0949-149X/91 $3.00+0.00
© 2011 TEMPUS Publications.

Cross-National Evaluation of Learning Assessment in First-Year Engineering Students: U.S. Experience Applied at Two Universities in Chile*

T. C. SHEAHAN
Department of Civil and Environmental Engineering, Northeastern University, 400 Snell Engineering Center, 360 Huntington Avenue, Boston, MA 02115, USA. E-mail: t.sheahan@neu.edu

E. J. MASON
Department Counseling and Applied Educational Psychology, Northeastern University, 404 International Village, 360 Huntington Avenue, Boston, MA, 02115, USA

D. M. QUALTERS
Department of Education and Human Services & Director, Center for Teaching Excellence, Suffolk University, 8 Ashburton Place, 711 Stahl, Boston, MA 02114, USA

P. V. POBLETE
School of Engineering and Science, University of Chile, Av. Beauchef 850, Santiago, Chile

X. VARGAS
Civil Engineering Department, University of Chile, Av. Blanco Encalada 2002, Santiago, Chile

The need to demonstrate the quality of engineering student learning outcomes has intensified in recent years as the assessment movement has spread into engineering education, been widely adopted by accreditation agencies, and has been incorporated into international accreditation and curricular agreements. Developing quality standards and measuring learning levels can be difficult enough from campus to campus within a particular country. However, exporting and adapting such standards across countries and educational traditions and cultures is even more challenging. The present paper describes the portability of one system of assessment developed at a U.S. institution as applied to two engineering faculties in Chile. The process followed during the initial training and development provides some good insights into how assessment and quality improvement processes can be quickly implemented. However, the implementation required cultural shifts in how the program approached, and administrators and faculty thought about, assessment for quality improvement. The experience in Chile provides some lessons on the opportunities and potential pitfalls of carrying out such a process in other countries and programs.

Keywords: assessment; standards; engineering education; outcomes assessment; international standards; performance testing

1. Introduction

The difficulty of implementing one country's specific educational practices in other countries with different educational traditions and cultures has been demonstrated in a number of disciplines and settings [1–4]. The complexity involved in these efforts is intensified by language and translation difficulties, national and institutional cultural and political impediments, faculty customs and relationships (with each other, the administration and the students), traditions of interdisciplinary collaboration in areas such as course development, and prescribed curricular sequences and expectations. A key element of the success of the cross-national implementation of educational programs seems to be the degree to which communication and sharing of ideas occurs within the development of policy and procedures for launching and operating the program [2, 4, 5].

In engineering education, international agreements such as the Washington Accord imply that there should be identifiable standards that apply to the training that engineers receive on university campuses across the world. Such educational standards have to be assessed in some manner, leading to the need for new approaches to develop and implement effective assessment procedures. In the United States, engineering programs are instituting assessment frameworks to determine the quality of engineering training outcomes, driven in part by recent accreditation requirements involving quality improvement promulgated by the Accreditation Board for Engineering and Technology [6]. A similar interest in standards and assessment has spread across the European Union (EU) requiring universities to comply with the Bologna Agreement of 1999 (formally, the *Joint Declaration of the Ministers of Education* [7]). The impact is extending to other countries. The motivation for these develop-

* Accepted 8 May 2011.

ments is two-fold. First, accreditation agencies including ABET are embracing the idea that student learning outcomes need to be assessed and then used as feedback to improve courses, programs, and learning. Meanwhile, the Bologna agreement seeks to achieve improvement in training as well as establishing common outcomes among courses (part of the so-called 'tuning' process) to facilitate transportability of university credits across borders.

It is now common to see on websites of universities that are part of the Bologna agreement uniformly presented objectives and learning outcomes for courses offered (e.g., [8]), while most universities in the U.S. still have this as a goal rather than an accomplishment. The development of objectives and outcomes is a first step in establishing effective assessment instruments that can measure the learning achievement of students. However, engineering schools have little experience or expertise in designing and evaluating learning outcomes [9]. For a given course, this includes identifying key learning goals for students, converting those broad goals into assessable outcomes (including the level of accomplishment desired), and then identifying and developing the methods to be used to provide evidence that students are achieving the desired level of accomplishment for each learning outcome.

The process of establishing learning objectives and outcomes, assessing the level of achievement with valid instruments, and using the results for quality instructional improvement is not easily implemented without considerable training, reflection and commitment by the instructional faculty, and support for the analysis and use of the assessment results. These intra-campus issues become even more complex when moving a program to other campuses and national settings. A system for accomplishing this kind of assessment has been developed at Northeastern University (NU) in Boston, Massachusetts, U.S.A. [10]. The present paper describes the portability of the system of assessment developed at NU to two engineering faculties in Chile: the University of Chile, Santiago (UC) and the Pontifical Catholic University of Chile (PCUC). The process followed during the initial training and development provides insights on how the assessment and quality improvement processes can be quickly implemented. However, success required cultural shifts in how the programs approached, and administrators and faculty thought about, assessment for quality improvement. The experience in Chile provides some lessons on the process and potential pitfalls of carrying out such an implementation in other countries and programs.

2. Background

The meaning and intent of quality improvement in engineering and other higher education programs presented significant problems in semantics and philosophy in the Bologna Process, and continues to be an issue in both U.S. and non-U.S. higher education institutions. A basic problem, as Adelman [11, p. xxi] delineates, is distinguishing between 'information' and 'accountability.' While he argues that quality improvement processes and quality assurance ('information') are bigger than accreditation ('accountability'), at least in the ABET model, these are inextricably linked and a program must have a working quality improvement process to achieve accreditation. Perhaps a better explanation of Adelman's statement is that achieving a working quality improvement process requires more cultural change, more intellectual resources and more ongoing 'maintenance' than complying with periodic accreditation audits and fulfilling certain curricular requirements. This process requires discipline reviews of relevant courses, programmatic reviews of discipline-level inputs, agreed-upon statements of student learning objectives and outcomes, minimally acceptable performance standards, and documented improvement of courses and the program based upon the inputs.

