, the expertise of a student in several subjects—but it cannot capture true action competence (2013, 9).

Two of the five proposals offered by Winther and Klotz to improve instruments in their domain of inquiry seem particularly generalizable and useful when conceptualizing an assessment of competence in the VET or HE sectors. They propose: (i) "Offering adequately authentic and complex test situations, such that the process-oriented, situated item setting aims to model real-life, authentic situations" and (ii) "Adopting a competence model that better depicts the development of competence through the learning process, moving from general competences (domain-related) to more specific competence components (domain-specific)" (2013, 10).

Moving away from content-based assessment to process-based tasks is the most straightforward means of moving towards the assessment of competence rather than of knowledge. In VET contexts, the most valid and reliable means of determined whether

a student has attained a desired competency would presumably be achieved by actually observing the student completing a technical task (such as welding) in an authentic scenario and evaluating the quality of the work they have done. However, in the context of international generalizable assessment, implementing such a program of assessment would be improbable to say the least.

However, process-based assessments where students are given the opportunity to demonstrate that they can apply their knowledge in real-world scenarios offers some middle ground for consideration between content-based tasks and actually demonstrating a learned skill. The two suggestions given by Winther and Klotz are at the core of what the OECD's Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study set out to achieve in the field of Civil Engineering. The AHELO Engineering framework articulates the concept of competence in terms of "Engineering Proficiency" (OECD 2012a). This construct is defined and ways of measuring it are offered with a focus on a process-based measurement instrument, coupled with a smaller proportion of content-based tasks to establish basic scientific grounding in the discipline.

The relevance of AHELO

This paper discusses the results of the Civil Engineering strand of the recent Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study, commissioned and completed by the Organization for Economic Co-operation and Development (OECD). This study was unique in scope and intent; the first of its kind to design and develop a robust approach for measuring learning outcomes of graduates worldwide. The aim being to achieve a valid and reliable means for measuring competencies across the diversity of cultures, contexts, languages, institutions and programs. AHELO is an interesting case study, particularly in the context of thinking about the measuring vocational competencies in an international setting.

The focus of the AHELO Civil Engineering assessment instrument was explicitly on the "above content" areas of the domain, and the assessment framework articulates certain domain-specific competence components. "Above content" means focusing on the ability of students to extrapolate from what they have learned and apply acquired competencies in new and unfamiliar contexts. The move to focus on an assessment pitched 'above content' stemmed from the concern that making comparisons across institutions, countries and contexts would be difficult with so many discipline related idiosyncrasies across the globe.

The instrument was developed with a strict mapping to the above content components and the items exemplify a way of measuring domain-specific competencies that are process-oriented, in authentic scenarios and contexts, all the while maintaining rigorous quality assurance processes and a focus on validity and reliability in assessment. This paper will specifically draw attention the Constructed Response Tasks (CRTs) in the engineering instrument. The CRTs were a module of assessment, which introduced an authentic context with a variety of stimulus information. Approximately 6–8 items followed, drawing on the context and building on different competencies with differing levels of sophistication. They were 'constructed-response' in the sense that candidates had to write detailed responses rather than select multiple-choice type single best answers. The paper then discusses the results of the study with a focus on the lessons to be learnt from the AHELO for future practice.

Although the study was initially referred to as a "PISA for Higher Education" (Tremblay et al. 2012, 56), the internationalizability of the processes and procedures of such an undertaking in a Higher Education setting remained the central focus of the research. There were many rationales for the study, which have been detailed elsewhere (Coates and Richardson 2011; OECD 2009; OECD 2010a; OECD 2010b). The main issue in the AHELO project was the notion of feasibility. Specifically, the project asked whether it is scientifically possible to produce, and feasible to implement, valid and reliable cross-linguistic, cross-cultural and cross-institutional comparisons of HE learning outcomes. All decisions on domains, frameworks, instruments, implementation, analysis and reporting were guided by this overarching ethos.

Although AHELO was administered and deployed in the HE sector, its relevance and connection with aspects of VET should not be forgotten. The lines between VET and Higher Education have become increasingly blurred. This is true especially of disciplines such as engineering, where there are divergent pathways to similar careers. In Australia, for instance, a civil engineer usually arrives in the professional through a university pathway, although increasingly, there are ways into the discipline from VET programs. Indeed, the more technical aspects of Civil Engineering typically require a great deal of practical, hands-on training. The question: "what can civil engineers *do* as they are prepared to enter the workforce?", is a question that is relevant to both the VET and HE sectors internationally. The lessons from the AHELO experience are therefore not only relevant to VET, but should inform future work in the area of measuring vocational competencies.

Methods

Design of the AHELO Feasibility Study

AHELO was a worldwide collaborative effort, with 17 countries (or more accurately, 'systems', as some were smaller provinces or regions of countries) participating in the development and validation of assessments from the three strands: Generic Skills, Economics and Civil Engineering. The design of the study was an iterative processes and involved much consultation and engagement with experts, governments, higher education bodies and institutions from across the globe. A contextual instrument was also developed, and administered in all three strands. This allowed rich contextual data to be analysed with the results and disseminated to participating institutions. The focus was on final year bachelor degree students, and the aim was to assess their capacity to apply their skills and knowledge to authentic, real-world problems associated with their specific discipline. Thus, the research question is whether it is possible to measure competencies relating to the practical work of a civil engineer at the end of higher education studies.

