A framework for validating post-entry language assessments (PELAs)

Ute Knoch
Cathie Elder
University of Melbourne

Introduction

It is now generally acknowledged that many students from the increasingly heterogeneous population of entrants to higher education in Australia may face problems with the English proficiency demands of their academic study and that this may have an impact on their performance. Competition between institutions to secure ever higher international student numbers means that the English entry requirements of many universities are set quite low. There are in any case a number of alternative pathways for student entry which may exempt students from these English requirements. Domestic students who have experienced English-medium secondary schooling may also experience problems with academic English, especially if they have chosen to study subjects that do not make strong language demands. Attention has focused quite strongly on these issues in recent times following the introduction of a Higher Education Standards Framework by the Tertiary Education Quality and Standards Agency (TEQSA, 2011), which includes a requirement that institutions make provision for English language development as a 'key graduate attribute'. To comply with these standards many universities have developed policies aimed at addressing the English language and academic literacy needs of their students including, in many cases, some form of post-entry English language assessment (hereafter PELA) to identify those who require support and to pinpoint the dimensions of language ability that may require such intervention.

A range of PELAs are currently operating in Australian universities and these vary not only in content and format but also in the manner of implementation, with some institutions offering them on a voluntary basis to particular "at risk" groups and others mandating them for all enrolled students, including both native and non-native English speakers. Depending on their PELA results, students are provided with different avenues for language development, whether in the form of credit-bearing courses or 'sheltered' tutorials offered as part of their academic program or as add-on courses/workshops or one-on-one assistance with assignments. While institutional policies vary, university

PELAS and associated support courses share the common goal of improving the quality of English and also the academic performance of enrolled students.

Australian PELAs, unlike the high stakes English tests (e.g. IELTS, TOEFL, PTE) used for university selection, tend to be developed in-house and are seldom professionally validated. Some notable exceptions for which validation evidence is available are the Diagnostic English Language Assessment (DELA) instituted at the University of Melbourne in the early 1990s (Brown & Lumley, 1991) and its subsequent incarnation, DELNA, at the University of Auckland (Elder & Erlam, 2001; Read, 2008). Nevertheless, as noted by Davies and Elder (2005) among others, validity is a relative rather than absolute concept, raising questions about how much validation is enough and where research efforts and resources are best directed if institutions are to claim integrity for their assessments.

One source of information about post-entry English language assessment is the recently completed Degrees of Proficiency project (Dunworth, 2013) funded by an Office for Learning and Teaching grant and the project website http://www.degreesofproficiency.aall.org.au/, set up to provide Australian universities with tools and strategies to develop students' English language capabilities. This website, while generally useful, falls short of explaining what validity might mean in this post-entry context or what the validation process entails, simply noting that PELA scores should be valid and defensible and that specialist expertise is required to achieve this. The structuring of the website, moreover, suggests a rather tenuous distinction between *validation*, viewed as a technical matter relating to a test's properties, and *evaluation*, characterised as broader in scope and encompassing a wider range of investigations including administration procedures, user perceptions, user uptake and the intended and unintended consequences of the assessment. This runs counter to current argument-based approaches to test validation, which consider *all* such investigations to be integral to the notion of validity (see e.g. Kane, 1992, 2006, 2013; Chapelle, Enright & Jamieson, 2008). What seems to be needed, then, is a principled framework outlining the various arguments that underpin validity claims in PELA contexts and the types of evidence that might be adduced to support these arguments.

Validity and argument-based approaches to validation

Before presenting such a framework, the following section presents some brief historical background on validity and validation and in particular on the argument-based approach expounded in this paper.

Conceptualisations of validity and approaches to validation have rapidly changed in recent decades (e.g. Chapelle, 1999; Xi, 2008). In the 1970s and early 1980s, validity was perceived as a number of disparate types (e.g., construct validity, content validity, predictive validity), with reliability, or consistency of measurement, seen as a pre-condition for rather than a component of validity. Estimates of reliability were the starting point for monitoring a test's functioning and subsequent validation processes were largely confined to statistical analyses indicating how well the test correlated with other measures designed to measure similar skills or skills in the relevant target language use situation. At that time, there was little interest in examining test use and test consequences, or rather, these were seen as a separate area of enquiry unrelated to validity. In the late 1980s, following the publication of Messick's (1989) seminal paper in which he described validity as a unitary construct with construct validity subsuming all other aspects including reliability, there was a shift in thinking about the scope of validity and the procedures required for validation. In particular, there was dissatisfaction with the traditional and more circumscribed view of validity, as this did not account for any consequences of test use. Messick, like many others who followed him (e.g. Fulcher & Davidson, 2007), believed that the intended consequences of a test should be articulated at the test design stage and evidence then be gathered to determine whether the actual test effects corresponded to what was envisaged. Importantly, Messick also maintained that validity did not reside in the test itself, but rather in the adequacy of the support for the inferences drawn from test scores and for the actions or decisions stemming from these inferences.