ABET seems to have had more of an influence on engineering program accreditation and related quality improvement processes in Latin America than in Europe, perhaps because of the former region's proximity to the U.S. At the same time, while Latin America is certainly paying attention to ABET and Bologna, the Latin American countries are creating their own path toward regional agreements and the evolution of common competencies and degree requirements. As noted by Phillips et al. [12], ABET and UNESCO's Regional Office for Science and Technology for Latin America and the Caribbean signed a memorandum of understanding (MOU) in 1995, and ABET has held numerous workshops in Latin and South American countries. The 'Engineer of the Americas' (later the 'Engineer for the Americas') movement, started in 2003 [13], sought to link educational institutions to industry and to emulate in Latin America the organizational efforts that were occurring (and continue to occur) in Europe. Participant countries developed a set of common competencies with the goal of producing engineering graduates who can be at the core of economic development and technical proficiency. And while Latin American countries may look toward ABET for guidance, they are more likely to emulate Bologna since, like Europe, Latin America does not have an entity large enough to dominate or define engineering competencies for the region—it seems likely that, if Europe's experience is any guide, any agreement about core engineering competencies will involve compromising on differences. The 'Washington Accord' serves as another model for Latin America in which signatories agreed to establish substantial equivalency of their respective country's accreditation systems to assess proficiency of their programs' graduates to practice engineering.

The recent history of higher education accreditation and quality assurance in Chile is highlighted by the formation in 1999 of the National Commission for Undergraduate Education (CNAP; [9, 14]), and changed to the National Accreditation Commission (CNA; [15]) in 2006. With this change came the introduction of multiple accreditation agencies supervised by CNA. CNAP/CNA established accreditation criteria and procedures, and included the participation of faculty, professional societies and end-users of graduates' services in particular fields. This is similar to the development and ongoing modification of ABET criteria. Also similar to ABET, the accreditation involves self-evaluation of programs and external review by national and international peer teams [14].

These developments in Chile, the rest of Latin America and throughout the world have not resolved some important questions [16]: (1) Under what conditions do continuous improvement concepts and practices become part of faculty culture? and (2) How transferrable are accreditation criteria and processes across national borders? In addition, as highlighted by Patil and Codner [17], most accreditation models that have been implemented in various parts of the world are characterized as outcomes-based models; however, actually demonstrating and/or proving that engineering graduates have attained a respective program's attributes and competencies is a feat that still appears in 'various stages of development.' Letelier and Carrasco [9], referring specifically to this issue in Chile, stated that one of the weak areas in the assessment process in Chile is 'evaluation procedures of main activities.'

As in other countries, including the U.S., better instruments for assessment of student learning outcomes are needed to demonstrate these competencies, while striking the right balance between summative evaluation and formative feedback for quality improvement. A further challenge, as posed by Prados et al. [16], is instilling such a balanced process into faculty culture at an academic institution.

3. Implementing an assessment instrument at the two universities in Chile

3.1 Initial workshop with instructors

In January 2007, the U.S. consultants from NU traveled to Santiago to work with administrators and faculty from two engineering programs, the University of Chile (UC) and the Pontifical Catholic University of Chile (PCUC), to develop, implement and evaluate a comprehensive assessment instrument for engineering students at those institutions based on the model program implemented at NU [10]. The specific goals for the two-day workshop were as follows.

- To provide and use as a model the mastery exam that had been administered to College of Engineering freshmen at NU. The mastery exam at NU was originally developed over a one-year period that included initial training, question (or item) development by individual instructors in four departments and deployment on-line using course delivery software. While the exam is designed to be related to the learning goals of the NU curriculum, it provided a basis for comparison and an example to help develop a similar test based on the learning objectives of the specific curriculum of the two institutions in Chile.
- To provide background on the overall exam development process, the ABET context for adopting a more robust assessment framework, and the analysis of exam results using a calibration approach know as Item Response Theory (IRT) [18]. Qualters et al. [10] provided background on the use of IRT for interpretation of assessment instrument results. An important aspect of this endeavor included a discussion of the culture of the faculties in the two institutions and how they might differ, and how they differed from that of NU where the initial mastery exam had been developed. This discussion was considered important to the successful implementation of the process.
- To introduce human learning models that had potential for utility in helping to identify appropriate exam questions. The team knew from their experiences implementing the mastery exam at NU that many engineering faculty are not familiar with such models, so they felt that it was important to provide background on this topic in Chile as well. This included background on how to formulate learning outcomes through the questions asked of faculty: 'What does it mean to learn something?' 'What would they want their students to say they learned in a course one year after taking it?' The learning models presented included Bloom's Taxonomy [19], Shulman's table of learning [20], and the Fink Taxonomy of Significant Learning [21]. One of the difficulties that the team faced was in presenting models that are laden with specialized terminology to those whose first language is not English. This took additional consideration when preparing for

the workshops. The main approach to deal with terminology was to emphasize concepts and the process conscientiously and to avoid jargon that might be unique to speakers of American English.