The feasibility study was commissioned and overseen by the OECD, after being conceived in 2007. The overall structure and strategy of AHELO, and how the Civil Engineering domain is embedded in the whole AHELO approach is detailed elsewhere (Tremblay et al. 2012). The operational work was undertaken by an international consortium, led by the Australian Council for Educational Research (ACER). A qualitative testing phase (Phase 1) ran from 2010 to 2011, and a larger-scale implementation with quantitative testing (Phase 2) ran from 2011 to 2012. (See Tremblay et al. 2012; OECD

2013a; OECD 2013b). The notion of value-added measurement was not part of Phase 1 or 2 operations, but has gained traction in the review of the feasibility study and is likely to play a part if an AHELO main study goes ahead (Braun 2013).

Civil Engineering was deemed to be a good domain for AHELO, being a scientific and professional discipline which enough commonality across cultures and contexts to assess the feasibility of the undertaking. Increasingly, civil engineers are expected to be able to transfer their skills across continents. Although the specifics of local environments, practical constraints and professional codes may vary across contexts, the basics of engineering sciences are fairly homogenized. And even is specific engineering codes differ across borders, the science underpinning the codes does not. In terms of generic skills for engineers, engineering education covers a broad range of competencies such as effective communication, good team-work, ethical and professional conduct, and so on. (Bons and McLay 2003; Walther et al. 2005; Gill et al. 2005).

Civil Engineering Assessment Framework

Following a common trend in engineering education, the assessment framework for the Civil Engineering strand was developed in line with the worldwide move to increase focus onto graduates' capacity to communicate effectively, to display ethical understanding and exercise professional judgement in authentic real-world type problems (Boles et al. 2006; West and Raper 2003). Engineering curricula across the globe are often articulated now in terms of learning outcomes and acquired competencies, as national engineering accreditors often require institutions to design and articulate their program outcomes in this way. Further, much international work has been done to date which finds significant commonality across borders between the articulation of engineering competencies. (Washington Accord 2013; European Network for Accreditation of Engineering Education (ENAEE) 2008; USA Accreditation Board for Engineering and Technology, ABET 2009; Engineers Australia (EA) 2011; UK Quality Assurance Agency (QAA) 2010; and EU Tuning Process (Tuning Project 2004)).

The AHELO Engineering Assessment Framework was developed through an iterative and collaborative process (OECD 2012a). It was guided by research, consultation with educators and industry, and other accreditation documentation. It was informed by the practice and processes of the Tuning process, and based initially on the AHELO-Tuning document (OECD 2009), the Tertiary Engineering Capability Assessment Concept Design (Coates and Radloff 2008) and several symposia. A provisional framework, which reflected international consensus regarding key competencies, was finalized and delivered to the OECD in May 2012.

By considering substantive, technical and practical considerations for the development of an assessment instrument to measure engineering student's competencies, the framework defines the domain to be tested. The overarching construct to be measured is Civil Engineering "proficiency", which is defined as the

> … demonstrated capacity to solve problems by applying basic engineering and scientific principles, engineering processes and generic skills. It includes the willingness to engage with such problems in order to improve the quality of life, address social needs, and improve the competitiveness and commercial success of society (OECD 2012a, 5).

Civil Engineering proficiency for final-year bachelor students is further broken down, and defined as the demonstrated capacity to solve problems by applying (i) analysis using basic engineering and scientific principles, (ii) engineering design and (iii) engineering practice skills. The skills are supported by generic skills which are assessed through the generic skills component of AHELO. Figure 1 (OECD 2012a) illustrates the key components of the Assessment Framework.

The schematic of the framework allowed items to be developed in accordance with specific proficiencies articulated under the different branches. A mapping between items and competencies ensured that the instrument could be designed to measure distinct aspects of engineering competence. Each component of the framework as depicted in Figure 1 was further broken down into more detailed domain specific competence components. There are two competencies under Engineering Generic Skills, seven competencies under Basic and Engineering Sciences, six competencies under Engineering Analysis, two competencies under Engineering Design, and six competencies under Engineering Practice.

