Setting Empirical Standards on the Diagnostic English Language Needs Assessment (DELA)

Ute Knoch and Kellie Frost
Language Testing Research Centre
The University of Melbourne
March 2017

Contents

1. Executive summary

The outcomes of the project are a new set of agreed passing standards for both undergraduate and post-graduate students entering the University of Melbourne on the Diagnostic English Language Assessment (DELA), a post-entry assessment used at the University of Melbourne. It is anticipated that suitable passing standards will have a flow-on effect of improving the experience of international students as well as other students from non-English-speaking backgrounds at the University of Melbourne. Suitable standards will ensure that the English-language development needs of students during their time at university are more appropriately met.

2. Introduction and Background

The Diagnostic English Language Assessment (DELA) has been used at the University of Melbourne since the early 1990s to provide a mechanism to identify students who might be at risk due to relative weakness in their academic English skills. Following the assessment, students are grouped into one of three groups: (1) 'at risk' for students who are deemed to require follow-up language support, (2) 'borderline' for students who may require language support and (3) 'proficient' for students who are deemed able to cope with the academic language demands of the university. In language testing terms, the score points on a language assessment that are used to divide students into different decision-making levels are referred to as cut-scores. When DELA was introduced, only English language experts, not content lecturers, were involved in determining the cut-scores that divide students into these three groups, and these scores have not since been revisited.

With an increasingly diverse student population at the university, including higher numbers of international students, local heritage language students, and students from low socio-economic backgrounds, attention within the university is focused more and more on ensuring appropriate support mechanisms are in place for students facing English language-related challenges. Internal discussions along these lines prompted proposed changes to the DELA policy. These proposed changes include introducing a requirement for all undergraduate and postgraduate students to undertake an English language communication skills course if they are entering a degree with a coursework component of 150 points or more with IELTS scores of lower than 7 overall (or equivalent). These students would have the option to test out of this requirement by taking the DELA and reaching at least the minimum acceptable threshold level of performance. This proposed policy would not only increase the stakes of the DELA, but would most probably also result in a significant increase in candidate numbers.

As a result, it became a priority to review the DELA cut scores used to determine if incoming students are likely to need additional English language support, in order to evaluate their appropriateness in the current university context. Furthermore, as the DELA is used to predict language performance in academic content courses, we sought to include both the views of academics with expertise in English language teaching and learning and the views of academics with expertise across the various disciplines taught within the university. This approach was aimed at capturing a wider and more valid range of perspectives concerning the level of academic English language proficiency needed by incoming students to succeed in university studies.

Standard-setting

One of the most challenging matters confronting language test developers today is the setting of performance standards. Standard setting, or mapping test scores to descriptions of language skills expressed within a scale of levels or competencies is a way of attributing meaning to test scores and of translating test scores into descriptions of test takers abilities (Kane 2012).

Extensive research has been carried out in the disciplines of educational measurement and psychometrics on standard setting (see Angoff, 1971; Cizek, 2001, 2012; Cizek & Bunch, 2007; Jaeger, 1982; Thurstone, 1927; Zieky, 2001), and a variety of methods exist to set cut scores for test performances (see Hambleton & Pitoniak, 2006; Cizek & Bunch, 2007; Zieky et al., 2008; Cizek, 2001, 2012).

In the current study, the Bookmark method (Cizek & Bunch, 2007) was used to set standards on the receptive skill sections of the DELA, reading and listening, and the Performance Profile method (Cizek, 2012) was used for the writing section. A 'twin-panel' approach was adopted (see Brunfaut and Harding, 2014), to compare the judgments of university lecturers (discipline experts) with the judgments of ESL teaching and academic skill support staff (English language experts), as a means of incorporating different perspectives and of cross-validating panel decisions.

Through the standard setting process for each of the three DELA skills, reading, listening and writing, the intention was to determine the minimum acceptable English threshold level needed by students to avoid enrolling in a language support course. Currently, the DELA has two cut-scores, placing students into three groups: (1) proficient, not requiring support; (2) borderline, may need support; and (3) at risk, in need of support. These cut-scores currently sit at 4.0 and 3.3 respectively on the scale of a possible 7. These cut-scores were not set empirically and for the new policy, if adopted, only one cut-score may be required. This will determine whether a student is required to take a language support program or not. As the new policy makes the DELA more of a high-stakes test than it has been in the past, it is important to review and validate the cut-scores used on the test before the policy is implemented.

