

## Chapter 2

# Examining the Validity of a Post-Entry Screening Tool Embedded in a Specific Policy Context

Ute Knoch, Cathie Elder, and Sally O'Hagan

**Abstract** Post-entry English language assessments (PELAs) have been instituted in many higher education contexts for the purpose of identifying the language needs of the linguistically and culturally diverse population of students entering English-medium universities around the world. The current chapter evaluates the validity of the Post-entry Assessment of Academic Language (PAAL), a PELA screening test trialled at two faculties (Engineering and Commerce) at a large Australian university. The chapter follows the approach to building an interpretative validity argument as outlined by Knoch and Elder (2013) and, by way of validity evidence, draws on data from the development phase of the test, its trial implementation and the evaluation phase of the trial. Evidence gathered during the development phase supports the first three inferences in the validity argument (evaluation, generalizability and explanation) and shows that the PAAL is technically adequate, relevant to the academic domain and effective in balancing the demands of efficiency and diagnostic sensitivity. Data from the trial however reveal problems in relation to the final two inferences in the validity argument concerning the relevance and appropriateness of decisions based on test scores and the consequences or perceived consequences of test use for the stakeholders involved. It is shown that these problems stem mainly from limitations in the institutional policy. We conclude that implementing the same test in a more hospitable policy setting might produce very different outcomes and assert the importance of evaluating tests in their policy contexts.

**Keywords** Post-entry language assessment • Screening test • Test validation • PAAL • English for academic purposes • Higher education

---

U. Knoch (✉) • C. Elder • S. O'Hagan

Language Testing Research Centre, University of Melbourne, Melbourne, Australia

e-mail: [uknoch@unimelb.edu.au](mailto:uknoch@unimelb.edu.au); [caelder@unimelb.edu.au](mailto:caelder@unimelb.edu.au); [sohagan@unimelb.edu.au](mailto:sohagan@unimelb.edu.au)

# 1 Introduction

While universities in major English-speaking countries have long-established English language entry requirements for international students in the form of a minimum score on an admissions test such IELTS or TOEFL, there is growing concern amongst academics that these minimum cut-scores may be too low to ensure that incoming students can cope with the language demands of their university study (Ginther and Elder 2014). Raising entry thresholds may not be deemed acceptable however, given the university's reliance on revenue from these fee-paying students on the one hand and, on the other, the fear of excluding otherwise academically talented students whose language skills might be expected to improve over time. In any case, as access to higher education broadens and diversifies, it has become evident that increasing numbers of domestic students, whether from English or non-English speaking backgrounds, may also be linguistically at risk in their academic study – even more so in some cases than their international student counterparts (Read 2015). Since limited command of English is associated with low retention rates and poor academic outcomes, addressing the issue has become a priority. In response to this situation many universities in Australia and elsewhere have instituted a Post-entry English Language Assessment (PELA) to identify the language needs of those who have been admitted to the university so that appropriate strategies can be devised to enhance their chances of academic success (Dunworth 2009).

Given the diverse range of students who may experience difficulties with academic English and the variable nature of the difficulties they face, the validity challenges for PELA design and use are considerable. A PELA needs to

- target the full range of potentially at risk students so that the chances of individuals falling through the net are minimized;
- capture relevant information from these individuals so that their particular language needs are clear;
- communicate this information to all relevant parties in a meaningful, timely and sensitive fashion;
- provide opportunities for the identified needs to be addressed; and, ultimately,
- consider whether the testing initiative is fulfilling its intended aim of improving student outcomes.

Different PELA models have been adopted by different institutions but evidence for their validity and efficacy is scarce and has been gathered in a piecemeal fashion. A paper by Knoch and Elder (2013) outlines a framework for PELA validation activity drawing on the influential argument-based models proposed by Kane (1992, 2006) and Bachman and Palmer (2010). The framework sets out a series of inferences for which supporting evidence needs to be collected to claim that the scores and score interpretations based on a PELA are valid. For this, Knoch and Elder (2013) formulated a series of warrants for each inference in the PELA interpretive argument (evaluation, generalizability, explanation and extrapolation, decisions and consequences) which might be applicable to a range of different PELA contexts.

The authors also showed how the institutional policy determines how the decisions and consequences associated with a certain PELA play out. This crucial role of policy in PELA implementation will be taken up later in this chapter.

This framework has subsequently been applied to the evaluation of the Diagnostic English Language Needs Assessment at the University of Auckland (Read 2015) and also to two well-established Australian PELAs claiming either explicitly or implicitly to be diagnostic in their orientation: the Measuring the Academic Skills of University Students (MASUS) test at the University of Sydney (Bonanno and Jones 2007) and the Diagnostic English Language Assessment (DELA) at the University of Melbourne (Knoch and Elder 2016). In the latter study, Knoch and Elder consider the different inferences (evaluation, generalization, explanation/extrapolation, decisions and consequences) that underpin claims about diagnostic score interpretation in a PELA context and the associated warrants for which evidential support is required. The findings of the evaluation revealed that support for some of these warrants is lacking and that neither instrument can claim to be fully diagnostic. Although each PELA was found to have particular strengths, the claim that each provides diagnostic information about students which can be used as a basis for attending to their specific language needs is weakened by particular features of the assessment instruments themselves, or by the institutional policies determining the manner in which they are used. The rich diagnostic potential of the MASUS was seen to be undermined by limited evidence for reliability and also by the lack of standardized procedures for administration. The DELA, while statistically robust and potentially offering valid and useful information about reading, listening and writing sub-skills, was undermined by the policy of basing support recommendations on the student's overall score rather than on the sub-skill profile. The authors concluded that the DELA 'really functions as a screening test to group students into three broad bands – at risk, borderline or proficient – and there is no obvious link between the type of support offered and the particular needs of the student' (p. 16).