Although the competencies are domain specific, they remain necessarily generic in a sense. This is perhaps due to the fact that the competencies needed to encompass Civil Engineering aspects from across the globe. Nevertheless, each competency can be grounded in a specific case or scenario. To explain further, Engineering Practice (iv) is "The ability to demonstrate understanding of the non-technical implications of engineering practice and commitment to professional ethics, responsibilities and norms of engineering practice". No specific non-technical aspects or ethical codes are mentioned, but it is assumed that given enough information an engineering graduate should be able to demonstrate appropriate awareness of these issues. For instance, in CRT16 students are asked to consider a scenario in which results of a geotechnical analyses

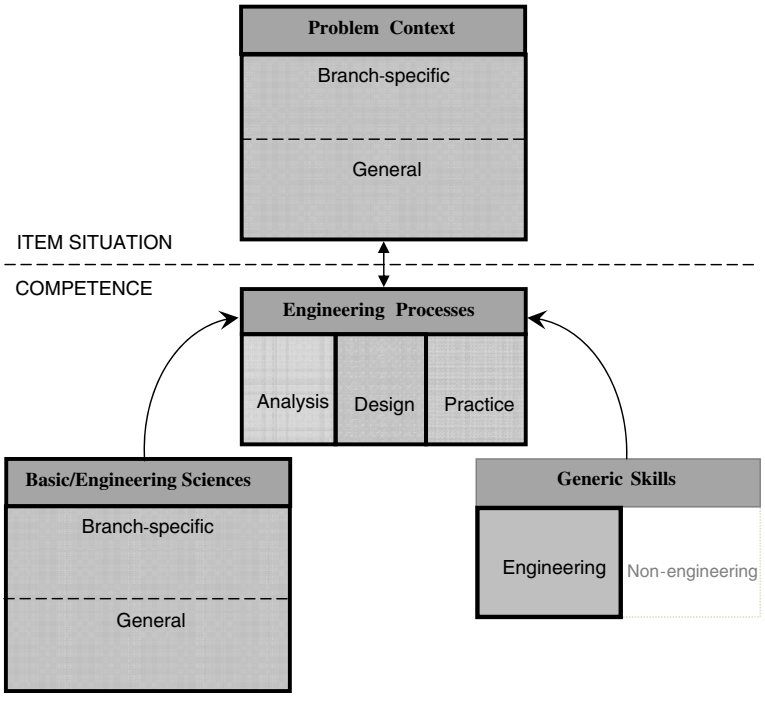

Figure 1 AHELO engineering assessment framework: key components. (OECD 2012a).

indicate that there is serious slope failure risk, but with low probability of occurrence, before constructing a dam to be built above a village. The student must discuss two aspects of engineering practice that favour re-designing the dam, even if this will incur greater cost (for the full item and scoring rubrics, See Tremblay et al. 2012, 261–263).

Another example is Engineering Analysis (iii): "The ability to select and apply relevant analytic and modelling methods"; even with no specific methods articulated, a graduate should be able to apply relevant methods in an authentic and unfamiliar context, again, given enough contextual information. This allows for specific scenarios or cases which map to these competencies to be introduced in a CRT. Designing the framework in this way emphasized that engineering problems occur in a diverse array of contexts. The next step was to design a sample of engaging contexts for items to assess the acquisition of the constituent components of engineering competency.

The tasks

The development of the engineering instrument has been discussed extensively by the OECD (2011b); Hadgraft et al. (2012); Tremblay et al. (2012) and Pearce (forthcoming). Here, the focus is on the tasks themselves. The Phase 2 assessment instrument was set at 90 minutes in length. Each student as take 25 Multiple Choice Questions (MQCs) which mapped to the Basic and Engineering Sciences component of the framework, and one Constructed Response Task (CRT), which mapped to the components under Engineering Processes. The CRTs introduced an authentic Civil Engineering scenario or context, and presented students with a set of information, followed by a set of items. As reporting was not designed to occur at the student level (only institutional and system level), 30 MCQs and 3 CRTs were administered to students through a rotated test design.

The MCQs were a fast and efficient way of collective data, and did not require human scoring. The CRTs are of more interest, as they arguably pushed the boundaries of competency assessment much further than the MCQs. The CRTs aimed to engage students with interesting real-world problems and issues that arise in the discipline. The items were geared towards finding out whether students could think like an engineer, perform appropriate tasks, and display the non-technical competencies required of practising engineers.

An example CRT is given in Tremblay et al. (2012, 252–264). This CRT begins with a presentation about the Hoover Dam in the U.S.A. Original construction plans are given, along with plans for the hydroelectric power station. The four items that follow ask students to identify aspects of the site that make it suitable for hydroelectric power generation; discuss design features that contribute to the dam's structural strength and stability; calculate the amount of water flowing through the turbines with given parameters; and explain environmental effects of the dam that an engineer would need to consider.

More stimulus information is then given, showing the disused Vajont Dam in Italy. Students are shown images relating to a landslide that occurred in 1963 which resulted in the death of around 2000 people in the valley below. The three items that follow ask students to explain geo-technical assessments that would have been done before the dam was constructed; ask them given reasons related to engineering practice supporting a possible decision made by an engineering team to redesign the dam (even at

greater cost) if they could reduce the risk of failure; and to outline town planning measures an engineer could have suggested to minimize risk to people.

The items which stem from the illustrative CRT above seems to meet the criteria given by Winther and Klotz, particularly in presenting students with an adequately authentic and complex test situation which is process-oriented. No strong technical knowledge is assumed and anything that an engineering graduate would have at hand to work through the items is effectively given. The goal is for the tasks to show minimal bias for different sub-groups in different countries. This is achieved by having all the tasks designed to see whether students can adapt to an authentic scenario and think like an engineer, when given enough background context. Further, the context surrounding dams is sufficiently generic for Civil Engineering that no one country is likely to have a greater predilection for such an emphasis in their curriculum.