Although Messick's theoretical model of the different facets of validity was hugely influential, it offered little by way of practical advice for practitioners and researchers on how to proceed with validation of their tests. To address this gap, Kane (1992) and Kane, Crooks and Cohen (1999) proposed an interpretative argument to guide validation work. This validity argument laid out the inferences and assumptions associated with the score interpretations systematically and the plausibility of these inferences and assumptions was evaluated using both theoretical and empirical evidence. Figure 1 below presents the main components of such an interpretative argument.

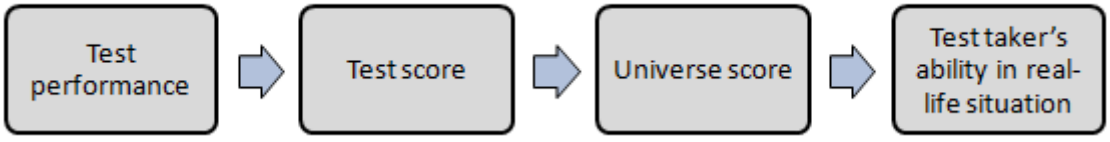

Figure 1. Basic building blocks of interpretative argument (adapted from Kane, 1992)

Each arrow in Figure 1 represents an inference. These inferences serve as bridges between one building block of the argument and the next. For a validity argument to be supported, each bridge needs to be supported by evidence. In this way the entire argument can be coherent and the final interpretation to be valid. Underlying each inference are a number of assumptions, which need to be stated explicitly to provide a basis for research supporting each inference.

The inference connecting the building blocks 'test performance' with 'test score' is the *evaluation* inference. Crossing this bridge successfully involves transforming a test taker's performance on the test into a defensible test score. Underlying assumptions which require evidential support are that the scoring on the test is conducted consistently and that the scoring procedures are valid reflections of the test construct or what the test is designed to measure. It is also assumed that the instructions on the test or the process of test administration have not introduced any construct-irrelevant variance (i.e., factors which might interfere with what the test is designed to measure). If these assumptions can be supported, the raw score can be considered defensible.

The inference linking a person's score on the test with the 'universe score' is called *generalisation*. Generalisation rests on the assumption that the observed test score derived from any particular test administration is representative of the score a test taker would receive on other versions of the test or on any other test occasion, regardless of the particular tasks or judges involved. This inference is akin to what has traditionally been termed 'reliability'.

The final inferences represented by the last arrow in Figure 1 above, the *explanation* and *extrapolation* inferences, connect the universe score with the theoretical construct that the test is designed to measure and, by implication, with the test takers' performance in a real life situation, also called the target score. The underlying assumptions requiring supporting evidence would include how well test tasks elicit strategies and processes by test takers as intended by the test designers and as implied in the model of ability on which the test is based (the 'explanation' inference) and how well the test tasks reflect the language demands of the relevant real-world domain (the 'extrapolation' inference).

Kane's earlier model, as set out in Figure 1 above, was limited as it did not take account of issues of test use or test consequences. This was addressed in later iterations (Kane, 2001, 2002, 2004, 2006) in which he included a further building block, decisions, with a further inference called *test use*.

Bachman and Palmer (2010), building on earlier work by Bachman (2005), proposed an assessment use argument. While Bachman's earlier work proposes two main parts to any validity argument: (1) the interpretation based on the results derived from the instrument in question and (2) a 'decision-based' interpretation, these two aspects were merged into a single assessment use argument in their 2010 work. Their work has extended what Kane proposed in the area of decisions and test consequences, offering a range of use and consequence-related assumptions that need to be supported for the score based interpretations and uses to be valid.

A framework for evaluating PELAs

For the purpose of the proposed framework, we have adopted a hybrid model, encompassing both Kane's (1992, 2001, 2004, 2006) models and parts of Bachman and Palmer's (2010) assessment use argument. The original building blocks of the two models are set out in Figure 2 below.

