Negotiating the boundary between achievement and proficiency: An evaluation of the exit standard of an academic English pathway program

Susy Macqueen
Australian National University
Sally O'Hagan
University of Melbourne
Brad Hughes
Insearch, University of Technology Sydney

Academic English programs are popular pathways into English-medium university courses across the world. A typical program design hinges on an established university entrance standard, e.g. IELTS 6.5, and extrapolates the timing and structure of the pathway stages in relation to the test standard. The general principle is that the course assessments substitute for the test standard so that successful completion of the course is considered equivalent to achieving the minimum test standard for university entrance. This study reports on an evaluation of such course assessments at a major Australian university. The evaluation undertook to determine the appropriateness of the exit standard in relation to an independent measure of academic English ability. It also explored the suitability of the course final assessments used to produce measures in relation to that standard: by investigating the robustness of the processes and instruments and their appropriateness in relation to the course and the target academic domain. The evaluation was revealing about the difficult relationship between best practice in achievement testing in academic English pathway programs and external proficiency test standards. Using the sociological concept of 'boundary object' worlds (Star & Griesemer, 1989), we suggest that program evaluations that arise from a specific institutional concern for meeting adequate language standards can be informative about interactions between assessments in use.

Email address for correspondence: susy.macqueen@anu.edu.au

Key words : English for Academic Purposes (EAP) pathway courses, achievement testing, proficiency testing, university language requirements, international students

Background

The rise in demand for English-medium university education has brought about a substantial university entrance industry which includes large-scale standardized English tests and language education provision in the form of English for Academic Purposes (EAP) 'pathway' courses. These two educational products (EAP pathway courses and standardized tests) have something of a dependent relationship; tests define, construct and measure the entry standard, and courses develop EAP-learner progress towards it from some lower proficiency point. This can be the case, even if the course does not explicitly include the use of a standardized test because institutions frequently link their course structures to one or more score levels of a well-recognized standardized test such as IELTS[2]. Such links are forged using the relevant university language requirement standard (e.g. IELTS 6.5) as the final pathway course exit level. The precise link point is the minimum passing grade for the final pathway stage. Since this minimum course grade is considered sufficient for university entry, it has de facto equivalence with the minimum test standard required for university entry. Furthermore, entry requirements for the range of pathway course stages prior to this final stage are extrapolated from the minimum university entrance test standard. This extrapolation is based on a general understanding that time periods of intensive EAP study can be related to predictable test score gains (despite considerable variability in research investigating this relationship, e.g. Elder & O'Loughlin, 2003; Green, 2005). In Australia, a typical EAP pathway course structure assumes that students will gain half an IELTS score band in 10-15 weeks' intensive EAP study.

Although the EAP course exit standard is considered equivalent to the required proficiency test score/s, an EAP pathway course obviously aims to do more than achieve a proficiency standard as narrowly defined by a standardized test instrument. Such courses typically develop academic literacy with degrees of specificity to students' target disciplines and a range of academic study skills (Dooey, 2010; Terraschke & Wahid, 2011). Aims such as the development of critical thinking and the ability to follow academic conventions in writing essays and reports for instance, are not typically

[2] International English Language Testing System (www.ielts.org)

addressed in standardized language proficiency tests. This necessitates the use of assessment tools which reflect the skills developed in the course. Such tools indicate achievement in relation to the course goals and content, rather than proficiency in relation to the target domain. Therefore they differ from standardized tests in terms of method, content and purpose, but not in interpretation or use. That is, ultimately final course grades are considered indicative of language proficiency because they are assumed (by virtue of time-score gain claims) to be equal to the standardized test/s score requirement for university entrance. This dual interpretation/use places contradictory pressures on course assessments. As achievement tests, they should show students how much they have learnt during the course through sampling a course-representative range of genres and skills. Few students should fail in a typical cohort if course placement procedures are sound and teachers are teaching the appropriate curriculum and carrying out regular classroom-based assessment processes to monitor and scaffold student progress. As proficiency-test equivalent measures, however, pathway course assessments should be highly discriminatory at the university entrance cut point (the minimum course passing grade), below which learners are not admitted to university. Theoretically, then, pathway course final assessments are subject to the disparate validity considerations of teacher-based assessment for learning and standardized assessments.