As the CRTs required human scoring, rigorous scoring processes were designed along with detailed scoring rubrics. In developing the scoring processes around the AHELO engineering assessment, issues of inter-rater and intra-rater reliability were integral to ensuring the quality of scoring. More information relating to the scoring design and processes is given by Pearce (forthcoming) and OECD (2011c). Importantly, the Civil Engineering scoring guide could not anticipate all possible responses from students for each item in the CRTs. Thus, the aim was to define the different avenues for receiving score points, and to give example responses which were exemplary, along with ones which while not as eloquent, yet provided sufficient information to receive the score points.

To look at the rubrics in more detail, CRT17 is now briefly presented. The task asks: "After construction of the Vajont Dam, and recognising the possibility of hillside failure, outline two planning measures an engineer could have suggested to minimise potential harm to people." (Tremblay et al. 2012, 263). Although only two planning measures as requested, the scoring guide presents seven possible ways that the students could receive the two score points. This initiative accounted for the cross-cultural concern that different students in different countries would think differently about how to approach the item. The cross-cultural aspects emerged when thinking about these planning measures throughout the iterative process of scoring design. In one country, the norm would be to set up sophisticated evacuation procedures. However, in another country, relocating the entire town was seen as the best response. The construct did not demarcate which option was best, as long as two planning measures were given, as this demonstrated the student's capacity to apply acquired competencies in a manner deemed 'above content'.

The final seven categories for CRT17 are: (a) Evacuation procedures; (b) Town planning; (c) Town evacuation; (d) Monitoring; (e) Operation; (f) Communication plan (public education) regarding the risks; and (g) Strengthening/reinforcement of the dam wall and/or hillside. Scoring teams in the participating countries, guided by a lead scorer, underwent scorer training to ensure that they were applying the rubrics correctly and consistently (Pearce forthcoming; OECD 2011c). Each of the seven categories in CRT17 included a description of what the student response had to "capture" to be a sufficient response, and then gave example responses that did this. Scoring teams were also given a larger training module which included more example responses that had pre-assigned scores.

For (a) Evacuation procedures, the description is: "Refers generally to the implementation of evacuation procedures/warning system (can include communication of these procedures)". An example student response which satisfies this is given: "Implement a warning system, whereby increased geological movement could trigger an evacuation alarm so that the town is warned of the possible disaster. Such warning is commonplace for large earth and rock-fill dams for instance." For (e) Operation, the description is: "Refers generally to changing the utilisation (operation) of the dam (including abandoning/dismantling the dam)". One of the example student responses satisfying this is given: "Implement a system whereby the dam capacity is reduced prior to the wet season (the likely period of failure)". (Tremblay et al. 2012, 264). A broad range of responses were therefore permissible under this design.

Implementation

The five countries that participated in Phase 1 expanded to nine 'systems' for Phase 2. These were: Australia, Canada (Ontario province only), Japan, Colombia, the Slovak Republic, United Arab Emirates (Abu Dhabi only), Egypt, Mexico and the Russian Federation. Participating systems were involved in the revision of the source version through feedback given by National Project Managers (NPMs) and domain specific experts. At this stage, minor modifications or 'enhancements' were made to the source version of the test instrument at this stage (particularly the scoring rubrics, to ensure completeness). The source version was subjected to a rigorous translation and adaptation process for all languages of implementation (English, Japanese, Spanish, Slovak, Arabic and Russian) (OECD 2011b; Pearce forthcoming).

Finally, the translated instrument was prepared for delivery through an online testing platform. This system was purpose built and took many months to perfect. It was able to handle all the alphabets of testing, and was required to operate seamlessly and securely. Candidates were given individual logins under invigilation, and undertook the assessment in lecture theatres or computer laboratories at designated times. Their responses were securely hosted in an external data servers, and their test login was close down once they had completed the test. Their responses were then collected and ready for scoring by the scoring teams.

Results and discussion

17 systems took part in the entire AHELO Feasibility Study, along with 248 Higher Education Institutions (HEIs) and approximately 23,000 students. In the Civil Engineering strand, 9 systems participated along with 92 HEIs and 6,078 students.

Validity and Reliability

Prior to analysing the data for the Civil Engineering strand, items that did not meet psychometric standards were deleted in accordance with standard practice (OECD 2012b). After calibration by Rasch modeling (Rasch 1960) and other traditional item analysis methods (including goodness-of-fit, item facility, item discrimination, point-biserial correlations, item characteristic curves, and differential item functioning) several items were removed from the analysis due to exhibiting bias. This is further elaborated in OECD 2013a. Of the 30 MCQs and 3 CRTs (comprising 6 to 8 items), between two and eleven

items were removed from different systems. However, most of these were MCQ items (OECD 2013a, 17).