A further problem with both of these PELAs is that they are not universally applied: MASUS is administered only in certain departments and DELA is administered only to categories of students perceived to be academically at risk. Furthermore, there are few or no sanctions imposed on students who fail to sit the test, on the one hand, or fail to follow the support recommendations, on the other.

Similar problems of uptake were identified with the Diagnostic English Language Needs Assessment (DELNA) at the University of Auckland, a two-tiered instrument which includes both an initial screening test involving indirect, objectively-scored items and a follow-up performance-based diagnostic component, with the latter administered only to those falling below a specified threshold on the former. One of the challenges faced was convincing students who had performed poorly on the screening component to return for their subsequent diagnostic assessment (Elder and von Randow 2008). While uptake of the diagnostic component has improved over time, as awareness of the value of the DELNA initiative has grown, the success of the two-tiered system has been largely due to the institution providing resources for a full-time manager, a half-time administrator and a half-time adviser whose

jobs include raising awareness of the initiative among academic staff and students, pursuing those who failed to return for the second round of testing, and offering one on one counseling on their English needs as required.

Given that many institutions are unwilling to make a similar commitment of resources to any PELA initiative, an alternative approach, which attempts to build on the strengths of previous models and to address their weaknesses, was devised by the authors of this chapter.

This paper offers an evaluation of the resulting instrument, known as the Post-entry Assessment of Academic Language (PAAL), based on the PELA evaluation framework referred to above. PAAL is the name adopted for a form of an academic English screening test known historically, and currently in other contexts, as the Academic English Screening Test (AEST). PAAL is designed to provide a quick and efficient means of identifying those students in a large and linguistically diverse student population who are likely to experience difficulties coping with the English language demands of academic study, while at the same time providing some diagnostic information. In the interests of efficiency it combines features of the indirect screening approach adopted at the University of Auckland (Elder and von Randow 2008) with a single task designed to provide some, albeit limited, diagnostic information, removing the need for a second round of assessment. It is based on the principle of universal testing, to allow for all at risk students to be identified, rather than targeting particular categories of students (as was the case for DELA) and builds on over 10 years of development and research done at the University of Melbourne and the University of Auckland.

## **2 Background to the Development and Format of the AEST/ PAAL**

The AEST/PAAL was developed at the Language Testing Research Centre (LTRC) at the University of Melbourne in early 2009. It was initially commissioned for use by the University of South Australia; however, the rights to the test remain with the LTRC.

The test is made up of three sections: a text completion task, a speed reading task and an academic writing task, all completed within a 1-h time frame. (For a detailed account of the initial development and trialling, refer to Elder and Knoch 2009). The writing task was drawn from the previously validated DELA and the other two tasks were newly developed. The 20-min text completion task consists of three short texts and uses a C-test format (Klein-Braley 1985), in which every second word has been partially deleted. Students are required to reconstruct the text by filling in the gaps. The speed reading task, an adaptation of the cloze-elide format used for the screening component of DELNA at the University of Auckland (Elder and von Randow 2008), requires students to read a text of approximately 1,000 words in 10 min and identify superfluous words that have been randomly inserted. The writing task is an argumentative essay for which students are provided with a topic and have 30 min

to respond with 250–300 words (see Elder et al. 2009 for further detail). The text completion and speed reading tasks, used for screening purposes, are objectively scored and the writing task, intended for diagnostic use, is scored by trained raters using a three-category analytic rating scale. The test scores from the two screening components place students in one of three support categories as follows:

- **Proficient.** Students scoring in the highest range are deemed to have sufficient academic English proficiency for the demands of tertiary study.
- **Borderline.** Students scoring in the middle range are likely to be in need of further language support or development.
- **At Risk.** Students scoring in the lowest range are deemed likely to be at risk of academic failure if they do not undertake further language support and development.

It is recommended, for the sake of efficiency, that the writing component of the AEST/PAAL be completed by all students but be marked only for those scoring in the Borderline and At Risk categories on the two screening components. The writing thus serves to verify the results of the screening components for Borderline students (where categorization error is more likely) and potentially yields diagnostic information for less able students so that they can attend to their language weaknesses. The AEST/PAAL was however designed primarily as a screening test which efficiently groups students into the three categories above. For reasons of practicality and due to financial constraints the test was not designed to provide detailed feedback to test takers beyond the classification they are placed into and information about the support available to them on campus.

Following the development and trial of the prototype outlined in Elder and Knoch (2009), three more parallel versions of the test were developed (Knoch 2010a, b, 2011), initially for use at the University of South Australia, as noted above. In 2012, following a feasibility study (Knoch et al. 2012a), the University of Melbourne's English Language Development Advisory Group (ELDAG) supported a proposal by the LTRC to put the test online for eventual use at the University of Melbourne. This was funded in 2012 by a university Learning and Teaching Initiative Grant. The online platform was then developed by Learning Environments, a group of IT specialists supporting online learning and teaching initiatives at the University. Following the completion of the platform, the online delivery was tested on 50 test takers who had previously taken the University's DELA (Knoch et al. 2012b). The students also completed a questionnaire designed to elicit their experiences with the online platform. This small trial served as an extra check to verify cut-scores (between the Proficient, Borderline and At risk levels) which had been set during the development of the test using performance on the DELA as the benchmark (see below, and for further detail, Elder and Knoch 2009).