- To guide instructors through the writing of course objectives and learning outcomes that were framed in the appropriate language for each category. Using the learning models cited in the previous section, but focusing primarily on Bloom [19], six categories of cognitive gain (or learning) were identified: (1) knowledge; (2) comprehension; (3) application; (4) analysis; (5) synthesis; and (6) evaluation. However, for practical reasons involving the difficulty in measuring the higher levels, the analysis, synthesis and evaluation levels were recommended to be combined into 'higher order thinking and problem solving.' It is noted that the writing of objectives and outcomes in the exam development process is always challenging for native English speakers, and it was even more so with instructors dealing with translation. This was certainly the case in Chile, where many of the nuances of objectives and outcomes were not easily translatable; however, this was facilitated somewhat by having a knowledgeable Spanish-speaking partner with us throughout the workshop who had spent considerable time in the U.S. and was familiar with the concepts in both languages.
- From the objectives and outcomes, to develop a blue print or *table of specifications* to define the domain of the test [22]. This two-dimensional table maps learning objectives with the levels of learning from the learning model. It is used to assign weights reflecting emphasis of specific content and levels of learning to course topics, thereby assuring that the test will then reflect the proportion of emphasis of the domain being tested. This approach, when done well, can contribute to the content validity of the test.
- To familiarize instructors with valid question (item) wording and syntax, as well as the development of the correct answer and distracters in multiple choice questions, and initiate question formulation. Group work was vital for this stage of the process: as items were developed, instructors were asked to share these with others in their discipline group to identify potential problems with the items such as misunderstandings by students, two or more correct answers, and poor or unclear phrasing. Again, language differences made this stage of the exam development challenging, and it was imperative to have a facilitator familiar with education terminology in both languages.
- To describe the framework for revising the items, and using the exam as feedback to instructors.

Based on the experience at NU, feedback to instructors based on the exam results was anticipated to be the greatest challenge in terms of altering faculty perceptions and culture regarding how they traditionally viewed an exam. The exam results must be shared with the disciplines involved, and those results explained. The manner in which this information is presented is critical, since it must be conveyed that the results are not intended as a judgment of the unit's teaching quality or even worse, personal criticism, but as a mechanism for more widespread improvement of student learning. In other words, it should be presented as a measure of what the students have learned without any message of blame. While the discussion may include teaching quality, the purpose of the feedback stage is for the responsible unit to take ownership of the exam results, and then take positive action to improve student performance. This can range from improving the exam items to altering the manner in which material is covered, to programmatic changes that will more adequately set students up for success. The exam results also provide a window on Ewell's [23] concept of a three-tiered curriculum: the one in the catalog, the one the faculty are teaching, and the one that students are actually experiencing. This is consistent with learning experiences designed by faculty, as described by Fink [21], versus student-centered learning advocated by Weimer [24]. Anecdotally, when compared with the experience at NU, in Chile there seemed to be a higher level of instructor awareness, bordering on anxiety, about the exam results and their use for feedback. In this case, it was very important to have an administrator provide unequivocal reinforcement that the results were intended to improve program delivery and teaching quality, and not to pass judgment on a faculty member's performance.

In addition to the workshop agenda segments, the team was available after the visit for responding to follow-up questions on developing the exam, and while the team analyzed the first data from the first administration of the exam using the MULTILOG IRT software package [25], the data were later reanalyzed by the UC/PCUC staff using IRT freeware [26].

3.2 Further development and deployment of the assessment instrument

After the workshop and the departure of the U.S. team, there was a joint committee formed from the two institutions to finalize the tables of specifications (TOS) that had been started during the workshop. While the TOS product is relatively simple in format (Table 1 provides an example), its development requires considerable reflection by the faculty—what are the major topics taught in a class,

Table 1. Example Table of Specifications [10]

Physics	Knows terms	Knows facts	Knows procedures	Comprehends principles	Applies principles	No. of items
Description of motion (velocity, acceleration, etc.)	×	×		× ×		2
Application of differential calculus to motion			× ×			1
Newton's laws	×	×	×	× × ×	× ×	3
Conservation of energy				× ×	× × ×	2
Conservation of momentum				×	× ×	1
Rotation of rigid bodies	×		×	× ×	× ×	2
Static equilibrium			× ×			1

and what weight should be given to each topic in an assessment instrument with a limited number of items? In the case of UC and PCUC, a TOS for each area to be tested (Chemistry, Computer Science, Math, and Physics,) was established to guide the item formulation. After this, items were developed and reviewed by the committee.

The exam consisted of 7 questions each in Chemistry and Computer Science, and 23 questions each in Math and Physics. An important point about the development of this particular exam is that a trial version of the exam was not tested prior to its initial administration to the students in the two programs. Such a trial can be valuable for resolving problems that inevitably cannot be foreseen by exam developers. Because the first exam administration was essentially the trial, problems arose that were addressed during the second exam administration. These issues are described in more detail in the next section.

A target group of students, consisting of those who had completed the common core at the two universities, was identified to take the first version of the exam. The exam was the same for both UC and PCUC. For UC students, the exam was mandatory for all those who had completed the core (n = 175 students), and for PCUC students, the exam was voluntary (n = 38 students). In both cases, the exam was administered with a proctor in the room.

3.3 Results from first exam administration

As noted previously, the exam data were analyzed by UC using free software [26] for applying IRT, and the results presented in terms of *item characteristic curves* (ICC). IRT is based on a logistic model that has as its inputs the student responses to each item on the exam, and the output is a probability (P) that a randomly chosen student with an ability level θ in a topic area will correctly answer a particular item. The ICC plots P versus θ for a particular item [27].