As alternative forms of evidence of language readiness for tertiary study, the equivalence of successful pathway course completion and standardised test results is uncertain. Indeed, due to the disparities in purpose and aims of achievement versus proficiency assessments, the measures they provide are not readily compared. The numerous studies on the predictive validity of proficiency tests such as IELTS, for example, have tended to find only a moderate correlation at best between IELTS scores and academic success (see O'Loughlin, 2015). Despite this, Oliver et al.'s comparison of academic outcomes according to how students had satisfied university English language entry criteria led the researchers to conclude that standardized tests probably provide "the best evidence for potential academic success" (2012: 553). This finding is supported by other studies, up to a point, such as that of Floyd (2015) who found higher levels of achievement amongst students who satisfied language entry requirements with a standardized test score. However, Floyd reported that the differences were small, they were shown to diminish over time, and they appeared to be contingent on many other factors including age, academic discipline and level of study. On the other hand, contrary findings are reported by O'Loughlin and Bailey (2006; unpublished report cited in O'Loughlin, 2015) who tracked a cohort of pathway program students, who had sat an IELTS test on entry and exit, and found that successful course completion was an indicator of future academic success, irrespective of whether the pathway students' IELTS exit scores showed gain, stagnation, or regression. Although the value of pathway programs in preparing students for tertiary study is well recognised (Leask, Ciccarelli & Benzie, 2003), mixed findings about the predictive validity of pathway entry, combined with a lack of standardization across the diversity of accredited programs (see Benzie, 2011), would suggest that careful review of pathway course assessments is warranted to ensure these measures of achievement, such as they are, provide appropriate evidence of readiness for tertiary study (Dyson, 2014).

The distinction between achievement tests and proficiency tests is a long-standing one in language testing (for example Hughes, 1989). Proficiency tests are generally defined as future-looking instruments which carry out measurement of a specific test construct that is irrespective of prior periods of instruction. The test samples performance that is then evaluated in relation to a particular use or domain. Achievement tests, on the other hand, are focused on the performances of individuals or cohorts in relation to prior periods of learning. In the case of pathway institutions, it is the course content and objectives which embody the future link to the domain. Davies (1990) describes how proficiency tests can become achievement tests as teaching programs which align with in-demand tests are developed. In the case of pathway courses, we would argue that the reverse is also true where course achievement assessments become proficiency tests by virtue of their cohabitation with the test standard in the university entrance space. This paper describes an evaluation of such a dual-purpose instrument, EAP pathway course final assessments. The paper aims to document the process of evaluation as well as issues that arise as a result of the tension between measuring 1) course achievement, a retrospective view of a test-taker's performance, and 2) language proficiency for university entrance, a prospective view.

Context and aim

This evaluation is of an EAP pathway course for students wishing to enter a major Australian university with a focus on the final course assessments. The course in focus is a ten-week instructional period which, on successful completion, enables direct entry to university degree courses. In terms of institutional structure, the instructional period comes at the end of five course levels, which are each linked to a particular proficiency test score (an IELTS score) on the understanding that 200 hours of tuition (per 10-week course) results in an increment of 0.5 in IELTS score levels. The levels proceed in increments towards the university entrance score of IELTS 6.5. Table 1 below shows the institutional structure and links to particular proficiency measures (IELTS, TOEFL iBT, CEFR). It is important to note that students may enter at any point in the course structure, which means that the links between proficiency test standards and course levels are held firmly in place by student intakes at different proficiency score levels along the way, including the possibility of entering for only the final course level.