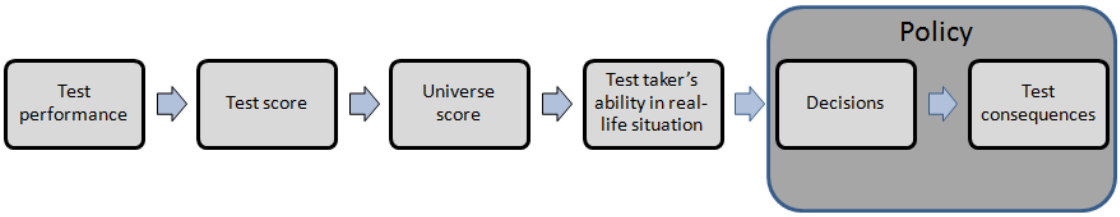

Figure 2. Building blocks of hybrid model based on Kane (1992, 2001, 2004, 2006) and Bachman & Palmer (2010)

As can be seen, the framework draws on the first few components and inferences from Kane's work and draws on aspects of Bachman and Palmer's framework to investigate decisions and test consequences. We chose to use the components of Kane's framework (i.e., test performance, test score, universe score and test taker's ability in the target language use domain) as the first building blocks in our model because they have been widely applied in validation studies, include all the necessary components to evaluate aspects directly related to the test and are directly relevant to PELAs.[1] Our decision to incorporate some aspects of Bachman and Palmer's (2010) assessment use argument in our framework was due to the particular emphasis these scholars place on decisions and test consequences. Both of these, we would claim, are critical in the evaluation of a PELA because the success of any PELA initiative

[1] Kane later divided the extrapolation inference into two (extrapolation and explanation) but we did not think that this further division was necessary for the evaluation of PELAs and therefore decided to maintain Kane's (1992) original sense of extrapolation as encompassing both test taker skills/processes/knowledge and their relevance to the construct on the one hand and how the assessment tasks relate to tasks used in the academic domain on the other.

relies on uptake of the advice stemming from test results. Expectations about decisions (e.g. about how results are presented to test takers) and also about the consequences of these decisions (e.g. what test takers make of their test results, how the results are used by the institution and how test takers will benefit from the follow-up support associated with test outcomes) need to be made explicit in the assessment use argument for PELAs and to drive its content and structure. While these elements are by no means absent from Kane's model, and indeed have been paid more attention in recent work (Kane, 2013) they are encompassed within the decision step and not sufficiently visible for our purposes

Finally, as well as going beyond Kane's Decisions and adding Bachman and Palmer's Consequences to the model, we have chosen to highlight the importance of institutional policy in framing Decisions and Consequences by placing it behind the model (see shaded area). This is because, in our experience, successful implementation of a PELA depends so crucially on the institutional policy surrounding its implementation and how the purpose of the test is conceived within that policy. This policy framing feeds into how we conceptualise good and bad test decisions and what we consider to be positive or negative consequences.

Thus, the way an institution's PELA policy is framed will govern such determinations as which student groups are targeted for testing, whether equal access is ensured, whether the testing requirement is mandated, how the requirements of the assessment are communicated to test takers, the level and type of support provided for at risk students and whether such students are required to attend such courses. Bachman and Palmer have linked policy issues to Consequences but our view is that policy governs Decisions as well. Acknowledging the policy dimension of PELA activities is in keeping with Kane's view that considering "the test-as-policy and the goals of this policy in achieving certain outcomes is critical and should be evaluated in terms of its perceived effectiveness" (Kane, 2002, p. 38). As McNamara and Roever (2006, p. 30) have pointed out, policy evaluations do not happen by default even in publicly accountable educational systems, so the importance of policy scrutiny needs to be emphasized..

To illustrate and visually summarise our framework with the accompanying inferences, we have chosen an upside-down pyramid (see Figure 3). We feel that this shape demonstrates the hierarchy of inferences and assumptions and the nature of the supporting evidence that is required. At the bottom end of the pyramid, the focus of the validation work is more narrowly on test reliability, the functioning of test items, the way the test is scored and how the test is

administered. As we move up and outwards, the focus moves to the consistency and generalisability of the results across several test administrations and across tasks and raters. The focus then moves to the relevance of the test tasks to the academic domain and to models of academic language ability and to the similarity of the test taker processes elicited by the test to those operating in the academic domain. Moving away from the test itself, validation work proceeds to examine the relevance and appropriateness of decisions based on test scores, and their intended and unintended consequences in light of the institutional policy which will determine how these decisions and test effects play out.

It is important to reiterate that test uses cannot be claimed valid if any of these inferences or bridges is not solid enough to buttress the remaining arguments. So if there are problems at the bottom of the pyramid, the higher aspects of the pyramid cannot be supported.