Exploratory and confirmatory factor analyses, along with result of analyses of fit statistics indicated that the MCQs and CRTs in the engineering instrument scale well onto the one-dimensional model, implying that the instrument was successful in measuring a common construct of 'Engineering Proficiency' (OECD 2013a, 23). It became clear that due to practical constraints (90 minutes of testing, etc.), and the breadth of the construct for measuring engineering competencies across the categories (Basic and Engineering Sciences, Engineering Analysis, Engineering Design, Engineering Practice and Engineering Generic Skills), more research would be required to uncover whether the construct could be reported according to these demarcations. As more than one competence dimension is assumed in the framework, it is disappointing that the validity of the construct could not be explored further. The main reason for this was that the focus of AHELO was to see whether the entire project was feasibility—an AHELO main study should place more attention on aspects such as validating dimensions of the framework.

Nevertheless, indicative psychometric analysis indicated that it would be possible to report according to these subsidiary sub-scales (OECD 2013a, 24). Given the constraints of the project (especially the overarching focus on *feasibility*), the number of items across the competencies were insufficient to run a detailed analysis of this feature of the construct. However, this highlights scope for future work in improving the construct validity of the Civil Engineering strand.

Content validity of the instrument was achieved through expert consensus in the iterative development of the design and development of the Engineering Assessment Framework, the instrument itself, and the comprehensive design and deployment of the scoring guide. Further content validity was gained through the positive student feedback. With the CRTs in particular, more than half of the students agreed, or strongly agreed, that the CRT tasks covered topics relevant to their programs. They also found the tasks "challenging, interesting, clear and comprehensive", and "very much related to the real world" (OECD 2013a, 24). It is true that these indicators are measures of acceptance, rather than scientific data validating the construct. Again, construct validation should play an integral part of future AHELO incarnations as there is indeed 'scope for further improving the construct validity of all three instruments' in AHELO (OECD 2013a, 24).

It is important to note that there was no large-scale trialling of the AHELO items. In a project such as PISA, the items would have been extensively field trialled prior to the instantiation of a main study. As AHELO was in a feasibility phase, no large-scale trialling was possible. Only small-scale qualitative data was gained in Phase 1. Thus, Phase 2 can be seen as a field trial of the items prior to final calibration of item statistics. Keeping this in mind, the final reliability of the instrument, with plausible values, was reported as 0.75 (OECD 2013a, 27). Although this falls below the recommended technical standards, it is comparable with other large-scale assessments of this kind and is remarkable considering that comprehensive psychometric data on the items was unavailable prior to Phase 2 testing.

Inter-rater scorer reliability for the CRTs in the engineering strand show that the percentage absolute agreement statistic sits around 80%, which is considered "fair to good"

(OECD 2013a, 28). A cross-country analysis of inter-scorer reliability was conducted, but the results must be "considered with caution" due to the limited amount of data. Due to this constraint, this data was not reported. Nevertheless, this study indicated that "it is feasible to score constructed-responses in a reliable way across countries, given then conditions such as the scoring rubric clarity, appropriate training and ongoing monitoring of reliability, are all put in place." (OECD 2013a, 29).

Item Difficulty and Targeting

As a whole, the cohort of engineering students worldwide (n = 6078) found the assessment items too difficult. Figure 2 shows an item map, with the distribution of the cohort on the left (marked as 'x's), and the item difficulty (logit) locations based on Rasch estimates on the right (OECD 2013a, 187). A candidate located at the same location as an item has a 50% probability of answering the item correctly. If they are located above the item location, the likelihood of answering the item correctly increases. Conversely, if they are located below the item, the likelihood of answering the item correctly decreases. If this data is treated as a field trial, then this information can be used to more adequately target the items to the cohort for a main study.

The item map indicates that several of the CRT items in particularly were far too difficult. There were consistent issues regarding difficulty with the CRTs across all participating countries. Some particular items were located between one and two and a half logits (item difficulty steps) away from the peak of the cohort distribution. For example, an average student would only have about a 8 percent probability of answering CRT35 correctly. Unfortunately, it is not possible to determine the technical features of this item as CRT3 has not been released.

Surprisingly, there was a large proportion of "zero" scores for the CRTs. Students received a score of zero if they attempted a task, but received no score points. This is different to a score of "missing", which represents an empty response where students have obviously not attempted the task at all. The range of zero scores was between 20 and 70 percent, indicating that the tasks were too challenging for a great deal of students (See Figure 3 (OECD 2013a, 188)). Figure 3 also shows that the range of zero scores for CRT2 was between 40 and 70 percent which is disappointing to say the least. Furthermore, in countries four and five, more than 40 percent of students obtained zero scores for any CRT attempted.

One possible alarming conclusion from these results is that many final year engineering students are severely lacking in many of the competencies required of them as they enter the workforce. However, it is equally possible that there was a mismatch between the Engineering Expert Group's expectations of students, and their actual abilities. The poor test targeting as illustrated in Figure 2 supports this latter conclusion. However, it seems that there were other factors at play.