Table 1. EAP course entry standards and exit pathway

Course	**IELTS entry standard**	**TOEFL iBT**[3]	**CEFR**[4]	**Pathway** (course completion allows entry to…)
EAP 5	6.0	75	B2	Undergraduate degree
EAP 4	5.5	55	B1	Diploma course / EAP 5
EAP 3	5.0	45	B1	Foundation course / EAP 4
EAP 2	4.5	35	A2	EAP 3
EAP 1	4.0	25	A2	EAP 2

The program evaluation was a recommended procedure following a curriculum renewal at the pathway institution. The curriculum and assessments were revised to address stakeholder concerns that the pathway course content and standard was not adequately preparing students for university study. The new curriculum was underpinned by principles of *assessment for learning* (see, for example, Rea-Dickins, 2001), in particular, the use of formative assessment in reading and writing skills development and an emphasis on the use of teacher feedback in the development of academic writing ability.

The scope of the evaluation was essentially to investigate the legitimacy of the claim that successful completion of the final EAP pathway course assessments is an indication of readiness to cope with university study. This investigation was to be undertaken in several phases over a 3-year period and at the time of writing was in its third year. From the outset, the project had two interrelated strands of activity which are expressed in the aims below:

i) to investigate the adequacy of the EAP pathway course exit standards for university entrance; and

ii) to investigate the suitability of the final course assessments for providing measures in relation to the exit standard.

[3] Internet-based Test of English as a Foreign Language

[4] Common European Framework of Reference for languages

The first aim was addressed by comparing students' final pathway course results with an external criterion measure of their academic English proficiency. The second was addressed by reviewing the final course assessment processes and instruments in terms of their robustness and their alignment with the course and the target academic domain. The following sections exemplify further how each aim was addressed in the first phase of the evaluation process.

Method

The evaluation was a collaborative and iterative process which was negotiated and jointly managed by the pathway institution staff and an external evaluation team. The first phase, the focus of this paper, commenced with an information-gathering stage which was followed by the assessment materials review. Information about the course exit standards was then sought through an external criterion measure, the Diagnostic English Language Assessment (DELA), a validated post-entry language assessment designed to determine which university students may need support in order to cope with their courses. This measure was selected as an external criterion because it is one of the few post-entry assessments supported by a range of validation evidence (for example, Brown & Lumley, 1991; Elder & Erlam, 2001; Read, 2008). The DELA was also a feasible test to use in the time constraints of the commissioned program evaluation. Used at the University of Melbourne since the 1990s, the DELA comprises discrete test components for each of the sub-skills of academic reading, listening and writing, with results averaged across the three sub-skill tests to an overall score out of six. Cut-scores for the DELA classify students into three academic language proficiency groups: 'at risk', 'borderline' and 'proficient'. The DELA was administered to a cohort of 90 students who were about to exit the pathway course and the DELA results were compared with the overall results for the same cohort on the final pathway course reading, writing and listening tests i.e. aggregated across the three tests to an overall grade on a seven-level scale and with a cut-score classifying students as either 'passing' or 'failing' the course (and therefore meeting or not meeting the university entry requirement). Two modes of comparison were used: i) classification patterns based on cut-scores and ii) rank order correlation.

The assessment materials review was carried out by the external evaluation team via a desk review of course and assessment materials, curriculum documentation, and policy and procedural documents. An evaluation of the end of course test tasks was included in the review and for the listening test only, it was possible to carry out further investigation using Rasch analysis to evaluate item functioning and reliability. Information on the functioning of the assessment tools and course standards was also gathered from teachers, materials developers, level coordinators and management staff in focus groups/interviews conducted by the external evaluation team.

Findings

Materials review

The assessments which contribute to the final grade for the EAP pathway course are one reading-writing assignment and three final tests: listening, integrated reading-writing and speaking. All instruments relate to the final course content, which is related to the theme of 'leadership'. The weightings for these assessments are shown in Table 2.