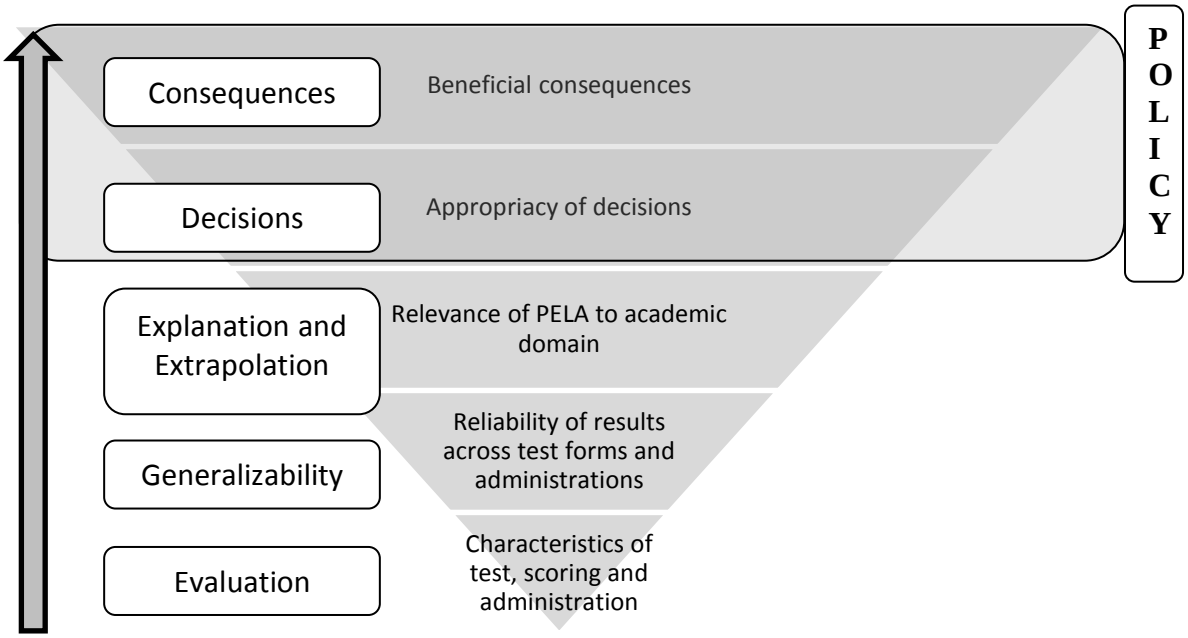

Figure 3. Overview of the building blocks and inferences of a PELA interpretive argument

Each inference in the model above is accompanied by a claim which is outlined in Table 1 below. Underlying a claim, are a series of underlying assumptions. The arrow in Table 1 shows the direction of the series of inferences, moving from the evaluation inference (and its accompanying claim) to test consequences and its associated claim. Research is then collected to support each of the assumptions. To guide this work, we have formulated a series of warrants which were written to be applicable to a range of different post-entry language assessment contexts. These will be discussed in more detail below.

Table 1. Inferences and claims of PELA validity argument

Inference	Claim
Consequences	The consequences of using the PELA and the decisions informed by the PELA are beneficial to all stakeholders.
Decision	Score-based decisions are appropriate and well communicated.
Explanation/Extrapolation	The assessment reflects the targeted language ability construct and provides information on test takers' skills/knowledge and characteristics that is relevant to the academic domain. The test tasks are adequate proxies for those performed in the academic domain.
Generalisability	The assessment yields results that are consistent across assessment contexts (e.g. across test forms, across tasks in the same form and test judges).
Evaluation	The score on the test is an adequate reflection of the observed test behaviour (i.e., scoring procedures are appropriate and clear, and test administration processes are carried out as intended by test designers).

Warrants and supporting data

In the following section, we outline a list of warrants we propose as a basis for PELA validation. The warrants were formulated based on a detailed review of the literature, including other studies using an argument-based approach to validation (e.g. see Chapelle, Enright & Jamieson, 2008; Kane, 2012), a review of the literature on post-entry language assessments and a series of meetings between the authors. We do not claim that the list of warrants is exhaustive or that a validation argument for PELA requires data to support all these warrants. For each warrant, we also offer suggestions on the type of data that can be collected as supporting (or disconfirmatory) evidence. The warrants were formulated to be generally applicable to different PELAs; however, as discussed later in this paper, some might be more relevant than others in certain policy contexts.

Table 2. Warrants and supporting evidence for the evaluation inference in a PELA validity argument

Evaluation inference	
Claim: The score on the test is an adequate reflection of the observed test behaviour	
Warrants	**Sources of supporting evidence**
1. Scoring criteria and rubrics capture relevant aspects of performance.	Review by language testing and domain experts.
2. Raters can implement scoring procedures consistently.	Statistical analysis of test scores.
3. Test administration conditions are clearly articulated and appropriate.	Student questionnaires/interviews; review of test administration protocol; observation of test sessions, interviews with test invigilators.
4. Instructions and tasks are clear to all test takers.	Student questionnaires/interviews
5. Test is pitched at appropriate difficulty level and test tasks/items discriminate consistently between more and less able candidates.	Statistical analysis of test properties (i.e., item difficulty, discrimination, internal consistency).