Based on the results of the small trial, a number of technical changes were made to the delivery of the test and in Semester 2, 2013, a trial on two full cohorts was undertaken (again funded by a Teaching and Learning Initiative grant) (Knoch and O'Hagan 2014). The trial targeted all newly incoming Bachelor of Commerce and Master of Engineering students as these two cohorts were considered to be

representative of students in very different disciplines and at the undergraduate and graduate levels. Following the trial, students were asked to complete an online questionnaire and a subset of students from both faculties took part in focus groups. For the purpose of the trial, the AEST was renamed as the Post-entry Assessment of Academic Language (PAAL) and this is the name we will use for the remainder of the chapter.

### 3 Methodology

As the overview of the historical development of the PAAL above shows, a number of trials and data collections have been conducted over the years. In this section, we will describe the following sources of data which we will draw on for this paper:

1. the PAAL development trial
2. the small trial of the online platform
3. the full trial on two student cohorts

#### 3.1 *PAAL Development Trial*

There were 156 students who took part in the development trial of the PAAL, 71 from the University of South Australia and 85 from the University of Melbourne. All students were first year undergraduates and were from a range of L1 backgrounds, including a quarter from English-speaking backgrounds.

Test takers at both universities took the following components of the PAAL:

- Part A: Text completion (C-test with 4<sup>1</sup> texts of 25 items each) – 20 min
- Part B: Speed reading (Cloze elide with 75 items) – 10 min
- Part C: Writing task – 30 min

Test takers at the University of Melbourne had previously taken the DELA as well, and therefore recent scores for the following skills were also available:

- Reading: 46 item reading test based on two written texts – 45 min
- Listening: 30 item listening test based on lecture input – 30 min

#### 3.2 *Small Trial of the Online Platform*

Fifty students from the University of Melbourne were recruited to take part in the small trial of the online platform developed by Learning Environments. The students completed a full online version of the PAAL from a computer or mobile

---

<sup>1</sup>A C-test with four texts is used for trial administrations whilst the final test form only includes three texts.

device at a place and time convenient to them. The final format of the PAAL adopted for the online trial and the full trial was as follows:

- Part A: Text completion (C-test with 3 texts of 25 items each) – 15 min
- Part B: Speed reading (Cloze elide with 75 items) – 10 min
- Part C: Writing task – 30 min

Following the trial, 49 of the participants completed an online questionnaire designed to elicit information about what device and browser they used to access the test, any technical issues they encountered, and whether the instructions to the test were clear and the timer visible at all times.

### ***3.3 Full Trial on Two Student Cohorts***

The full trial was conducted on two complete cohorts of commencing students at the beginning of Semester 2, 2013: Bachelor of Commerce (BCom) and Master of Engineering (MEng).

In the lead-up to the pilot implementation, extensive meetings were held with key stakeholders, in particular Student Centre staff from the respective faculties. It became evident very early in these discussions that universal testing is not possible as there is no mechanism to enforce this requirement on the students. Although it was not compulsory, all students in participating cohorts were strongly encouraged through Student Centre communications and orientation literature to complete the PAAL. At intervals during the pilot period, up to three reminder emails were sent by the respective Student Centres to remind students they were expected to take the test.

On completing the PAAL, all students were sent a report containing brief feedback on their performance and a recommendation for language support according to their results. The support recommendations were drafted in consultation with the Student Centres and Academic Skills<sup>2</sup> to ensure recommendations were in accord with appropriate and available offerings. The reports were emailed by the Language Testing Research Centre (LTRC) to each student within 1–2 days of their completing the PAAL. Cumulative spreadsheets of all students' results were sent by the LTRC to the Student Centres on a weekly basis throughout the pilot testing period.

The PAAL was taken by 110 BCom students, or 35 % of the incoming cohort of 310 students. In the MEng cohort, PAAL was taken by 60 students, comprising 12 % of the total of 491. The level of uptake for the BCom cohort was reported by the Commerce Student Centre as favourable compared with previous Semester 2 administrations of the DELA. Lower uptake for the MEng cohort was to be expected since traditionally post-entry language screening has not been required for graduate students at the University of Melbourne.

---

<sup>2</sup>The unit responsible for allocation and delivery of academic language support at the University.

The full trial was followed up with an evaluation in which feedback was sought from University stakeholders in order to develop recommendations for the best future form of the PAAL. Feedback came from students in the trial by means of a participant questionnaire and focus groups. Face-to-face and/or email consultation was used to gather feedback from Student Centres and Academic Skills.

Student consultation commenced with an online survey distributed to all pilot participants 2 weeks after taking the PAAL. Responses were received from 46 students, representing a 27 % response rate. Survey respondents were asked for feedback on the following topics: the information they received about the PAAL prior to taking the assessment; their experience of taking the PAAL (technical aspects, task instructions, face validity and difficulty of tasks); the PAAL report (results and recommendations); and the options for support/development and follow-up advice after taking the PAAL.

To gather more detailed feedback on these aspects of the PAAL, and to give students an opportunity to raise any further issues, four focus groups of up to 60 min duration were held: 20 students attended 1 of 4 faculty-specific focus groups, with an average of 5 students in each group. Group discussion was structured around the themes covered in the survey and participants were given the opportunity to elaborate their views and to raise any other issues relating to the PAAL that were of interest or concern to them.