Table 2. EAP pathway course assessment weighting

Assessment		Weighting	Description
Writing Assignment		30%	Comparative review of 3 input texts
Final Tests	Reading-Writing	40%	Comparative review of 3 input texts
	Listening	15%	Two listening texts; 20 items
	Speaking	15%	Interlocutor and peer interactions

Final pathway course results are reported on a seven-level scale, comprising four 'pass' levels (all of which enable university entrance), and three 'fail' levels. Writing is the most heavily weighted skill (70% of final grade) with both the written assignment and test requiring integration of specified reading material; there is no separate reading assessment. Listening is assessed separately in a test which is marked numerically using a marking guide. Writing and speaking tasks are rated using criteria and descriptors on the seven-level reporting scale. These are converted automatically to numerical scores which combine with the listening test results to produce one final course result. One concern raised by teachers was that the standard directly below the minimum pass was a very narrow percentage band and this might contribute to students' achieving the bare minimum numerical score even though they may have been rated as performing inadequately on qualitative criteria and descriptors.

Overall, the pathway course assessment adheres to the principles of *assessment for learning*; these are evident in the priority given to drafting, self-assessment and the facilitation of student engagement with learning goals and rating criteria. Successful completion of the course also includes completion of portfolio tasks. Throughout the course, rich feedback is provided to students; it is built into the assessment process, via criterion descriptors, a checklist of areas for improvement and free written comment. Summaries of the key points from the evaluation of the separate assessment components are below.

i) Writing-reading assessments

There are two writing assessments, a teacher-scaffolded assignment and an test, both of which require synthesis and integration, including the incorporation of citations of academic input texts. The input texts are related to the course theme. The writing assignment receives teacher feedback and is redrafted prior to rating. Both assessments are marked on criteria which can be readily linked to academic writing skills, including understanding and appropriately paraphrasing and referencing source texts. The assignment grade includes a 'process' criterion which is an evaluation of a student's response to feedback and his/her ability to edit and self-assess. Other than this criterion and the conditions of performance, the assignment task and the test task are essentially the same. Routine rater training and rater moderation processes are in place.

Recommendations were that consideration be given to the development of a separate reading test as the assessment of reading ability is minimally represented in the writing criteria, although it is integrated (appropriately) in the writing performance. It was also recommended that the writing test be less of a replication of the assignment to allow for a more independent sample of writing.

ii) Speaking assessment

The speaking test is carried out in groups of three students with an interlocutor (also the assessor) who has a script to follow. It is divided into three parts: question-response, long turn and discussion. The speaking genres – long turn and leading group discussion – are highly relevant to the target context where oral presentation and discussion skills are highly valued. Topics are related to the course theme. The rating scale is composed of criteria and descriptors that are well-balanced in their focus on broader discourse features such as interaction (e.g. sensitivity to turn-taking norms and ability to initiate and respond) as well as grammar, vocabulary and pronunciation. Rater training was perceived to be regular and thorough.

Recommendations were that some form of routine monitoring of rater reliability, e.g. double rating a random sample, be carried out. It was also recommended that teachers do not rate their own classes and consideration be given to principled ways of grouping students for the test so that they are not advantaged/disadvantaged by their peers' interactive abilities (e.g. O'Sullivan, 2002).

iii) Listening assessment

There are two forms of the listening test, each comprising two texts and 20 items (multiple choice, true/false, short answer). These are based on the course theme and are relevant to the listening skills demands of tertiary contexts, utilising relevant text types and genres (monologic academic lecture and interview dialogue). Overall, in terms of content, structure and delivery, the texts have an authentic feel. The production of written test specifications was recommended in order to standardize the listening subskills targeted and the text lengths and types, as was a process of statistically equating the two forms.