Table 2 above summarises the warrants as well as the type of supporting evidence needed to support or refute the claim that accompanies the first inference in the interpretative argument. The evidence alluded to in this table focuses on design aspects of the test including scoring rubrics, evidence that the test instructions and tasks are clear to students, data on test administration conditions and statistical evidence of test and item properties to ensure that the test is working as it should. Only if these warrants are supported, is the overall claim of the inference supported. The kinds of evidence listed above, while routinely collected in higher stakes testing contexts, are often lacking for Australian PELAs or at least not made publically available.

Table 3. Warrants and supporting evidence for the generalisation inference in a PELA validity argument

Generalisability inference	
Claim: The assessment yields results that are consistent across assessment contexts	
Warrants	**Supporting evidence**
1. Different test forms are parallel in design.	Review of test specifications and test materials.
2. Appropriate equating procedures are used to ensure equivalent difficulty across forms.	Review of equating reports and statistical procedures used.
3. Sufficient tasks are included to provide stable estimates of test taker ability.	Statistical analysis of scores from a trial test population.
4. Test administration conditions are consistent	Review of procedures; interviews with students and invigilators.

Table 3 above shows the list of warrants and a summary of the type of supporting evidence that needs to be collected to support the generalisability inference. The warrants in this table focus on the generalisability of a single test form and administration in relation to other test forms and administrations of the test. The underlying assumption here is that PELAs exist in parallel forms and that the forms are parallel in test content and statistically equated, such that the result yielded for any candidate will be the same or highly similar regardless of which test form he or she has taken. Supporting evidence for this claim can be found in a review of test documentation, including test specifications, test materials and technical reports as well as a thorough review of test administration procedures and practices. Again, in our experience, such issues are sometimes neglected in PELA contexts and different test forms, if at all available, are often used interchangeably without ascertaining that they are equivalent in difficulty. Only when both the evaluation inference and the generalizability inference have been supported, can the focus move to the explanation and extrapolation inferences, which is presented in Table 4 below.

Table 4. Warrants and supporting evidence for the explanation and extrapolation inferences in a PELA validity argument

Explanation and Extrapolation inferences Claim: The assessment provides information on test takers skills/knowledge and characteristics that is in keeping with understanding of academic language proficiency and relevant to the academic domain. The test tasks are adequate proxies for those performed in the academic domain.	
Warrants	**Sources of supporting evidence**
1. Test results are good predictors of language performance in academic domain.	Correlations between PELA scores and academic performance esp. language-related academic tasks (e.g. essays, oral presentations, GPAs and WAMs).
2. Characteristics of test tasks are similar to those required of students in the academic domain (and those in the language development courses students are placed in).	Comparison of test materials and course materials/course assessment requirements.
3. Linguistic knowledge, processes, and strategies employed by test takers are in line with theoretically informed expectations and observations of what is required in the corresponding academic context.	Test taker verbal protocols gathered during test performance; responses on strategy questionnaires gathered after the test.
4. Scores derived from the test provide sufficient information about candidates' academic language proficiency (i.e., no construct under-representation).	Review of test materials to ensure adequate coverage of academic language domain.
5. Performance on PELA relates to performance on other assessments of academic language proficiency.	Correlation between test scores and scores derived from other validated test instruments measuring similar abilities; correlation

	between test scores and teacher rankings of language proficiency.
6. Tasks do not unfairly favour certain groups of test takers	Expert 'sensitivity' review of test content, statistical bias analyses.

This table presents a range of warrants for the explanation and extrapolation inferences. These focus on comparability of the test tasks and language elicited by the tasks with the type of tasks and language used in the academic domain as well as how accurately the test tasks predict language performance in the academic domain or other tests designed to assess academic language proficiency. The kinds of supporting evidence that can be sought ranges widely, including, for example, a content comparison of the test tasks with language-related tasks in the academic domain (for one such study in another testing context see Moore, Morton & Price, 2012) and correlations between test takers' results and academic performance on particular subjects as well as grade point averages (GPAs) or weighted average marks (WAMs). This latter evidence needs to be interpreted with caution, however, as different academic subjects make different language demands and predictive validation studies generally show a fairly weak overall correlation accounting for no more than 10% of GPA score variance (e.g. see Elder, Bennett & Bright, 2007). Bias investigations (Warrant 6), while rarely reported, are of particular relevance to PELAs given the heterogeneous nature of the test taker population. A validation study by Elder, McNamara and Congdon (2003), for example, investigates the possibility that particular tasks or items on the DELNA might be functioning differently for native and non-native speaker test takers for reasons unrelated to the constructs that the test is designed to measure. Comparisons between PELA scores and results on other similar academic English tests may offer useful evidence that the test is targeting relevant skills, although it may also be useful to compare test takers' cognitive processes and strategies when engaging with PELA tasks to the processes and strategies employed when performing similar language-related tasks in the academic domain. A detailed analysis of student background information in relation to their test results, including an analysis of cohort results from different entry pathways, may be useful in confirming or disconfirming expectations about relative degrees of academic language proficiency in each cohort. For example, we would expect international students from non-English speaking backgrounds to perform less well overall on a PELA than other cohorts made up primarily of native speakers (despite wide variability among the latter group). If results of such investigations accord with expectations we can take this as supporting the validity of our test scores.