Questions of Motivation and Effort

In reality, it is difficult to ascertain whether these items were truly too difficult or if there were motivational issues at play, especially due to the low-stakes nature of the assessment implementation. It could be that Figure 2 is not an accurate representation of the cohort's true capabilities. The low-stake environment appears to have influenced

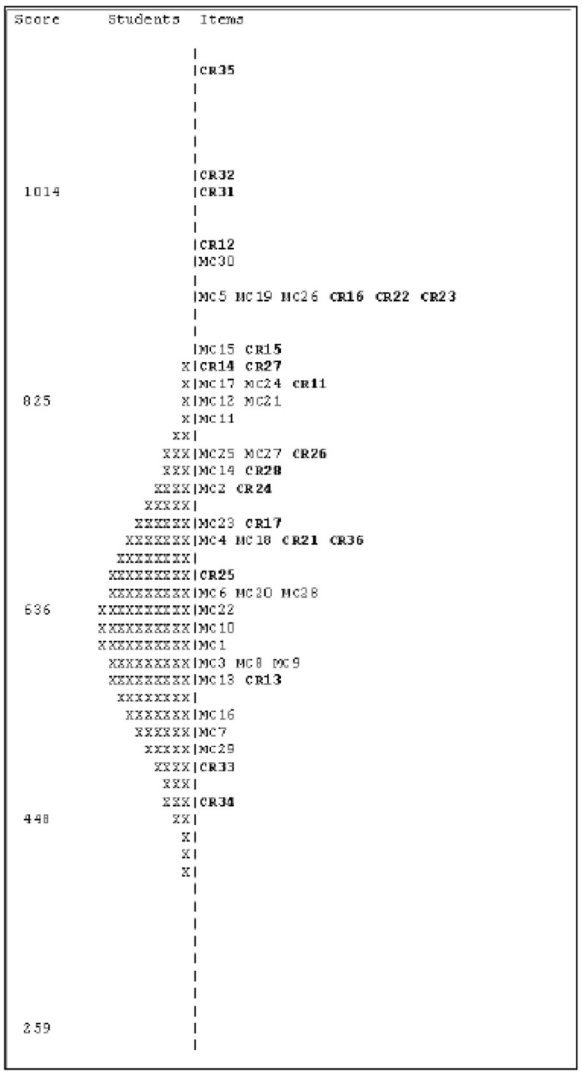

Figure 2 Item map engineering strand (n = 6078). (OECD 2013a, 187).

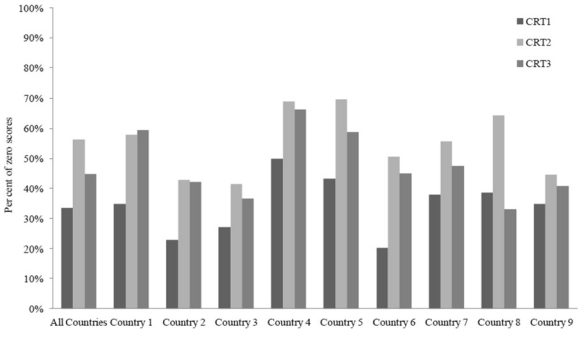

Figure 3 Engineering assessment zero scores, by CRT and country (n = 6078). (OECD 2013a, 188).

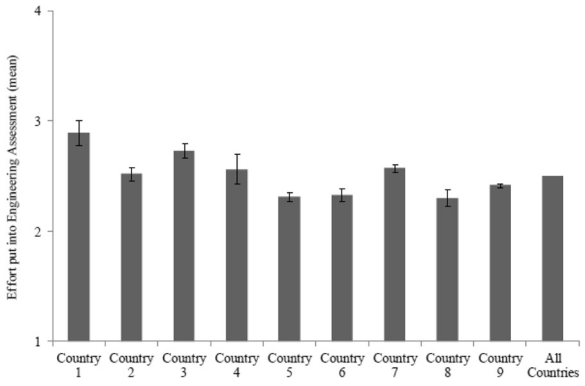

Figure 4 Self-reported effort put into the engineering assessment by country (n = 6078). (OECD 2013a, 180).

the results. The case for this claim can be made further by investigating the influence on the effort put into the assessment by students. Figure 4 (OECD 2013a, 180) shows that student self-reported effort into the instrument was only moderate. Self reporting a value of one indicated that a student put 'little or no effort' into the instrument; two represented 'some effort'; three represented 'close to my best effort'; while a four represented 'my best effort'. These results are surprising considering the fact mentioned earlier, that more than half of students across all countries felt that the CRT tasks were relevant to their programs and interesting and clear. But they also did note that they found the tasks "challenging" (OECD 2013a, 24)—a fact which, when coupled with the reality of undertaking the tasks in a low-stakes environment, may have resulted in the low levels of effort.

The CRTs were designed and developed through a long, iterative process of revision and consultation. The Engineering Expert Group deemed the items to be adequate to measure the competencies of graduating civil engineers across cultures and contexts. The items were mapped to the framework, developed to measure the anticipated spectrum of student ability, and designed to be pitched above curriculum content, emphasizing the process-oriented structure of competencies. The fact that students on the whole found it challenging to demonstrate the capacity to think like an engineer and apply their skills onto authentic real-world contexts is concerning, regardless of whether these results were influenced by motivational considerations.