The statistical properties of one Listening test form were investigated with Rasch analytical procedures using the Winsteps program (Linacre, 2014) to establish overall test reliability and to evaluate the functioning of individual test items. The analysis showed that overall test reliability was very low, as indicated by the reliability statistic of 0.51 for Cronbach's alpha/KR-20. Item analysis revealed a number of items with problematic functioning in terms of item discrimination (i.e. how well items are able to discriminate among the ability levels of the test takers) and item facility (i.e. the levels of difficulty of the items). However, the analysis did show that students' abilities and item difficulties were, on the whole, matched predictably (with all infit mean square statistics in the range of 0.7 to 1.30, all items showed 'good fit'). The test analysis results suggested that substantial changes to the Listening test were needed to improve overall test reliability and to reduce the number of poorly functioning items. Increasing the number of items to gain a larger listening performance sample was also recommended.

External criterion

The results on the DELA and the pathway course tests (i.e. aggregated scores on the writing, listening and speaking tests reported as an overall grade on a seven-level scale), were compared for their similarity of classification patterns of students in terms of readiness for tertiary study. Using a further mode of comparison, a correlation coefficient was computed to determine the extent to which student performances on the two tests are positively correlated.

For the pathway course, the relevant classifications were considered to be:

- Pass grade and above = qualifying for 'direct entry' to university course
- all failing grades = 'not direct entry'.

For DELA, the relevant classifications of students were:

- 'proficient' = sufficient academic English proficiency for the demands of tertiary study
- 'borderline' = likely to be in need of further language support or development
- 'at risk' = at risk of academic failure without further language support or development.

The classifications of students according to the pathway course final tests and the DELA were compared to show whether the classification patterns according to the two assessment tools are sufficiently similar i.e., students who pass the pathway tests would be more likely to be classified 'borderline' or above by the DELA, and less likely to be deemed 'at risk' by the DELA. The classification patterns based on overall results on each test are shown in Table 3.

Table 3. Classification of student test takers (N=90) by EAP pathway tests and DELA

	DELA classifications			
EAP 5 aggregated test scores	At risk	Borderline	Proficient	TOTAL
EAP 5 Pass and above	57	23	7	87
EAP 5 Fail	2	-	1	3
TOTAL	59	23	8	90

As can be seen from Table 3, the classification patterns of the two tests are different. Although most students (7 of 8) deemed 'proficient' by the DELA also passed the pathway tests, it is also the case that the majority of students (57 of 59) deemed 'at risk' by the DELA also passed the pathway tests, thereby succeeding in the university direct entry pathway. One implication of this finding, to be drawn with caution, is that the standard may not be high enough for the demands of tertiary study.

In interpreting this finding, on the one hand, it is important to keep in view the context of each test: the pathway tests are assessments for university entrance, while the DELA is taken post-entry. However, the target domain for the two is nevertheless essentially the same. Therefore, one would expect that the two assessments of academic English language proficiency would sort students similarly, a lower motivation for success on the no-stakes DELA for these students notwithstanding.

In addition to the classification patterns, the results of the two tests were compared taking an approach independent of the DELA standards. Based on the assumption that the EAP course assessments and the DELA should sort students similarly by level of academic English proficiency, it would be expected that the two sets of test scores would be positively correlated. The relationship between DELA overall results (i.e. aggregated scores reported out of six) and final EAP results (aggregated scores reported as a grade on a 7-level scale) was investigated by computing a rank order correlation coefficient (Spearman's *rho)*. Although a positive correlation was found between the two variables, the correlation was small and not statistically significant, $r_s = .150$, n = 90, $p = 0.159$. The correlation coefficient shows that the scores co-vary in a positive direction i.e. as DELA scores increase, so do EAP 5 scores; however, it also shows that the strength of this relationship is weak. Overall, this result shows only a limited relationship between DELA overall results and final EAP 5 test results. Given the common target domain of the two tests, this result runs contrary to expectations.