Figure 1 shows four conceptual ICCs, based on a three-parameter model, to illustrate graphically the significance of the item characteristics [10]. For an item i, the parameter a_i quantifies the item's effectiveness at discriminating among abilities, and is the slope of the ICC where it crosses the $P(\theta)$ = 50% level. Referring to Fig. 1, the two solid line graphs (marked ① and ②) have relatively high slopes at $P(\theta)$ = 50%, indicating that these two questions are better discriminators of student ability compared with the dot–dash line (marked ③). The parameter b_i is the ability level θ where the ICC crosses the $P(\theta)$=50% level, and is used to represent item difficulty level. In Fig. 1, the two 'good discrimina-

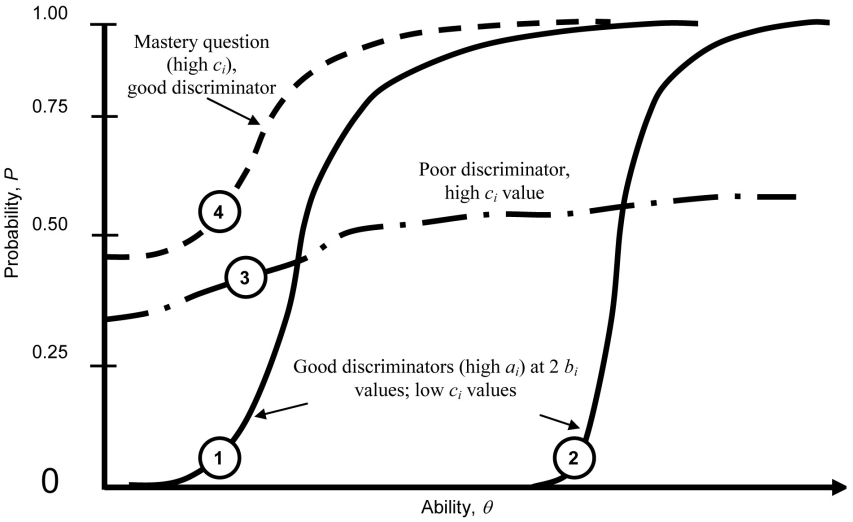

Fig. 1. Example Item Characteristic Curves (ICC) [10].

tor' items (marked ① and ②) have very different difficulty levels, and the 'poor discriminator' item (marked ③) has a difficulty in between those two. The parameter c_i is known as the *pseudo-chance-level parameter* (or 'guess factor'), and quantitatively is a lower bound asymptote for the ICC. The value of c_i indicates the probability that an examinee of lower ability can answer the item correctly; however, in an exam designed to assess the degree of subject mastery, there should ideally be some items that produce relatively high c_i values. This helps to support the concept that minimally acceptable levels of mastery were achieved by a large number of examinees; an example of an ICC for such an item is shown in Fig. 1 by the dashed line (marked ④). Thus, for each question or item on the assessment exam, an ICC can provide both a graphical and an indexed quantitative depiction of the item's difficulty, its ability to discriminate among ability levels, and its suitability for use as a subject mastery question.

Figures 2 through 5 show the ICC item groups for each of the four subject areas for which items were administered (Chemistry, Computer Science, Math and Physics, respectively). These ICCs represent aggregate results for the two institutions, so it cannot be determined how the results may have differed between mandatory (UC) and voluntary (PCUC) exam participants. Referring to Fig. 2 (Chemistry) to further explain ICC composition, the value of *P(θ)* (probability that a student of ability θ will answer the item correctly) for a given item logically varies between 0 and 100%, and the ability level for the item varies ranges from -3.00 to +3.00, representing standard deviations about a mean ability level. The three-parameter model described previously was used for the results presented herein.

Figure 2 shows examples of two typical ICC behaviors: one that has a high *P(θ)* intercept or guess factor, and very low slope, indicating that the item does not provide a strong element of discrimination among student ability levels; and two ICCs that show very strong discrimination tendencies, as indicated by their steep slopes at *P(θ)*= 50%, with difficulty ratings in the 1.0 to 2.0 range (reasonably high difficulty). The last prototypical ICC results from a mastery item (students of lower ability have a relatively high probability of answering the item correctly) and indicates that a large proportion of test-takers were able to answer the question correctly.

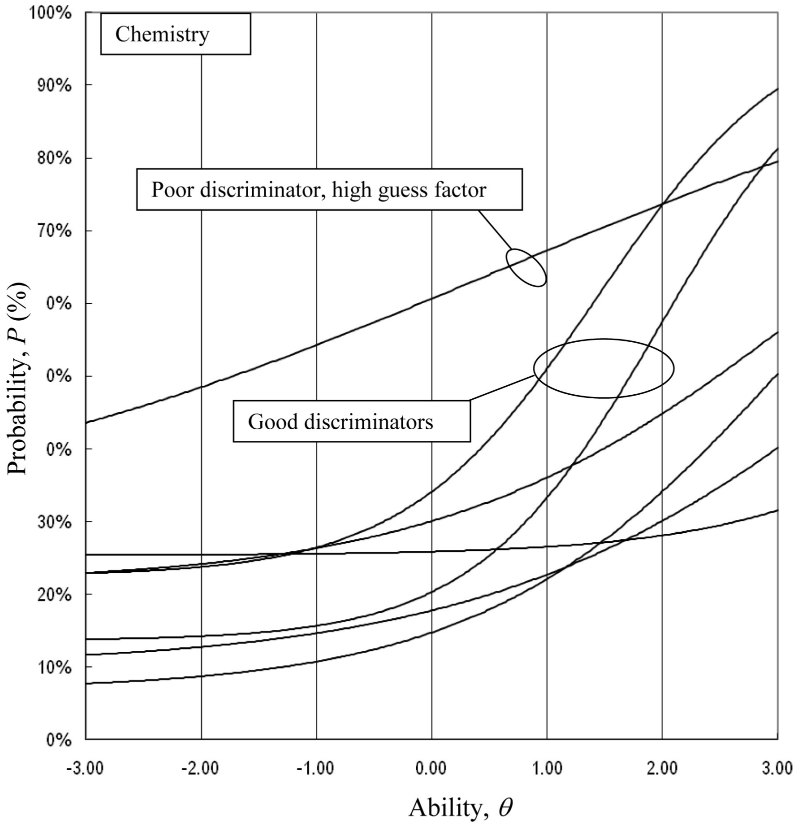

Fig. 2. Item characteristic curves for first exam administration Chemistry questions, UC and PCUC aggregated results, indicating the probability, P, that a student with ability level θ will correctly answer a particular question in that topic area.