The main aim of this study was therefore to set new, empirically based standards on the DELA. This was achieved by conducting formal standard-setting workshops with academics and language support staff from across the university. Setting appropriate passing standards on language assessments is a topic that is often under-researched but is very important. As mentioned, a large number of methods have been proposed in the educational testing literature but no research to date has applied these methods in a principled way to a post-entry language assessment such as the DELA. Passing standards are often set ad hoc with little understanding of the implications.

3. Methods

Participants

A twin panel approach was adopted in the current study, with one panel comprised of English language and academic skills support academics (Panel 1), and the other comprised of lecturers and teaching academics from a variety of disciplines in the university, including Architecture, Arts, Business and Economics, Engineering, Health Sciences and Veterinary and Agricultural Sciences (Panel 2).

Panel 1 consisted of 9 participants for both the reading and listening test standard setting sessions and 10 participants for the writing session. Panel 2 consisted of 12 participants for the reading session, 9 for listening and 12 for writing.

Procedures

Reading and Listening

Two standard setting sessions were conducted with each panel, one for reading and one for listening. The Bookmark method (Cizek & Bunch, 2007) was used for standard setting on both sub-tests. In the bookmark method, standard setting participants are provided with an 'Ordered item booklet', which is a booklet comprised of a set of test items placed in order of difficulty, from easiest to hardest. The task for participants is to review each page of the booklet to decide if the test item on the page is likely to be answered correctly by a minimally qualified candidate at the ability level in question. When participants reach a page where the answer is no longer 'yes', the last item for which the answer was 'yes' becomes the cut score point. In other words, a minimally qualified candidate is expected to have a good chance of correctly answering all items before and at the bookmarked page, but is considered likely to find all items after the bookmarked page too difficult.

An overview of the ordered item booklets for each sub-skill will first be provided below, followed by a description of how the standard setting sessions were conducted.

Individual ordered item booklets were compiled for each of the sub-tests, reading and listening. DELA reading test version 4, "Dogs and Games" and DELA listening test version 7, "New Media", were selected for the purpose of conducting standard setting on the basis of reliability indices. For each sub-test, Rasch analysis was conducted using Winsteps version 3.81.0[1] on test response data from 100 randomly selected test takers from a previous live administration of DELA. Item difficulty indices were converted into theta ability values for a response probability of two-thirds[2]. 24 test items (out of a total of 46 for reading and 45 for listening) were selected for each sub-test to ensure a suitable spread

[1] Linacre, J. M. (2016). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

[2] Theta ability value = item difficulty index + 0.693 when a response probability of 2/3 is used.

of difficulty, and these items were compiled in order of ability values, from lowest to highest, into each of the ordered item booklets[3]. Each page of the booklets consisted of a test item, an answer key, and the section of reading or listening text, the latter of which was transcribed, to which the item related.

Both reading and listening standard setting sessions began with an introductory presentation, in which an overview of the DELA was provided, as well as an overview of the purpose of setting standards and of the procedures involved in the Bookmark method. Participants were then provided with and asked to complete a sample version of the relevant sub-test. For the sub-skill in question, participants were asked to imagine and discuss the characteristics of minimally qualified undergraduate and postgraduate students at two ability levels: a student with just enough language ability to be admitted to university but in need of compulsory language support (at risk), and a student with just enough language ability to not need a compulsory language program, but for whom support should be recommended (borderline).

Two decision making rounds were conducted in each session. In round one, participants were asked to complete their ordered item booklets individually. Their task was to review each page in the booklet and to stop at the pages where they judged:

- an 'at risk' undergraduate student would have less than a 2/3 chance of answering the item correctly
- an 'at risk' postgraduate student would have less than a 2/3 chance of answering the item correctly
- a 'borderline' undergraduate student would have less than a 2/3 chance of answering the item correctly
- a 'borderline' postgraduate student would have less than a 2/3 chance of answering the item correctly

Each participant was thus asked to make four decisions, two relating to undergraduate students and two relating to postgraduate students. They were instructed to place their bookmarks on the page immediately before each of the four pages they identified, so that their bookmarks corresponded to the last item for which the imagined students at 'at risk' and 'borderline' ability levels were thought to have a good chance of providing a correct response.

The four decision points for each participant were then compiled by the session coordinator and presented to the group. This was followed by a group discussion whereby participants provided reasons for their judgments and negotiated over disagreements. In round two of decision making, participants were invited to reflect on and if they wished, revise their initial decisions. They were asked to write down their final decisions and justifications for these decisions on the decision form provided.

Writing

[3] The order of items is the same whether ordered by difficulty indices or theta values when Rasch analysis is used.