## 4 Results

The remainder of the chapter will present some of the findings from the multiple sources of evidence collected. We will organize the results under the inferences set out by Knoch and Elder (2013) and have summarized the warrants and evidence in a table for each inference at the beginning of each section. Below each table, we describe the different sources of evidence in more detail and present the results for each.

### 4.1 *Evaluation*

Table 2.1 summarizes the three key warrants underlying the Evaluation inference. Evidence collected to find backing for each warrant is summarized in the final column.

To find backing for the first warrant in Table 2.1, the statistical properties of the PAAL were evaluated in the original trial as well as during the development of subsequent versions. Table 2.2 summarizes the Cronbach alpha results, which are all fairly consistent across the four versions. We also found a consistent spread of candidate abilities between new versions and the prototype version and a good spread of item difficulty.



**Table 2.1** Warrants and related evidence for the Evaluation inference

Evaluation	
Warrants	Evidence
1. The psychometric properties of the test are adequate	Psychometric properties of the test as reported in the initial development report (Elder and Knoch 2009) and subsequent development reports (Knoch 2010a, b, 2011)
2. Test administration conditions are clearly articulated and appropriate	Responses to feedback questionnaires from the small trial and full trial (Knoch and O’Hagan 2014)
3. Instructions and tasks are clear to all test takers	Responses to feedback questionnaires and focus groups from the full trial (Knoch and O’Hagan 2014)

**Table 2.2** Reliability estimates for the PAAL test versions

	Text completion (k = 75)	Speed reading (k = 75)	Combined screening (k = 150)	Writing
Version 1 (prototype)	.95	.96	.97	.883
Version 2	.92	.96	.97	n/a
Version 3	.93	.98	.98	n/a
Version 4	.95	.97	.98	n/a

The reliability statistics for the writing task used in this trial were also within acceptable limits for rater scored writing tasks, as was the case for previous versions (Elder et al. 2009).

To examine the second warrant in Table 2.1, we scrutinized the responses from the feedback questionnaires from the small and the full trial. The small trial of the online capabilities showed that there were several technical issues that needed to be dealt with before the test could be used for the full trial. For example, slow loading time tended to be an issue and some participants had experiences of the site ‘crashing’ or losing their connection with the site. The trial also indicated the need to further explore functionality to enable auto-correction features of some browsers to be disabled and adjustments to be made to font size for small screen users. In addition, feedback from trial participants indicated that fine-tuning of the test-taker interface was required. For example, some participants reported problems with the visibility/position of the on-screen timer and with the size of the text box for the writing task. The results of the trial further showed that there were variations in functionality across different platforms and devices (most notably, the iPad). Following this trial of the online capabilities of the system, a number of technical changes were made to the online system before the full trial was conducted.

Overall, the findings of the student survey and focus groups conducted following the full trial indicated that students’ experiences of the online testing system were mostly positive in terms of accessibility of the website, clarity of the task instructions, and timely receipt of their report. Few students reported any technical problems, although there were a small number of students who found the system ‘laggy’, or slow to respond to keystrokes, and some reported that they had lost their internet

connection during the assessment. Overall, the purpose of the assessment and the benefits of taking it were clear to participating students and almost all of them appreciated being able to take the assessment from home in their own time.

The final warrant investigates whether students understood all the instructions and whether the task demands were clear. The questionnaire results from the two trials show that the students commented positively about these two areas.

In sum, the Evaluation inference was generally supported by the data collected from the different sources. The statistical properties of the PAAL were excellent, and the administration conditions suited the students and were adequate for the purpose, despite a few smaller technical problems which may have been caused by the internet rather than the PAAL software. The task demands and task instructions seemed clear to the test takers.

## 4.2 Generalizability

Table 2.3 lists the key warrants and associated supporting evidence we will draw on in our discussion of the Generalizability inference.

The first warrant supporting the Generalizability inference is that different test forms are parallel in design. The PAAL currently has four parallel forms or versions (and two more are nearly completed), all of which have been based on the same specification document. The psychometric results from the development of Versions 2, 3 and 4 show that each of these closely resembles the prototype version (Version 1).

The second warrant is that appropriate equating methods are used to ensure equivalence of test forms. The development reports of Versions 2, 3 and 4 outline the statistical equating methods that have been used to ensure equivalence in the meaning of test scores. Each time, a new version was trialed together with the anchor version and Rasch analysis was used to statistically equate the two versions.

**Table 2.3** Warrants and related evidence for the Generalizability inference

Generalizability	
Warrants	Evidence
1. Different test forms are parallel in design	Review of test features and statistical evidence from reports of the development of parallel versions (Knoch 2010a, b, 2011)
2. Appropriate equating procedures are used to ensure equivalent difficulty across test forms	Review of equating evidence from reports of the development of parallel versions (Knoch 2010a, b, 2011)
3. Sufficient tasks are included to provide stable estimates of test taker ability	Psychometric properties of the test as reported in the initial development report (Elder and Knoch 2009) and subsequent development reports (Knoch 2010a, b, 2011)
4. Test administration conditions are consistent	Discussion of test delivery and results from the survey of the full trial (Knoch and O'Hagan 2014)

Statistically equating the writing tasks is more difficult as no suitable anchor items are available and only one writing task is included. However, the developers of the writing task attempt to closely stay true to the test specifications and small trials of new writing versions are carefully evaluated by a team of test developers to ensure they are as equivalent in design as possible and are eliciting assessable samples of writing performance from test candidates. Successive administrations of writing tasks for the DELA (from which the PAAL writing task is drawn) have shown stable estimates over different test versions as noted above.