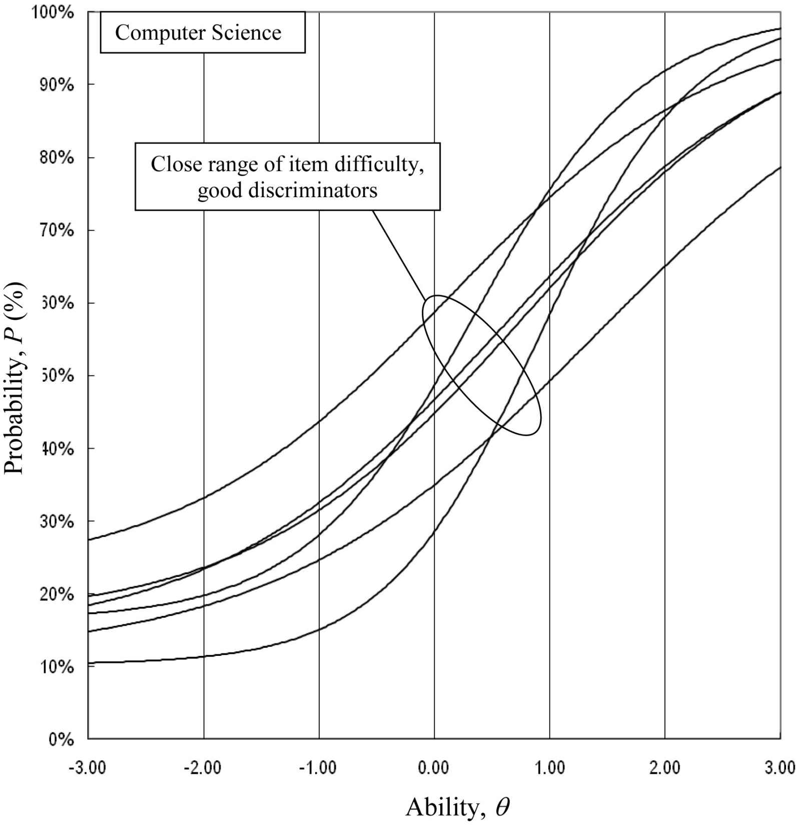

Fig. 3. Item characteristic curves for first exam administration Computer Science questions, UC and PCUC aggregated results, indicating the probability, P, that a student with ability level θ will correctly answer a particular question in that topic area.

Referring to Fig. 3 (Computer Science), the steep slope and relatively low guess factor associated with each item indicates that it was good at discriminating between the ability levels for each item. However, one issue that should be considered from this data is that the item set will not allow the faculty to distinguish among different ability levels given the close grouping of item difficulty levels. With the exception of the item marked with ICC denoted with a solid line, all items had an ability rating between about -0.50 and 1.00. This could lead to faculty discussion about what topics being assessed are more basic versus those that are more difficult, and thus may lead to higher difficulty items.

The Math questions (ICCs shown in Fig. 4) were identified by the UC and PCUC faculty, prior to IRT analysis, as having higher difficulty by far than any other item group based simply on the number of students who answered the question incorrectly or simply left them unanswered. The ICCs further bear this out by the number of items that were highly discriminating, but at difficulty levels that generally exceed those of the other subject areas. There were also three other non-discriminating items that needed to be rewritten. The faculty expressed their 'surprise' at the low number of students who were able to answer the Math questions correctly, and this was related to their observation that it is a challenge to predict item difficulty prior to the first exam administration. This observation highlights the previous observation that, without a trial exam administration, misunderstood questions, answer and distractors cannot be identified, and the exam cannot be 'calibrated' on any level, even with limited data. Thus, the first administration of this exam represented the trial and the Math items were identified as a group that needed to be re-examined in light of these results.

Finally, the Physics questions (Fig. 5) indicate a relatively wide range of difficulty levels and all are reasonably discriminating in terms of their utility in determining student ability levels.

4. Discussion

4.1 Modification of the exam after first administration

Because the first administration of the exam functioned primarily as a trial version, the faculty concluded that no definitive action should be taken based on the exam results. While they referred to the lack of 'reliability' in the exam results as a means to

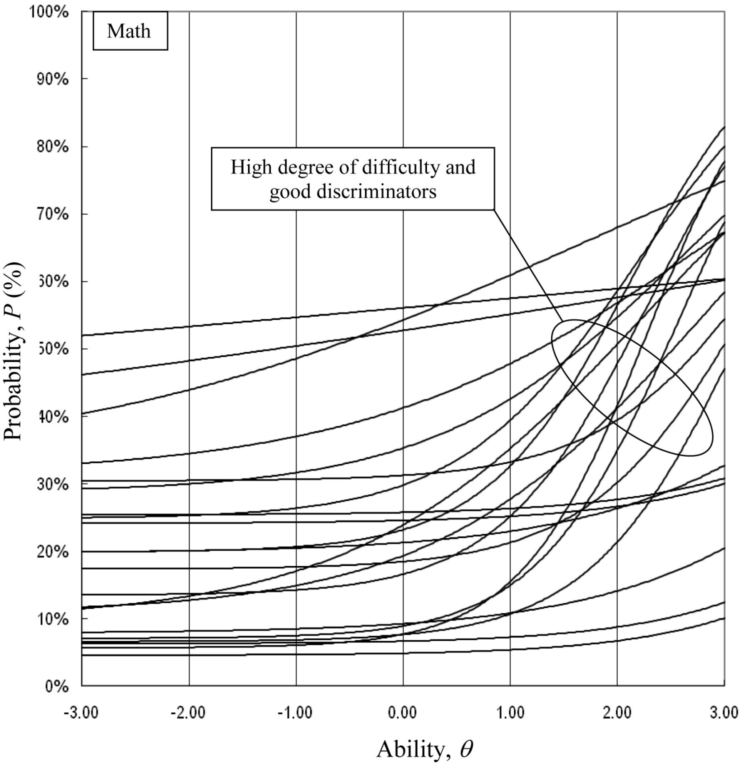

Fig. 4. Item characteristic curves for first exam administration Math questions, UC and PCUC aggregated results, indicating the probability, P, that a student with ability level θ will correctly answer a particular question in that topic area.