Conclusions

To return to our first evaluation aim of investigating the adequacy of the EAP pathway course exit standards for university entrance, the external criterion findings (the comparison of final pathway results with DELA) suggest that concern about EAP pathway exit standards may be warranted. The majority of pathway students who are deemed ready for university entrance would be classified by a post-entry academic English test as being at risk of academic failure without language support. While there are certainly caveats to consider, such as the different purposes of the testing programs and the motivation of the pathway students in undertaking DELA purely for the purposes of the program evaluation, it would be nevertheless reasonable to expect similar patterns of success and failure to emerge in the comparison. The unexpectedly weak correlation between DELA and pathway scores suggests an area for further investigation: perhaps via a repeat of the comparison with a future cohort in the first instance. Washback, potentially arising from revisions to the pathway assessments (following recommendations of the materials review), could also contribute to different results in a future comparison. The finding that successful pathway students were classified as 'at risk' according to an instrument designed for post-entry assessment on the other hand, may indeed underscore the need to recognise cut-scores as minimum requirements for entry. In other words, evidence of the level of readiness to commence tertiary study–independently, or with some degree of future support–needs to be determined post-entry. The approach taken so far to investigating the pathway exit standards could be complemented by other indications that students have been equipped with adequate linguistic resources to cope with university course demands. Such indicators could include: tracking pathway graduates' academic performance using course achievement data, such as GPA (grade point average) or WAM (weighted average mark), more specific grades achieved on 'language-rich' university assessments, a fine-grained linguistic analysis of such writing in relation to that produced in pathway course assessments, and students' own perceptions of their university readiness (e.g. Dyson, 2014).

In relation to the second aim, to investigate the suitability of the final course assessments for providing measures in relation to the exit standard, there were clear links perceived by the materials reviewers between the pathway assessment tools and the target domain practices in terms of skills sampled and content represented. There were also aspects of the assessment tools and the assessment processes which might be improved. Accordingly, specific recommendations were made concerning the diversity of assessments, the improvement and monitoring of reliability, and the development of test specifications. These recommendations were acted upon by the pathway institution, as detailed below.

Institutional response

The EAP institutional response to the findings of the DELA and EAP comparison combined with the review of assessment materials would be embodied in one of the aims of subsequent phases in the program evaluation: to build a more robust exit assessment program. In line with the recommendations arising from the materials review, several activities guided by this aim were undertaken including i) a comprehensive listening test redevelopment to address the low reliability of the test, ii) a separate reading test development to address the underrepresentation of reading skills, iii) revisions to the writing test to distinguish it from the assignment task and iv) the development of test specifications for all assessments to increase the equivalence of test forms and clarify the targeted assessment construct. It was envisaged that once these materials developments were undertaken, a standard-setting exercise would be carried out. This would use actual EAP Level 5 assessment samples to elicit stakeholder judgements (e.g. undergraduate lecturers) about an appropriate language standard. Such an exercise, it is hoped, will result in a locally-contextualised standard and a more meaningful use of EAP course scores. In this way, the standard would be set through direct interaction between the receiving institution and the course assessments. This more direct standard would relate to the expectations of the two institutions (EAP pathway institution and university) as well as maintaining a relationship with the IELTS standard as a result of continued use of IELTS-defined EAP course entry points and assumptions about time and proficiency development.

A further effect of the findings was a follow-up investigation of the exit standards of the penultimate pathway course. This reflects a trickle-down effect of the concern for the ultimate standard to prior steps in the pathway program which are also held in place by the IELTS score-gain per time-period assumptions.

In addition, the pathway institution invested considerably in the assessment literacy of the staff through professional development in Rasch analysis and test writing. Teams of teachers were also involved in consultations about exit standards and test revisions. Other responses from the pathway institution to the evaluation include the implementation of double rating procedures for the writing test and statistical analysis of rater severity and treatment of criteria. It is hoped that these activities will result in a more defensible assessment program.

Discussion

The evaluation recommendations and the institutional responses are essentially a standardizing reform in which each assessment instrument is made more uniform through measurement methods such as statistical properties and behaviour constraints such as rater monitoring and speaking examiner scripts. As a means of high-stakes decision-making, it is reasonable to expect that final pathway assessments are subject to standardization since they are otherwise at risk of operating unfairly for students expecting uniform assessment treatment and receiving institutions expecting uniform indications of linguistic readiness. However, implementing standardization methods requires some investment on the part of the EAP institution. Two important questions arise: 1) how much standardization is enough to be accountable to the stakeholders, and 2) how much standardization is enough to balance the different purposes of retrospective-looking course achievement assessments and prospective-looking proficiency tests.