The third warrant is that sufficient tasks are included to arrive at stable indicators of candidate performance. Each PAAL has 150 items, 25 for each of the three texts which make up the C-test and 75 in the cloze elide, as well as one writing task. As the PAAL is a screening test, the duration of 1 h is already at the upper limit of an acceptable amount of administration time. It is therefore practically impossible to add any more tasks. However, the trials have shown that the test results are fairly reliable indicators of test performance, with students being classified into the same categories when taking two parallel forms of the test.

The final warrant supporting generalizability relates to the consistency of the test administration. As students can take the test in their own time at a place of their choosing, it is likely that the conditions are not absolutely consistent. For example, a student might choose to take the test in a student computer laboratory that is not entirely free of noise, or at home in quiet conditions. However, due to the low stakes of the test, any differences in test taking conditions are probably not of great concern. Due to the fact that the test is computer-delivered, the timing and visual presentation of the test items are likely to be the same for all students.

By and large, it seems that the Generalizability inference is supported by the evidence collected from our trials.

### ***4.3 Explanation and Extrapolation***

Table 2.4 presents the warrants underlying the Explanation and Extrapolation inferences as well as the evidence we have collected.

The first warrant states that test takers' performance on the PAAL relates to their performance on other assessments of academic language proficiency. During the development of the prototype version of the PAAL, the cohort of students from the University of Melbourne had already taken the Diagnostic English Language Assessment (DELA) and their results could therefore be compared directly with their performance on the PAAL.

Table 2.5 presents the correlational results of the two PELA tests. It can be seen that overall screening test results correlated significantly with both the DELA overall raw scores and the DELA scaled scores.

The second warrant states that the scoring rubric captures relevant aspects of performance. The scoring rubric used to rate the writing performances has been developed on the basis of test developers' intuitions from their experience in EAP

**Table 2.4** Warrants and related evidence for the Explanation and Extrapolation inferences

Explanation and Extrapolation	
Warrants	Evidence
1. Performance on the PELA relates to performance on other assessments of academic language proficiency	Correlational results from the development report (Elder and Knoch 2009)
2. Scoring criteria and rubrics capture relevant aspects of performance	Review of the literature on academic writing
3. Test results are good predictors of language performance in the academic domain	No data collected
4. 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)	No data collected
5. 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	No data collected
6. Tasks do not unfairly favor certain groups of test takers	No data collected

**Table 2.5** DELA/AEST correlations (N=156)

	C-test	Cloze elide	Screening total
DELA average (raw scores)	.772**	.721**	.809**
DELA average (scaled scores)	.775**	.699**	.797**

\* $p < .05$ , \*\* $p < .01$

contexts as well as on a careful review of current practice in assessing writing in the academic domain. The criteria on the scale (organization and style, content and form) are commonly used in the assessment of academic writing and the level descriptions have been refined over the years to assist raters in better differentiating between candidates. Due to the nature of the task and the time limit of 30 min, the writing task captures only a limited sample from the candidates and there is no criterion specifically measuring the use of input reading material (which is in any case limited on this task – consisting only of a series of dot-point statements giving ideas for and against the proposition around which the argument is to be formulated). The ability to integrate reading input in writing is of course an important skill in academic writing but the benefits of assessing this ability needed to be weighed against the costs of devising and rating a more elaborate and time-intensive task involving extensive reading input. Although the task and its scoring rubric may somewhat under-represent the academic writing construct, the writing rating scale goes at least some way towards measuring relevant writing skills for the academic domain.

The third warrant states that test scores are good predictors of performance in the academic domain. No data in support of this warrant was collected as part of this study; however, an unpublished internal report (Group 2012) examining the relationship of the Diagnostic English Language Assessment (DELA) and students'

performance in their first year (as measured through WAMs<sup>3</sup>) shows that the DELA (which correlates strongly with the PAAL) is a very strong predictor of WAMs. The study clearly shows that a higher score on DELA is associated with higher WAMs and that a higher DELA score is associated with a lower risk of failing.

The fourth warrant states that the task types in the PELA are similar to those required of students in the academic domain. In the case of the PAAL, the test designers set out to develop a screening test which would be automatically scored and practical for test takers. It was therefore not possible to closely model the kinds of tasks test takers undertake in the academic domain (e.g. listening to a lecture and taking notes). However, the tasks chosen were shown to be very good predictors of the scores test takers receive on the more direct language tasks included in the DELA and it was therefore assumed that these indirect tasks could be used as surrogates. Similarly, warrant five sets out that the test takers' cognitive processes would be similar when taking the PELA and when completing tasks in the academic domain. Again, due to the very nature of the test tasks chosen, backing for this warrant might be difficult to collect. However, studies investigating the cognitive processes of test takers completing indirect tasks such as C-tests and cloze elide (e.g. Matsumura 2009) have shown that test takers draw on a very wide range of linguistic knowledge to complete these tasks, including lexical, grammatical, lexico-grammatical, syntactic and textual knowledge.

As for the final warrant, potential evidence has yet to be gathered from a larger test population encompassing students from different backgrounds, including native-English speaking (NES) and non-native speaking (NNES) students, and those in university Foundation courses, who may be from low literacy backgrounds or have experienced interrupted schooling. A previous study by Elder, McNamara and Congdon (2003) in relation to the DELNA screening component at Auckland would suggest that such biases may affect performance on certain items but do not threaten the validity of the test overall. Nevertheless, the warrant of absence of bias needs to be tested for this new instrument.