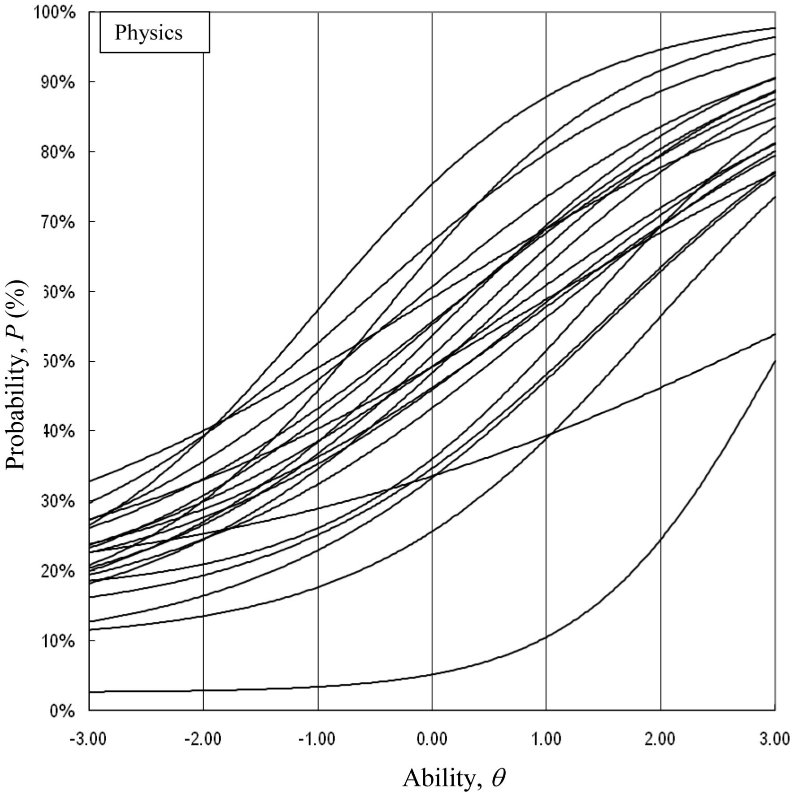

Fig. 5. Item characteristic curves for first exam administration Physics questions, UC and PCUC aggregated results, indicating the probability, P, that a student with ability level θ will correctly answer a particular question in that topic area.

Original Question	**Revised**
$\sum_{i=1}^{47}\frac{n^2+n-1}{(n^2+1)(n^2+2n+2)}$	$\sum_{n=1}^{15}\ln(1+\frac{1}{n})$
is equal to: $a)\frac{47^2}{2(48^2+1)}$ (correct answer) $b)\frac{47^2}{2(48^2-1)}$ $c)\frac{48^2}{2(47^2+1)}$ $d)\frac{48^2}{2(47^2-1)}$	is equal to: (a) ln 15 (b) ln 14 (c) 4 ln 2 (correct answer) (d) 5 ln 2

Fig. 6. Sample of how an item on the mathematics test was altered after analysis suggested the difficulty level was too high for the time available.

take programmatic action, it was really the validity of the data and its interpretation using IRT that was at issue. In August, 2009, PCUC administered a modified test (primarily the Math questions), and again detected some issues in the Math section that continue to show that these items are consistently more difficult than those in the other sections. Figure 6 shows an example of a question that was modified based on the results of the first exam feedback.

The experience by the faculties of the two institutions is indicative of the progression in developing and implementing such an instrument for educational assessment by those who are not experts in exam design. The instrument is initially the focus, with most energy devoted to refining the exam and managing the logistics of its administration and analysis. Once the exam is sufficiently refined and institutionalized, more 'mature' questions regarding teaching and learning can be asked and addressed using the instrument as valid feedback data [10, 28].

As of this writing, the exam has been administered a second time with the next group of students completing the first year curriculum at both institutions, and a revised exam is expected to be administered yearly at both institutions. PCUC has begun to require parts of the exam for all students finishing the first year engineering curriculum. The faculty responsible for teaching the common core engineering courses has reported that they have found the assessment results to be useful for informing curriculum decisions and for assessing programmatic instructional changes. As a result, both universities plan to continue the assessment procedure in the future.

4.2 Lessons for cross-national implementation of assessment instruments

The successful implementation of a student learning outcome assessment procedure in Chile, originally designed for use in the U.S.A., provides some useful lessons for institutional assessment of undergraduate engineering programs in general, and especially for cross-national implementation of assessment systems designed to meet the requirements of accreditation. These lessons seem to exist in three areas. Specifically, with the approach outlined: (1) Ownership of the academic program properly remains with the campus faculty who are involved at every decision and whose decision-making is informed by the assessment process; (2) The advisory group assisting with the development of the assessment assumed the role of consultant and thus do not share ownership of the procedures developed with the faculty. This is necessary because when the consultants leave, the faculty will be responsible for maintaining the system; (3) Feedback from the assessment system is directed at the faculty who can make the changes in curriculum, sequence, scope, and/or method of instruction that the data suggest to them and to students to help them gain a realistic sense of what they have mastered; (4) Responsibility for the assessment process, from initialization to gathering of the test scores, interpretation of the data, and implementation of curricular or instructional changes, is in the hands of the local interdisciplinary faculties directly responsible for teaching the students. This last point enables the process to occur with deference to local educational culture, traditions and language issues that others have found important in attempting to install educational practices cross-nationally [1, 4, 5]; and (5) Administrative support was a necessary ingredient to the success of the program on all three campuses. Commitment and support from the deans of the colleges in the form of resources, funds, and faculty release time commitments were vital to the successful implementation of the assessment system.