If they do not we may need to question whether our test is capturing the relevant abilities.

Table 5. Warrants and supporting evidence for the decision inference in a PELA validity argument

Decision inference Claim: Score-based decisions are appropriate and well communicated.	
Warrants	**Sources of supporting evidence**
1. Students are correctly categorised based on their test scores.	Interviews with key stakeholders (e.g. students, academic staff, learning and teaching staff); review of test results; review of standard-setting activities to set cut-scores; review of academic outcomes for students classified above and below the cut-score.
2. The test results include feedback on test performance and a recommendation.	Review of policy and practice.
3. Recommendation is closely linked to on-campus support.	Review of language development options; interviews with key stakeholders including students.
4. Assessment results are distributed in a timely manner.	Review of practice; interviews with key stakeholders.
5. The test results are available to all relevant stakeholders.	Review of policy and practice.
6. Test users understand the meaning and intended use of the scores	Review of policy and practices, including test website; interviews with test users

Table 5 above summarises the warrants and supporting evidence for the decisions inference. The warrants for this inference focus on the results and recommendations following the assessment. This includes how accurately students are categorized based on the test (Warrant 1), an issue that is often neglected in PELA contexts. Decisions about what constitutes a satisfactory and less than satisfactory level of performance on the test need to be made on a principled basis through a process known as standard-setting which includes consultation with stakeholders and/or comparisons with performance on another validated measure (e.g. Elder & von Randow, 2008; Knoch & Elder, 2009). It is also important to consider how soon the test results are received by students (taking into account their busy academic schedules and the need to enrol in any follow-up language development courses), to examine the structure of the results and recommendations given to test takers, how closely the test recommendation is linked to on-campus support and who has access to the results with reference to available documentation and practice as well as interviews with key stakeholders. The link between test results and on-campus support (Warrant 5) and questions as to who has access to the test results (Warrant 3) are of course a function of institutional policy and the resources

that are made available to implement that policy. Warrant 6 relates to test users' understanding of the policy driving test implementation, and evidence derived from them will shed light on how effectively that policy has been communicated - whether students and their teachers are fully informed about the test, the reasons for testing, how they will be classified and the decisions that will be made based on these classifications. PELAs, as indicated earlier in this paper, are used not for selection but as a mechanism for determining who might benefit from additional language development opportunities. Making this test purpose plain to the test takers and other users is clearly critical if the test is to achieve its goal of enhancing academic success. Read (2008) for example has pointed to the need to avoid stigma in the wording of any communications about the test and argues that this is critical for student uptake.

From test decisions flow consequences – beneficial or otherwise – and these are the subject of the final inference in the PELA validity argument (see Table 6). Consequences, like decisions, are governed by policy.

Table 6. Warrants and supporting evidence for the consequences inference in a PELA validity argument

Consequences inference Claim: The consequences of using the PELA and the decisions informed by the PELA are beneficial to all stakeholders	
Warrants	**Sources of supporting evidence**
1. All targeted test takers sit for the test.	Analysis of test administration data.
2. The test does not result in any stigma or disadvantage for students.	Interviews with students regarding their test attitudes and experiences following the test.
3. Test takers' perceptions of the test and its usefulness are positive.	Interviews with students.
4. The feedback from the test is useful and directly informs their future learning.	Interviews with students and language support teachers.
5. Students act on the test recommendation (i.e. take up the proposed language development strategies).	Review of student uptake data. Interviews with key staff and students regarding reasons for compliance or non-compliance.
6. Follow-up language development options provided for students are appropriate.	Interviews with key stakeholders including students.
7. Learners taking up support options improve their English over the course of their studies.	Comparison of pre- and post-test scores.
8. Students who fail to act on test recommendations are more likely to struggle in their academic studies.	Comparison of academic results of compliant and non-compliant students.