In sum, it would seem that the warrants for which evidence is available are reasonably well supported, with the caveat that the scope and screening function of the PAAL inevitably limits its capacity to fully represent the academic language domain.

## 4.4 Decisions

Table 2.6 sets out the warrants and associated evidence for the Decisions inference.

---

<sup>3</sup>WAM (weighted average mark) scores are the average mean results for students' first year course grades. The results of this in-house study are from an unpublished report undertaken for the University of Melbourne English Language Development Advisory Group committee which oversees the English language policy of the University of Melbourne.

**Table 2.6** Warrants and related evidence for the Decisions inference

Decisions	
Warrants	Evidence
1. Students are correctly categorized based on their test scores	Review of standard-setting activities
2. The test results include feedback on test performance and a recommendation	Evidence from the questionnaires and focus groups of the full trial
3. The recommendation is closely linked to on-campus support	Review of institutional policy and evidence from focus groups of the full trial (Knoch and O'Hagan 2014)
4. Assessment results are distributed in a timely manner	Review of test documentation and evidence from focus groups of the full trial
5. The test results are available to all relevant stakeholders	Review of test procedures
6. Test users understand the meaning and intended use of the scores	Evidence from questionnaires and focus groups of the full trial

The first warrant in the Decision inference states that the students are categorized correctly based on their test score. Finding backing for this warrant involved two standard-setting activities. The first was conducted as part of the development of the prototype of the PAAL. A ROC (Receiver Operating Characteristics curve) analysis, a technique for setting standards, was used to establish optimum cut-scores or thresholds on the screening components of the test (c-test and cloze elide). A number of alternative cut-scores were proposed using either a specified DELA Writing score or an overall DELA score (representing the average of reading and listening and writing performance) as the criterion for identification of students as linguistically at risk. While these different cut-scores vary in sensitivity and specificity (see Elder and Knoch 2009 for an explanation of these terms), they are all acceptably accurate as predictors, given the relatively low stakes nature of the PAAL. Moreover, the level of classification accuracy can be improved through the use of the writing score to assist in decisions about borderline cases.

A further standard-setting exercise was conducted in preparation for the full trial. To set the cut-scores for the three result categories outlined in the previous section (i.e. 'proficient', 'borderline', 'at risk'), a standard-setting exercise was conducted with a team of trained raters at the Language Testing Research Centre using the writing scripts from the small trial. All 50 writing scripts were evaluated by the raters individually, evaluations were compared, and rating decisions moderated through discussion until raters were calibrated with each other and agreement was reached on the placement of each script in one of three proficiency groups: high, medium or low. To arrive at the two cut-scores needed, i.e. between 'proficient' and 'borderline', and between 'borderline' and 'at risk', we used the analytic judgement method (Plake and Hambleton 2001), a statistical technique for identifying the best possible point for the cut-score. Based on this, the cut-scores from the development trial were slightly shifted.

The second warrant states that test takers receive feedback on their performance and a recommendation. The feedback component following the PAAL is minimal, a fact that was criticized by the participants in the full trial. Students expressed disappointment with the results statement given in the report, describing it as somewhat generic and lacking in detail. Students in general would have preferred more diagnostic feedback to guide their future learning. Many also indicated they would have liked to discuss their report with an advisor to better understand their results, and to learn more about support opportunities.

Concerns were also raised about the vagueness of the support recommendation given in the report, with many students wanting a clearer directive for what was required of them. Many students stated they would have liked to receive follow up advice on how to act on the recommendation, with many reporting that they did not know how to access an advisor, or that they had received advice but it had not met their expectations.

The third warrant states that the recommendation is closely linked to on-campus support. Availability of appropriate support was identified as a problem by students who felt that offerings were not suited to their proficiency, level of study or academic discipline, or were otherwise not appropriate to their needs. Some students were also concerned that, in accessing the recommended support, they would incur costs additional to their course fees. Where a credit-bearing course was recommended, students expressed concerns about the implications for their academic record of failing the course.

The fourth warrant states that the assessment results are distributed in a timely manner. This was the case during the full trial, with results being distributed within 1–2 days of a student taking the assessment. Accordingly, in the evaluation of the full trial, students commented positively on the timely manner in which the results were distributed.

The next warrant relates to the availability of assessment results. During the full trial implementation, all accessible stakeholders were made aware of the fact that assessment results could be requested from the Language Testing Research Centre. Students were sent their results as soon as possible, and the Student Centres of the two cohorts were regularly updated with spreadsheets of the results. Unfortunately, because of the size of the cohorts, it was not possible to identify lecturers who would be responsible for teaching the students in question, and therefore lecturers may not have been aware of the fact that they could request the results.

The final warrant states that test users understand the meaning and the intended uses of the scores. The results of the evaluation of the full trial indicate that this was not an issue, at least from the test-taker perspective. The purpose of the assessment and the benefits of taking it were generally clear to students and students in the focus groups indicated that they all understood that the test was intended to be helpful, that it was important but not compulsory and that it did not affect grades.

Overall, the Decisions inference was only partially supported. We could find support for some aspects, including the categorization of the test takers, the expedient handling of test scores and that test takers generally understood their meaning. No data was collected from other test users, however, so the extent to which they

understood the meanings and intended uses of the scores requires further investigation. Other aspects relating to the feedback profile, the recommendation and the close link to on-campus support were not supported.