The IRT method of scaling tests described in this paper provides certain advantages for implementing standards to assess international training. The IRT method enables interpretation of a test score differently than the classical testing approach in which a student's score is compared with a group average. That is, the IRT method calibrates the test to the ability or knowledge domain being measured [18, 29]. In addition, methods have been developed for test equating and correction for bias using IRT calibrated tests [29, 30]. Further, the technical qualities and the mathematical basis underlying item response-based testing and analysis seemed to increase the appeal of the procedure for science, mathematics, and engineering faculty.

There were certain differences between the campus setting in the U.S. and those in Chile that impacted the implementation of the procedure that are worth mentioning. The first involved the level of participation by the administration. As stated earlier, support from the administrations at the various institutions was both significant and necessary. However, the administrators of both Schools of Engineering in Chile were relatively more 'hands on' in their involvement (participating in the workshop and being involved in the implementation) than was true at NU. Second, there seemed to be less familiarity with multiple choice formats in the Chilean setting. Item response theory accommodates other formats, but for reasons discussed elsewhere [10], the multiple choice format was chosen for this implementation. Third, there were differences in how the two Chilean universities ultimately were able to use the information provided by the test on their campuses, including a sense that the results could be used in a summative manner regarding instructional quality. Fourth, translation from English to Spanish seemed resolvable, partly aided by the cooperative atmosphere that existed before, during, and after the workshop in Chile, but also relied heavily on the presence of some people with high levels of fluency in both languages at the workshops. Since some of the concepts of test theory and the IRT methodology were somewhat new to the engineering faculty, the presence of colleagues who could translate conceptually at a high level was instrumental in the success of implementation. Finally, a spirit of collaboration across faculty academic specialties proved as important in Chile as it had on the NU campus. Cooperation, support and group collaboration across faculties seemed essential for the success of the implementation and the use of the data to improve the quality of student education on the US and Chilean campuses.

5. Conclusions

This paper has presented a case study on the process for implementing an assessment instrument, used successfully in the U.S., at two universities in Chile. In both the U.S. and Chilean implementations, the instrument was designed to assess learning in lower division engineering curricula in which course work is significantly delivered in departments outside of engineering (i.e., Chemistry, Math, Physics, and Computer Science), with the goal of using the assessment results to improve learning and identify areas of improvement in instructional delivery. A workshop was developed to provide instructors from the host departments with basic knowledge of learning models, methods for developing valid, multiple choice items that comprehensively test the most important course concepts (using a table of specifications), a quantitative analysis method, Item Response Theory (IRT), for item evaluation, and the use of analysis results for feedback and quality improvement. The authors found that some cultural and language differences had the potential to present barriers to successful implementation of the assessment instrument. The cultural barriers included less familiarity with multiple choice questions in Chile than in the U.S., and question nuances tied partially to language issues. However, it was found that a cooperative atmosphere and support by administrators proved critical in the successful deployment of the instrument and its subsequent use for feedback. The success of this implementation could prove to be a valuable lesson as engineering program accreditation that includes continuous quality improvement based on measurable learning outcomes (e.g., that proscribed by ABET) is adopted internationally.

Acknowledgments—The activities described were funded by the Chilean Ministry of Education through MECESUP Project UCH0403 and the Faculty of Physical and Mathematical Sciences of the University of Chile and Faculty of Engineering of Pontifical Catholic University.

References

1. G. L. Ooi and K. C. Goh, Networking for the region and beyond—Role of the Southeast Asian Geography Association. *International Research in Geographical and Environmental Education*, **17**(4), 2008, pp. 292–297.