Pathway courses inevitably invoke a journey metaphor; crossing the boundaries of social worlds. The assessment instruments that communicate between the worlds are boundary objects which need to satisfy the informational requirements of both worlds (Star & Griesemer, 1989). That is, they should indicate how well the course was completed in the prior EAP pathway context on the basis of sampled skills, topic and text types and vocabulary and grammatical structures covered in the course, as well as predict success in the new university context on the basis of a language use sample. Bowker and Star (1999) describe how boundary objects are buffeted between worlds. In their conceptualisation, it is through this buffeting process that boundary objects ramp up to become more uniform, naturalized mechanisms or 'standards' over time. In this program evaluation we can glimpse this process in action whereby the concerns of stakeholders bring about standardizing activities so that everyone 'understands' the meaning of the object in between. The meaning of the pathway assessments are intertwined with the IELTS standard which looms large in the finding that the pathway students were not sorted similarly by another academic English test (the DELA) from a university similarly bound to IELTS-derived standards.

Standardization in human behaviours such as language testing involves diverse individuals doing the same thing so that differential sorting may occur. Teaching, however, despite occurring in groups, is very much concerned with the diverse trajectories of individuals. In the case of the evaluation components reported here as well as its subsequent phases, teachers' concerns for the language development of individual students is well in evidence. The emphasis on assessment for learning and strong feedback processes built into the formative assessments and the writing assignment also reflect this orientation to the development of the individual. Thus, an effective pathway assessment program must encompass the full range of assessment mechanisms. These mechanisms range from sensitively-targeted individual feedback during class to multiple test forms which offer evidence that substitutes adequately for both the narrow proficiency remit of other test standards, as well as proves that the broader academic aims of the course have been achieved. It is important, however, to monitor standards as they operate in local contexts. In this sense, pathway course assessments, once sufficiently standardized, can themselves be used to set an appropriate minimum entry standard, rather than through reference to third-party standards such as IELTS. This would involve the two institutions negotiating directly over an appropriate minimum level standard and is likely done most meaningfully when subject matter is similar (e.g. business course teachers judge the adequacy of samples derived from course assessments on business-related topics) and when the context of the sample (e.g. task and conditions) is made highly apparent. A proficiency test does not do the same thing as a pathway course and perhaps recognising this by interacting separately with these mechanisms is a constructive course of action.

A final note relates to the role of program evaluator. Because boundary objects, such as language tests, intersect different worlds and maintain coherence between them (Macqueen, Pill, & Knoch, 2016; Star & Griesemer, 1989), the program evaluator plays a role in the process of between-world negotiation. This involves adaptation of methods to context and sensitivity to the possible consequences of misinterpretation of recommendations (Lynch, 1990). In this role, evaluators are constrained by multiple forces, including but not limited to practical concerns such as time, staffing and project financing, on the one hand, and, on the other, the demands of their own discipline, including the need to provide warrants for their claims about the program in the form of accepted research tools, processes and methods (Freeman, 2009). In this project, for instance, as evaluators, our decisions about whom to consult, how to review the assessment materials and which external criterion measure to use were all, to some extent, a balance of practical concerns and disciplinary standards. To use the IELTS test itself as the external criterion would of course have been a more direct way to compare the EAP pathway standard and the IELTS standard. However requiring the students to sit IELTS in an external venue in a timely manner in addition to their end of course assessments was a prohibitive factor. In any case, the academic construct and diagnostic purpose of DELA may even make it a more commensurate measure of linguistic readiness for academic study, albeit a less direct one. Similarly, as with many program evaluations, the materials review process was simply an informed appraisal of the program, carried out by a team of two to five external evaluators with relevant experience and expertise. Such exercises do not necessarily use knowledge-building methods which might employ, for instance, a specific theoretical framework and a rigorous triangulation procedure. The questions asked in program evaluations such as this one are tailored to particular institutional needs rather than to the more stringent requirements of academic research. This institutional specificity may prevent rich local experience such as that documented in evaluation reports, from contributing to disciplinary knowledge. Just as a laboratory-tested drug may perform differently once it is in interaction with the individual physiologies of humans, the food they eat and other drugs they take, language assessments should also be examined in terms of contextual interactions. By providing glimpses of the actual practice of program evaluations such as the one as described in this paper, we can understand more about the unpredictable effects of multiple assessments in use; their inevitable interactions and potential cross-purposes.