One of the most obvious intended consequences of a PELA policy is that those identified as needing to be tested will show up for the testing session (Warrant 1). However, as Ransom (2009) has demonstrated, based on an analysis of the test-taker population at the University of Melbourne, a policy decision to make the DELA test mandatory did not result in universal uptake. Gathering evidence of such failures or limitations is important in refining the institutional policy, perhaps with an eye to more effective communication of the test purpose and potential benefits to users. Other warrants worthy of exploring in the PELA context are how students perceive the test taking experience, what they make of the feedback provided and whether they take notice of the advice given (Warrant 3). The success of any PELA initiative will depend also on how well students are guided to appropriate follow-up support on campus, whether appropriate courses are available on a continuing basis and can be taken within the limitations of busy course schedules. If, for example, students feel that the language support offered are inconveniently timed or add unhelpfully to their study burden they are unlikely to comply. Monitoring compliance and exploring reasons for any resistance to institutional mandates is essential if claims about the beneficial consequences of a PELA program are to be supported. Bright and von Randow (2004) in a longitudinal study of 18 DELNA test takers cited lack of time and pressures of work among the reasons for non-compliance, along with an expectation by students that lecturers and tutors would be able to give the required support in the mainstream study context. Findings such as these serve to highlight policy issues that need to be addressed in relation to how language development courses are delivered and promoted.
The type of supporting evidence to support or refute the warrants listed above is drawn from interviews with students and other stakeholders, and a review of the available language development courses. Evidence of improved language proficiency as a result of language development courses (Warrant 7) is perhaps the hardest to capture, due to the other factors, such as English exposure and for opportunities practice outside the university setting. Nevertheless such evidence, however tentative, needs to be sought if claims about the benefits of any PELA initiative are to be upheld.

Adapting the framework for local use

The number of warrants listed under each claim in the above framework may be seen as rather daunting to those charged with developing and implementing PELA initiatives within particular institutions. There are inevitably practical constraints on what test developers and administrators can achieve in relation to post entry language assessment and these will influence the type of assessment and the manner in which it can be delivered in a certain context. Scholarly conceptualisations of validity and the validation process have

acknowledged issues of practicality (see e.g. Bachman & Palmer, 2010,) and we understand that these will always play a key part in any evaluation of a PELA. Below, we provide some examples of how different institutions have adapted their local policy in light of practical constraints and consider the implications of these adaptations for the validation process.

With regard to the test taker population, institutions have taken different approaches to which students they target for post-entry assessment. The University of Melbourne, for example, has opted to administer its Diagnostic English Language Assessment (DELA) only to potentially at-risk groups of students, including international students, and those entering the university via non traditional pathways, such as Foundation courses, rather than to the entire student cohort, thereby containing the costs and resources expended.

Other institutions see language problems as being more broadly distributed within the student population. To be sure of identifying all students at risk they deem it important to test all incoming students, regardless of their language or educational background. Because the volume of students that need to be tested in such a scenario is very high, Knoch and Elder (2013), for example, have argued that it is acceptable to consider using indirect (screening) tasks as part of a PELA rather than more time-consuming direct tasks (e.g. listening to a lecture) used in tests such as the DELA.

Some institutions opt to reduce costs of test administration by forgoing the double marking of writing scripts (widely recognised as important for reliability and hence generalisability of inferences) and settling for a single rater only. Or they may decide to exclude from the assessment any skills (e.g. listening) for which direct language development support is not available on campus. The latter approach has obvious implications for any inferences involving extrapolation from the test score to the real-world context, given that such skills may be important in coping with the demands of academic study.

Therefore, when evaluating a PELA, one way to account for practicality constraints is to mount an argument which justifies leaving certain claims unsupported. It is important in each of these cases, however, that the institution closely monitors the area where compromises have been made. At the University of Melbourne, for example, it will be important to examine whether students not targeted by the initial assessment could be at risk (Decision inference, Warrant 1). Similarly, it is important to examine whether a policy to target certain student groups for PELA testing is creating a view of those students as deficient with respect to their peers. This could thereby be seen as discriminatory and hence violates the claim associated with the Consequence

inference, that 'The consequences of using the PELA and the decisions informed by the PELA are beneficial to all stakeholders' (and see in particular Warrant 4.)

There is, however, a risk associated with this process of adapting and selecting from the framework in light of local goals and circumstances, namely that the selection of evidence to be gathered will be based merely on convenience and the desire to prove that the test is valid (a tendency described by Cronbach, 1988 as 'confirmation bias') rather than on a genuine attempt to put the test under the spotlight and explore any validity concerns. Indeed one of the weaknesses of our framework is that it offers little assistance in prioritising between different warrants. When attending to the extrapolation inference for example, should we prioritise correlations with other measure over qualitative evidence of test taking processes? How much validity evidence is enough to support the inferences drawn? Here we must rely on common sense judgements about what evidence is likely to be most crucial and illuminating given available resources.