2. M. Molphy, C. Pocknee and T. Young, Online communities of practice: Are they principled and how do they work? *Proceedings ascilite2007. http://www.ascilite.org.au/conferences/singapore07/procs/molphy.pdf*, Accessed July 27, 2010.
3. K. Ringwald, Transferring management knowledge in Anglo-Chinese Higher Education Collaboration: Are we speaking the same language? *Industry and Higher Education,* **22**(5), 2008, pp. 315–326.
4. M. Shibata, Assumptions and implications of cross-national attraction in education: the case of 'learning from Japan,' *Oxford Review of Education*, **33**(5), 2006, pp. 649–663.
5. C. S. Sankar, P. K. Raju and H. Clayton, Preparing students for global research experiences: U.S.–India summer projects, *International Journal of Engineering Education*. **25**(5), 2009, pp. 1046–1058.
6. Accreditation Board for Engineering and Technology (ABET), *Criteria for Accrediting Engineering Programs*, Location reference: www.abet.org, 2010, 29 pp.
7. European Ministers of Education. *Joint Declaration of the Ministers of Education*, European Higher Education Area, http://www.bologna-berlin2003.de/pdf/bologna_declaration.pdf, accessed March 20, 2011.
8. University of Edinburgh, Study Abroad in Edinburgh, Course Finder, Fluid Mechanics (Civil) http://www.ed.ac.uk/studying/visiting-exchange/course-finder?course=CIVE09014&session=1&subject=CIVE&year=2010/1&cw_xml=courseinformation.php, accessed March 20, 2011.
9. M. F. Letelier and R. Carrasco, Higher education assessment and accreditation in Chile: state-of-the-art and trends, *European Journal of Engineering Education*, **29**(1), 2004, pp. 119–124.
10. D. M. Qualters, T. C. Sheahan, E. J. Mason, D. S. Navick and M. Dixon, Improving learning in first-year engineering courses through interdisciplinary collaborative assessment, *Journal of Engineering Education*, **97**(1), 2008, pp. 37–45.
11. C. Adelman, The Bologna Process for U.S. eyes: Re-learning higher education in the age of convergence, Institute for Higher Education Policy, Washington, DC, 2009, 118 pp., Location reference: www.ihep.org/research/globalperformance.cfm.
12. W. M. Phillips, G. D. Peterson and K. B. Aberle, Quality assurance for engineering education in a changing world, *International Journal of Engineering Education*, **16**(2), 2000, pp. 97–103.
13. J. Lucena, G. Downey, B. Jesiek and S. Elber, Competencies beyond countries: The re-organization of engineering education in the United States, Europe, and Latin America, *Journal of Engineering Education*, **97**(4), 2008, pp. 433–443.
14. M. J. Lemaitre, Transnational higher education in Chile: A new development. In M. Martin (ed.) *Cross-border Higher Education: Regulation, Quality Assurance and Impact, Chile, Oman Philippines, South Africa,* Vol. 1, UNESCO International Institute for Educational Planning, Paris, 2007, pp. 56–128.
15. A. Patil and P. Grey, (eds.) *Engineering Education Quality Assurance. A Global Perspective*, Springer, New York, 2009, 316 pp.
16. J. Prados, G. Peterson, and L. Lattuca, Quality assurance of engineering education through accreditation: The impact of Engineering Criteria 2000 and its global influence, *Journal of Engineering Education*, **94**(1), 2005, pp. 165–184.
17. A. Patil and G. Codner, Accreditation of engineering education: Review, observations and proposal for global accreditation, *European Journal of Engineering Education*, **32**(6), 2007, pp. 639–651.
18. S. E. Embretsen and S. P. Reise, *Item Response Theory for Psychologists*, Erlbaum, Mahwah, NJ, 2000, 384 pp.
19. B. S. Bloom, H. Englehart, W. Hill, E. Furst, and D. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain,* Longmans, Green, New York, 1956.
20. L. S. Shulman, Making differences: A table of learning, *Change*, **34**(6), 2002, pp. 36–44.
21. L. D. Fink, *Creating Significant Learning Experiences: An Integrated Approach to Designing College Courses,* Jossey-Bass Adult and Higher Education Series, San Francisco, 2003, 320 pp.
22. R. M. Thorndike and T. Thorndike-Christ, *Measurement and Evaluation in Psychology and Education*, 8th edn, Pearson, Boston, 2010, 528 pp.
23. P. T. Ewell, National trends in assessing student learning, *Journal of Engineering Education,* **87**(2), 1998, pp. 107–113.
24. Weimer, M., *Learner-Centered Teaching: Five Key Changes to Practice*, Jossey-Bass, San Francisco, 2002, 288 pp.
25. D. Thissen, W.-H. Chen and R.D. Bock, *MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory* [Computer software], Scientific Software International, Inc., Lincolnwood, IL, 2003.
26. B. A. Hanson, *IRT Command Language*, http://www.b-a-h.com/software/irt/icl/, last accessed August 20, 2010.
27. R. K. Hambleton, H. Swaminathan and H. J. Rogers, *Fundamentals of Item Response Theory*, Sage Publications, Inc., Newbury Park, CA, 1991, 184 pp.
28. D. Bolt, The present and future of IRT-based cognitive models, *Journal of Educational Measurement*, **44**(4), 2007, pp. 377–383.
29. R. P. McDonald, *Test Theory: A Unified Treatment*, Erlbaum, Mahwah, NJ, 1999, 504 pp.
30. M. J. Kolen and R. L. Brennan, *Test Equating, Scaling and Linking: Methods and Practices*, Springer-Verlag, New York, 2004, 576 pp.

Thomas C. Sheahan, Sc.D., P.E. is a professor and undergraduate program director in the Department of Civil and Environmental Engineering at Northeastern University. He was a member of the GE Master Teachers Team from 1999 to 2003, developing initiatives to improve teaching and learning in first-year engineering courses. He went on to lead the follow-on grant that developed the Northeastern engineering mastery exam, and serves as the training and education director for major grants funded by the National Science Foundation and the National Institute of Environmental Health Sciences. He has published widely in his technical area of geotechnical engineering, and has presented work on engineering education and education technology.

Emanuel J. Mason, Ed.D. is a professor in the Department of Counseling and Applied Educational Psychology at Northeastern University. Dr. Mason has authored texts on research methodology and computing in schools, and was co-editor of a series on recruiting and retaining minorities for education. He has also published and presented numerous research papers on reasoning, assessment, and school psychology-related issues. His current research focuses on teaching science and technology, and developmental cognition.

Donna M. Qualters, Ph.D. is Chair of the Department of Education and Human Services, and the Director of the Center for Teaching Effectiveness at Suffolk University. She previously served as Director of the Center for Effective University Teaching and Associate Professor of Education at Northeastern University (CEUT). The teaching center oversees faculty development and student assessment activities. Her research focuses on creating educational change and she has published in the area of assessment, pedagogy, teacher identity/change, experiential education and reflective practice. She has been recognized by the Professional Organization and Development Network in Higher Education (POD) for her innovative faculty development activities including the book *Chalk Talk.* Dr. Qualters is a speaker on higher education teaching and learning.

Patricio V. Poblete is Professor of Computer Science and Director of the School of Engineering and Science at the University of Chile. His research is in the areas of Design and Analysis of Algorithms and Data Structures. As Director of the School, he has led an effort to redesign the curricula and the teaching methodologies, to focus on learning outcomes and to introduce active learning methods. The impact of these changes have been recognized by awards by SOCHEDI (the Chilean Engineering Education Society) and Colegio de Ingenieros (the Professional Engineering Society of Chile).

Ximena Vargas is a Professor in the Civil Engineering Department, University of Chile. She participates in courses related to Hydrology for civil engineering and master's degree students. Her research field deals with climate change, hydrologic modeling and forecasting of surface flow. She is also involved with groups interested in projects dealing with active learning engineering education and curricula redesign.