References

Benzie, H. (2011). A pathway into a degree program: Forging better links. *Journal of Academic Language & Learning 5*(2), A107-A117.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, Mass.: The MIT Press.

Brown, A. & Lumley, T. (1991). *The University of Melbourne ESL Test. Final report.* Language Testing Research Centre, University of Melbourne.

Davies, A. (1990). *Principles of language testing*. Oxford: Blackwell.

Dooey, P. (2010). Students' perspectives of an EAP pathway program. *Journal of English for Academic Purposes, 9*(3), 184-197.

Dyson, B. (2014). Are onshore pathway students prepared for effective university participation? A case study of an international postgraduate cohort. *Journal of Academic Language & Learning 8*(2), A28-A42.

Elder, C. & Erlam, R. (2001). *Development and validation of the Diagnostic English Language Needs Assessment (DELNA)*. Auckland: The University of Auckland.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS *IELTS research reports* (Vol. 4, pp. 207-254). Canberra: IELTS Australia.

Floyd, C. B. (2015). Closing the gap: International student pathways, academic performance and academic acculturtion. *Journal of Academic Language & Learning* 9(2), A1-A18.

Freeman, D. (2009). What makes research 'qualitative'? In J. Heigham & R. A. Croker (Eds.), *Qualitative research in applied linguistics: A practical introduction* (pp. 25-41): Palgrave Macmillan.

Green, A. (2005). EAP study recommendations and score gains on the IELTS Academic Writing test. *Assessing Writing, 10*(1), 44-60.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Leask, B, Ciccarelli, A. & Benzie, H. (2003). Pathways to tertiary learning: a framework for evaluating English language programs for undergraduate study. *English Australia Journal, 21*(1), 17-29.

Linacre, J. M. (2014). Winsteps® (Version 3.74.0). Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Lynch, B. K. (1990). A context-adaptive model for program evaluation. *TESOL Quarterly* 24(1), 23-42.

Macqueen, S., Pill, J., & Knoch, U. (2016). Language test as boundary object: Perspectives from test users in the healthcare domain. *Language Testing, 33,2:* 271–288.

Oliver, R., Vanderford, S. & Grote, E. (2012). Evidence of English language proficiency and academic achievement of non-English-speaking background students. *Higher Education Research & Development, 31*(4), 541-555.

O'Loughlin, K. (2015). 'But isn't IELTS the most trustworthy?': English language assessment for entry into higher education. In A. Ata & A. Kostogriz (Eds.), *International education and cultural-linguistic experiences of international students in Australia*. Samford Valley, QLD: Australian Academic Press, 181-194.

O'Loughlin, K. & Bailey, A. (2006). *An evaluation of an intensive English language preparation program for postgraduate study in education.* Unpublished report.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277-295.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing, 18*(4), 429-462.

Read, J. (2008). Identifying academic language needs through diagnostic assessment, *Journal of English for Academic Purposes*, *7*(3), 180–190.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science, 19*(3), 387-420.

Terraschke, A., & Wahid, R. (2011). The impact of EAP study on the academic experiences of international postgraduate students in Australia. *Journal of English for Academic Purposes, 10*(3), 173-182.