Conclusions

Lessons from AHELO for future practice

The AHELO Civil Engineering instrument provides an example of a move away from content-based assessment to process-oriented tasks. For VET contexts, this type of assessment could be adapted to measure vocational competence rather than knowledge, without moving towards an observational assessment instrument of students performing a technical task. An assessment of vocational competencies of this kind has the potential to be instigated in international contexts. However, as the AHELO study shows, this type of undertaking is neither trivial nor straightforward.

The CRTs provided process-based tasks where students can demonstrate their ability to apply their knowledge and skills to real-world scenarios. These tasks were distinct

when compared with MCQs in that students could be more creative in their responses, and the scoring rubrics were built to ensure that "thinking like an engineer" would be rewarded appropriately. Thus, the assessment instrument is a clear case of an internationalizable outcomes assessment of Civil Engineering proficiency.

One of the most important lessons to draw from the AHELO experience, is that full scale field trialing is essential to determine the correct anticipate difficult of items and target the test appropriately. The AHELO Technical Advisory Group (TAG) concluded that while the CRTs "were of high technical quality, they were simply too difficult for many students to effectively engage and perform well." Further, "more time for piloting and field trials" may have revealed this issue "in time for it to be rectified." (OECD 2013a, 169). If Phase 2 of AHELO is seen as a field trial, then the data produced provides statistical and psychometric information sufficient to ensure the validity and reliability of a main study. If a full-scale AHELO goes ahead, rigorous trialing of the items is a must.

Although there were challenges in designing and developing the CRTs (notwithstanding the added complication of ensuring that the items could be delivered in multiple languages and contexts), the OECD report that experts felt that the CRTs were a positive step in competency assessment, in that they assessment higher level skills and provided a more comprehensive instrument. (OECD 2013b, 44).

However, producing process-oriented items of this type was a time consuming and expensive process. If these types of items are implemented in smaller-scale projects to measure vocational competencies, some of the added constraints may not apply (translation and adaptation, for instance). However, the amount of input required to develop sound scoring procedures and scoring rubrics—as well as the investment and infrastructure required to actually run the scoring—should not be underestimated.

By demonstrating that it is indeed feasible to develop and implement instruments with valid and reliable results across the diversity of countries, languages, cultures and institutional settings, the AHELO Feasibility Study achieved its main goal. However, there is much room for improvement for future large-scale competency based assessments of this kind. In designing future work for measuring vocational competencies, the lessons learnt from AHELO should be duly considered.

Competing interests

The author is an employee of the Australian Council for Educational Research, the organisation commissioned by the OECD to managed the AHELO Feasibility Study. The author was specifically involved in the development and implementation of the engineering strand.

Acknowledgements

I would like to thank Daniel Edwards for feedback on an earlier version of this paper. I would also like to thank the OECD for allowing this publication to go ahead. I am grateful to all colleagues who made the Feasibility Study possible (especially Hamish Coates, Sarah Richardson, Julian Fraillon, Daniel Edwards and Roger Hadgraft), and to all staff and students across the globe who participated.

Received: 20 August 2014 Accepted: 20 January 2015
Published online: 12 February 2015

References

ABET (2009) Criteria for accrediting engineering programs: effective for evaluations during the 2010–2011 cycle. Engineering Accreditation Commission. http://www.abet.org/uploadedFiles/Accreditation/Accreditation_Process/Accreditation_Documents/Archive/criteria-eac-2010-2011.pdf. Accessed 3 Feb 2015

Blömeke S, Zlatkin-Troitschanskaia O, Kuhn C, Fege J (2013) Modeling and measuring competencies in higher education: Tasks and challenges. In: Professional and VET learning: Vol. 1. Sense Publ, Rotterdam

Boles W, Murray M, Campbell D, Iyer M (2006) Engineering the learning experience: Influences and options. Proceedings of the 17th Annual Conference of the Australasian Association for Engineering Education, Auckland, 10–13 Dec

Bons W, McLay A (2003) Re-Engineering Engineering curricula for tomorrow's engineers: Engineering education for a sustainable future. Proceedings of the 14th annual conference for Australasian Association for Engineering Education and 9th Annual Women in Engineering Forum, Melbourne, 29 Sept-1 Oct

Braun H (2013) Value-added measurement: report from the expert group meeting. In: OECD. Assessment of higher education learning outcomes ahelo: feasibility study report, volume 3. further insights. OECD, Paris

Coates H, Radloff A (2008) Tertiary engineering capability assessment: concept design. Group of Eight Universities, Canberra

Coates H and Richardson S (2011) An international assessment of bachelor degree graduates' learning outcomes. Higher Education Management and Policy. 23(3). doi:10.1787/17269822