4.5 Consequences

Table 2.7 outlines the warrants of and associated evidence for the Consequences inference.

The first warrant underlying the Consequences inference states that all targeted test takers take the test, since this was the idea behind the streamlined test design (which was designed to be administered universally to all new students). During the full trial it became evident that institutional policy at the University of Melbourne makes it impossible to mandate such an assessment because it goes beyond content course requirements. Of the undergraduate Bachelor of Commerce cohort, only 110 students out of 310 (35 %) took the assessment. The numbers were even lower for the Master of Engineering cohort, where only 60 students out of 491 (12 %) of students took the PAAL. Evidence from the focus groups also showed that students differed in their understandings of whom the PAAL is for, with many believing it to be intended for ‘international’ students only. There was overall no sense among students that the assessment was meant to be universally administered.

The next warrant states that test takers’ perceptions of the assessment are positive and that they find the assessment useful. The data from the full trial, some findings of which have already been reported under the Decision inference above, show mixed results. Students were generally positive about the ease of the test administration and the information provided prior to taking the test, but were less positive about the level of feedback provided and the follow-up support options available. Therefore, this warrant is only partially supported.

The next warrant states that the feedback from the assessment is useful and directly informs future learning. It is clear from the data from the full trial that the

Table 2.7 Warrants and related evidence for the Consequences inference

Consequences	
Warrants	Evidence
1. All targeted test takers sit for the test	Evidence from the full trial
2. The test does not result in any stigma or disadvantage for students	Evidence from the questionnaires and focus groups
3. Test takers’ perceptions of the test and its usefulness are positive	Evidence from the questionnaires and focus groups
4. The feedback from the test is useful and directly informs future learning	Evidence from the questionnaires and focus groups
5. Students act on the test recommendation	Evidence from the questionnaires, focus groups of the full trial and follow-up correspondence with Student Centres



students did not find the feedback from the assessment particularly useful; however, it is important to remember that the PAAL is intended as a screening test, which is designed to identify students deemed to be at risk in minimum time and with minimum financial expenditure. When the test was designed, it was clear that no detailed feedback would be possible due to financial as well as practical limitations. While the analytically scored writing task potentially allowed for more detailed feedback, the resources to prepare this feedback were not available for a large cohort of students such as the one that participated in the full trial. More suitable online or on-campus support options would also have improved the chances of this warrant being supported. Unfortunately, offering more varied support provisions is costly and, in the current climate of cost-savings, probably not a viable option in the near future.

The final warrant states that students act on the score recommendation. Approximately 15% of students taking the PAAL as part of the full trial were grouped into the 'at risk' group. It is not clear how many of these students acted upon the recommendation provided to them by enrolling in a relevant English language support course but historically the compliance rates at the University of Melbourne have been low. We suspect that the same would apply to this cohort for many reasons, including the limited array of support options and the lack of any institutional incentive or requirement to take such courses.

Overall, it can be argued that the Consequences inference was either not supported by the data collected in the full trial or that the relevant evidence was lacking.

## 5 Discussion and Conclusion

The chapter has described a new type of PELA instrument, which builds on previous models of PELA adopted in the Australian and New Zealand context. The online PAAL, taking just 1 h to administer, was designed to be quick and efficient enough to be taken by a large and disparate population of students immediately following their admission to the university, for the purpose of flagging those who might face difficulties with academic English and identifying the nature of their English development needs. Various types of validity evidence associated with the PAAL have been presented, using an argument-based framework for PELA validation previously explicated by the first two authors (Knoch and Elder 2013) and drawing on data from a series of trials.

The argument-based framework identifies the inferences that underlie validity claims about a test, and the warrants associated with each inference. The first is the *Evaluation* inference with its warrants of statistical robustness, appropriate test administration conditions and clarity (for test-takers) of tasks and instructions. These warrants were generally supported by the different sources of evidence collected, with an item analysis of each test component yielding excellent reliability statistics, the writing rater reliability being within expected limits, and feedback from test takers revealing that instructions were clear and tasks generally

manageable, apart from a few remediable technical issues associated with the online testing platform.

Warrants that were tested in relation to the second, *Generalizability*, inference were that different forms of the test were parallel in design and statistically comparable in level of difficulty, that there were sufficient tasks or items to provide stable estimates of ability and that test administration conditions were consistent. Results reported above indicate that different forms of the test were comparable, both in content and difficulty and that candidates were sorted into the same categories, regardless of the version they took. The online delivery of the PAAL moreover ensures consistency in the way test tasks are presented to candidates and in the time allowed for task performance.

As for the third *Explanation and Extrapolation* inference, which has to do with the test's claims to be tapping relevant language abilities, correlational evidence from the development trial showed a strong relationship between the PAAL scores for Parts A and B and the more time-intensive listening and reading items of the previously validated DELA, which had been administered concurrently to trial candidates. The warrant that the writing criteria capture relevant aspects of the academic writing construct is supported by research undertaken at the design stage. The predictive power of the writing component test is also supported by in-house data collected at the University of Melbourne showing the strong predictive power of DELA scores in relation to WAM scores. Other warrants associated with this inference have yet to be tested, however, and it is acknowledged that the length of the test and the indirect nature of the screening tasks in Parts A and B somewhat constrain its capacity to capture the academic language ability construct.