Furthermore, the question of who is responsible for gathering validation evidence needs to be considered. Clearly validation activity as specified for the evaluation, generalisation and extrapolation inferences should be part and parcel of the test development process. Warrants relating to decisions and consequences however relate to institutional policies and demand evidence from a wider range of sources. Collecting the relevant evidence cannot be left to test developers alone. We propose, then, that when articulating a validation argument for PELA within a specific institution, responsibilities for carrying out key validation activities be clearly specified and allocated in advance. We would also advocate that validation be seen not as a one-off, all-or-nothing initiative but rather as an ongoing process geared not only to demonstrating validity, but also to highlighting areas for improvement of both the test and the policy in which it is embedded.

Concluding remarks

We have argued in this paper for setting up a systematic validation framework which outlines the goals of a PELA initiative, the inferences and interpretations to be made in relation to PELA scores, the warrants associated with these inferences and the evidence required to support or refute the warrants. We believe framework we have laid out is sufficiently general to be applicable to all PELA contexts but, as we have indicated above, will need to be adapted for particular institutional contexts. Such a framework will assist institutions in planning validation activity and keeping track of what has been done and

needs to be done to ensure that the test scores are used and interpreted appropriately and that associated language programs are meeting expectations.

Application of this framework requires ongoing commitment, not only from language test developers and administrators, but also from policy makers in higher education, who need to ensure that validation is seen as part and parcel of the PELA initiative and that adequate resources are available to implement the validation plan. Making such a commitment is necessary if we are to improve on the haphazard processes operating in many institutions and to ensure that PELA initiatives actually achieve their goal of building the quality of students' academic English and ultimately, their capacity to engage with and succeed in their academic studies.

References

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1–34.

Bachman, L. F. & Palmer, A. S. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Bright, C. & von Randow, J. (2004, September.). *Tracking language test consequences: The student perspective.* Paper presented at the Ninth National Conference on Community Languages and English for Speakers of Other Languages (CLESOL 2004), Christchurch.

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

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254–272.

Chapelle, C. Enright, M. & Jamieson, J. (2008) *Building a validity argument for the Test of English as a Foreign Language*. London, UK: Routledge.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun, (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davies, A. & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel, (Ed.). *Handbook of research in second language learning* (pp. 795–814). Mahwah, NJ: Lawrence Erlbaum.

Dunworth, K. (2013). Degrees of proficiency: Building a strategic approach to university students' English language assessment and development. Final report. Canberra: Office for Learning and Teaching.

Elder, C., Bennett, S. & Bright, C. (2007). The role of language proficiency and academic success: perspectives from a New Zealand University**.** *Melbourne Papers in Language Testing*, *12*, 25–57.

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

Elder, C., McNamara, T. & Congdon, P. (2003). Rasch techniques for detecting bias in performance assessments: an example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement*, *4*(2), 181–197.

Elder, C. & von Randow, J. (2008). Exploring the utility of a web-based screening tool. *Language Assessment Quarterly*, 5(3), 173-194.

Fulcher, G. & Davidson, G. (2007). *Language testing and assessment: An advanced resource book*. New York, Routledge.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement* 38, 319-42.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective*, *2*(3), 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan, (Ed.), *Educational Measurement* (4th ed.) (pp. 17-64). Westport, CT: Greenwood Publishing.

Kane, M. (2012). Articulating a validity argument. In G. Fulcher & F. Davidson, (Eds.), *The Routledge handbook of language testing* (pp. 34–47). New York, NY: Routledge.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. Educational *Measurement: Issues and Practice*, *18*, 5–17.

Knoch, U. & Elder, C. (2009). Report on the development and trial of the Academic English Screening Test (AEST). Language Testing Research Centre, University of Melbourne: unpublished report.

Knoch, U. & Elder, C. (2013). Knoch, U. & Elder, C. (July 3-5, 2013*). Post-entry English language assessments at university: How diagnostic are they?* Language Testing Research Colloquium, Seoul, Korea.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA and Oxford: Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: American Council on Education and Macmillan.

Moore, T., Morton, J. & Price, S. (2012). Construct validity in the IELTS academic reading test: A comparison of reading requirements in IELTS test items and in university study, (pp. 120-211). In L. Taylor, (Ed.), *Studies in language testing 34*, Cambridge: Cambridge University Press.

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

Ransom, L. (2009). Implementing the post-entry English language assessment policy at the University of Melbourne: rationale, processes and outcomes. *Journal of Academic Language and Learning*, *3*(2), 13–25.

TEQSA (2011). Higher Education Standards Framework (Threshold Standards). Tertiary Education Quality and Standards Agency Act 2011. Retrieved from http://www.comlaw.gov.au/Details/F2013C00169/Download

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. Hornberger, (Eds.), *Encyclopedia of language and education* (2nd ed.) (Vol. 7: Language Testing and Assessment). New York: Springer.