Engineers Australia (2011) Stage 1 Competency Standard for Professional Engineer https://www.engineersaustralia.org.au/sites/default/files/shado/Education/Program%20Accreditation/110318%20Stage%201%20Professional%20Engineer.pdf. Accessed 20 July 2012

European Network for Accreditation of Engineering Education (2008) EUR-ACE® Framework Standards for the Accreditation of Engineering Programmes. www.enaee.eu. Accessed 16 Jul 2009

Gill J, Mills J, Sharp R, Franzway S (2005) Education beyond technical competence: gender issues in the working lives of engineers. Proceedings of the 4th ASEE/AaeE Global Colloquium on Engineering Education, Sydney

Hadgraft R, Pearce J, Edwards D, Fraillon J, Coates H (2012) Assessing Higher Education Learning Outcomes in Civil Engineering: the OECD AHELO Feasibility Study. Proceedings of the 2012 AAEE Conference, Melbourne, Victoria

OECD (2009) A Tuning-AHELO Conceptual Framework of Expected/Desired Learning Outcomes in Engineering. OECD, Paris, http://www.oecd.org/dataoecd/46/34/43160507.pdf. Accessed 12 Feb 2010

OECD (2010a) AHELO Feasibility Study Analysis Plan. OECD, Paris, http://www.oecd.org/edu/highereducationandadultlearning/ahelodocuments.htm. Accessed 23 Jul 2012

OECD (2010b) AHELO Assessment Design. OECD, Paris, http://www.oecd.org/edu/highereducationandadultlearning/ahelodocuments.htm. Accessed 23 Jul 2012

OECD (2011b) Engineering Assessment Development Report. ACER consortium and AHELO consortium, Group of National Experts on the AHELO Feasibility Study, 8th Meeting of the AHELO GNE, Paris, 28–29 November 2011. http://www.oecd.org/edu/highereducationandadultlearning/ahelodocuments.htm. Accessed 15 Jul 2014

OECD (2011c) International Scoring Manual. ACER consortium and AHELO consortium, Group of National Experts on the AHELO Feasibility Study, 8th Meeting of the AHELO GNE, Paris, 28–29 November 2011. http://www.oecd.org/edu/highereducationandadultlearning/ahelodocuments.htm. Accessed 15 Jul 2014

OECD (2012a) Engineering Assessment Framework. ACER consortium and AHELO consortium, Group of National Experts on the AHELO Feasibility Study, 8th Meeting of the AHELO GNE, Paris, 28–29 November 2011. http://www.oecd.org/edu/highereducationandadultlearning/ahelodocuments.htm. Accessed 15 Jul 2014

OECD (2012b), AHELO Feasibility Study Technical Standards, ACER consortium and AHELO consortium, Group of National Experts on the AHELO Feasibility Study, 9th Meeting of the AHELO GNE, Paris, 19–20 March 2012. http://www.oecd.org/edu/highereducationandadultlearning/ahelodocuments.htm. Accessed 15 Jul 2014

OECD (2013a) *Assessment* of Higher Education Learning Outcomes AHELO: Feasibility Study Report, Volume 2. Data Analysis and National Experiences. OECD, Paris

OECD (2013b) Assessment of Higher Education Learning Outcomes AHELO: Feasibility Study Report, Volume 3. Further Insights, OECD, Paris

Pearce J (forthcoming) Ensuring quality in AHELO item development and scoring processes. In Musekamp F, Spöttl G and Saniter A (Eds) Competence Assessment in Engineering Science. Peter Lang Publishers, Reihe Berufliche Bildung in Forschung

Quality Assurance Agency (2010) Subject benchmark statement: Engineering. http://www.qaa.ac.uk/en/Publications/Documents/Subject-benchmark-statement-Engineering-.pdf. Accessed 3 Feb 2015

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research, Copenhagen

Tremblay K, Lalancette D, Roseveare D (2012) Assessment of Higher Education Learning Outcomes AHELO: Feasibility Study Report, Volume 1. Design and Implementation, OECD, Paris

Tuning Project (2004) Tuning educational structures in Europe: Generic competencies. http://www.unideusto.org/tuningeu/. Accessed 30 Apr 2009

Walther J, Mann L, Radcliffe D (2005) Global Engineering education: Australia and the Bologna Process. Proceedings of the 4th ASEE/AaeE Global Colloquium on Engineering Education, Sydney, 26–29 Sept

Washington Accord (2013) Graduate Attributes and Professional Competencies. http://www.ieagreements.org/IEA-Grad-Attr-Prof-Competencies.pdf. Accessed 3 Feb 2015

West M, Raper J (2003) Cultivating generic capabilities to develop future engineers: An examination of 1st year interdisciplinary Engineering projects at the University of Sydney. Proceedings of the 14th annual conference for Australasian Association for Engineering Education and 9th Annual Women in Engineering Forum, Melbourne, 29 Sept-1 Oct

Winther E, Klotz VK (2013) Measurement of vocational competences: an analysis of the structure and reliability of current assessment practices in economic domains. Empiric Res Vocational Educ Train 5:2, doi:10.1186/1877-6345-5-2