The *Decision* inference, the fourth in the argument-based PELA framework, encompasses warrants relating to the categorization of students based on test scores and the way test results are reported and received. Here the evidence presented gives a mixed picture. Standard-setting procedures ensured that the test's capacity to classify candidates into different levels was defensible. The meaning and purpose of the testing procedure was well understood by test users and score reports were made available to them in a timely manner. However, test-taker feedback revealed some dissatisfaction with the level of detail provided in the score reports and with the advice given about further support – perhaps because the avenues for such support were indeed quite limited. The fact that feedback was gathered from only a portion of the potential test taker population may also be a factor in these reactions as it tends to be the more motivated students who participate in trials. Such students are more likely to engage with the testing experience and expect rewards from it, including a full description of their performance and associated advice.

Evidence supporting the warrants relating to the fifth, *Consequences*, inference is even more patchy. Although the PAAL is designed to be administered to all incoming students, participation in the testing was by no means universal in the Faculties selected for the trial. In addition, there were mixed feelings about the usefulness of the initiative in informing future learning, due partly to the limited diagnostic information provided in the score reports but, more importantly, to the lack of available support options linked to these reports. Whether many students in the 'at

risk' category actually acted on their recommendation to enroll in support courses is unclear, but past history at the University suggests this is unlikely.

In general then, it can be seen that while the design of the PAAL and the information it provides about students' needs appears sound and indeed an improvement on previous PELA models in terms of its efficiency, there are issues associated with its utilization that require attention. Most of these issues are related to the policy environment in which the various trials were implemented rather than to the nature of the test itself.

For such a test to achieve its purpose of enhancing students' chances of academic success by identifying their particular English learning needs (or the lack of any such need), it has to be embedded in a more enlightened university policy which places a premium on the provision of opportunities for English language development, makes these opportunities accessible to students from any discipline by offering appropriate advice about avenues for action, and makes the consequences of inaction plain to test users, whether by mandating the test and enforcing its recommendations or through individual post-test counseling and follow-up tracking and monitoring of students. While the resources required to implement a fully-fledged language development policy alongside the test are considerable, the expenditure may well pay off in terms of student retention and outcomes which in turn would contribute to the reputation of the institution concerned. As well as suggesting such directions for policy reform, the findings of this study point to the necessity of construing the policy context as an integral dimension of validity, rather than merely as a set of external constraints.

## References

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Bonanno, H., & Jones, J. (2007). *The MASUS procedure: Measuring the academic skills of university students – A diagnostic assessment*. Sydney: Learning Centre, University of Sydney. Available at: [http://sydney.edu.au/stuserv/documents/learning\\_centre/MASUS.pdf](http://sydney.edu.au/stuserv/documents/learning_centre/MASUS.pdf)
- Dunworth, K. (2009). An investigation into post-entry English language assessment in Australian universities. *Journal of Academic Language & Learning*, 3(1), A1–A13.
- Elder, C., & Knoch, U. (2009). *Report on the development and trial of the Academic English Screening Test (AEST)*. Melbourne: University of Melbourne.
- Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly*, 5(3), 173–194.
- 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., Knoch, U., & Zhang, R. (2009). Diagnosing the support needs of second language writers: Does the time allowance matter? *TESOL Quarterly*, 43(2), 351–359.
- Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of the TOEFL iBT test, the International English Language Testing System (Academic) Test, and the Pearson Test of English for Graduate Admissions in the United States and Australia: A case study of two university contexts* (ETS research report series, Vol. 2). Princeton: Educational Testing Service.

- Group, E. L. D. A. (2012). *Performance in New Generation degrees and the University's policy on English language diagnostic assessment and support*. Melbourne: University of Melbourne.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport: American Council on Education/Praeger.
- Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing*, 2(1), 76–104.
- Knoch, U. (2010a). *Development report of version 2 of the Academic English Screening Test (AEST)*. Melbourne: University of Melbourne.
- Knoch, U. (2010b). *Development report of version 3 of the Academic English Screening Test*. Melbourne: University of Melbourne.
- Knoch, U. (2011). *Development report of version 4 of the Academic English Screening Test*. Melbourne: University of Melbourne.
- Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessments (PELAs). *Papers in Language Testing and Assessment*, 2(2), 1–19.
- Knoch, U., & Elder, C. (2016). Post-entry English language assessments at university: How diagnostic are they? In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim*. Newcastle-upon-Tyne: Cambridge Scholars Publishing.
- Knoch, U., & O'Hagan, S. (2014). *Report on the trial implementation of the post-entry assessment of academic language (PAAL)*. Unpublished paper. Melbourne: University of Melbourne.
- Knoch, U., Elder, C., & McNamara, T. (2012a). *Report on Feasibility Study of introducing the Academic English Screening Test (AEST) at the University of Melbourne*. Melbourne: University of Melbourne.
- Knoch, U., O'Hagan, S., & Kim, H. (2012b). *Preparing the Academic English Screening Test (AEST) for computer delivery*. Melbourne: University of Melbourne.
- Matsumura, N. (2009). *Towards identifying the construct of the cloze-elide test: A mixed-methods study*. Unpublished Masters thesis, University of Melbourne.
- Plake, B., & Hambleton, R. (2001). The analytic judgement method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah: Lawrence Erlbaum.
- Read, J. (2015). *Assessing English proficiency for university study*. London: Palgrave Macmillan.

Post-admission Language Assessment of University  
Students

Read, J. (Ed.)

2016, X, 243 p. 23 illus., 14 illus. in color., Hardcover

ISBN: 978-3-319-39190-8