The standard-setting method chosen for the writing sessions was the performance profile method (Cizek, 2012). Two separate meetings were convened, one for the lecturers and one for English language and academic skills academics. The procedures followed in each case were identical.

At the beginning of each session, participants were provided with information about the purpose of the standard-setting session and the background to the DELA test. They were then provided with a sample DELA writing task to review (the same task which was used as the basis for the writing samples reviewed later in the session).

Participants were then asked to discuss the typical writing ability of a minimally qualified undergraduate and postgraduate student at each of the following two ability levels: (1) a student with just enough language ability to be admitted to university but in need of compulsory language support (at risk), and (2) a student with just enough language ability to not need a compulsory language program, but for whom support should be recommended (borderline). Participants discussed these characteristics as a group.

Participants were then presented with a booklet of 14 DELA writing essays which were pre-ordered based on the average DELA writing score the scripts had received in the administration of the test. The weakest script appeared first in the booklet, and the highest scored script last.

Participants were asked to read each essay and then decide whether the writing was minimally acceptable without compulsory language support and record their decision in Table 1 below, first imagining that the writer was a newly arriving undergraduate (Column 2) and a new postgraduate coursework student (Column 3). Once these decisions were recorded, participants were asked to discuss where they placed the first 'no' and provide reasons. Participants were also asked whether they agreed with the ordering of the scripts. Following this discussion, participants were provided with the opportunity to change their judgements if they desired.

Table 1. Writing decision table

Script ID	Undergraduate Minimally Acceptable without compulsory language support - Yes or No?	Postgraduate coursework Minimally Acceptable without compulsory language support - Yes or No?
1		
2		
3		
4		
…		
…		
14		

Following this, participants were presented with another form on which they were asked to record which writers of the same 14 scripts would benefit from non-compulsory support to arrive at the

second set of cut-scores. The same procedures as described above were also used for this round of decisions.

Data analysis

Reading and Listening

Final decisions from round two were collated for each of the two panels and data were entered into separate excel spreadsheets for each panel and each sub-skill. Mean theta ability values were calculated for each of the four decision points (two for undergraduate students and two for postgraduate students) for each panel group separately, and for the combined group. These mean values were then converted into DELA band scores using existing raw score-to-band score conversion formulas for each sub-test.

Writing

The first step in the analysis was to work out the means of the scripts at each of the three levels: requiring compulsory support, optional support, and not requiring support. The means were calculated using the mean DELA writing scores for each script. Once the mean for each group had been established, the mean between two adjacent means was calculated to arrive at the cut-score between two levels. The same process was followed for both undergraduate and post-graduate cut-scores. We also calculated the means and cut-scores first separately for each participant group as well as combined.

4. Results

The results are presented in three sections. Firstly, findings from the reading test standard setting sessions are shown, followed by the listening session results. We conclude this section with the results for the writing standard-setting.

DELA reading

Tables 2 –4, below, show findings for panel 1, panel 2 and the combined group, respectively.

Table 2. Panel 1 – ESL and Academic Skills group

n=9	Undergraduate 1*	Undergraduate 2#	Postgraduate 1*	Postgraduate 2#
Mean theta (ability)	-0.20	1.13	0.78	1.76
SD	0.48	0.25	0.61	0.66
Theta value range	-0.55 to 1.02	0.81 to 1.51	-0.24 to 1.32	0.81 to 2.72
DELA band score	3	4	4	5

*'Undergraduate 1/Postgraduate 1' = 'at risk': in need of compulsory language support.

'Undergraduate 2/Postgraduate 2' = 'borderline': support is recommended but not compulsory.

Table 3. Panel 2 – Lecturers and teaching staff group

n=12	Undergraduate 1*	Undergraduate 2#	Postgraduate 1*	Postgraduate 2#
Mean theta (ability)	0.13	1.43	0.79	1.75
SD	0.74	0.78	0.90	1.0
Theta value range	0.55 to 2.12	0.02 to 3.37	-0.67 to 2.72	-0.24 to 3.37
DELA band score	3	4	4	5

*'Undergraduate 1/Postgraduate 1' = 'at risk': in need of compulsory language support.
'Undergraduate 2/Postgraduate 2' = 'borderline': support is recommended but not compulsory.

As shown in tables 2 and 3, above, mean theta values for both the ESL and academic skills group (panel 1) and the lecturers and teaching staff group (panel 2) indicated, as expected, that lower ability levels were associated with 'at risk' undergraduate and postgraduate students than 'borderline' students at each course level. For undergraduate students, mean theta values in panel 2 were higher at both decision points than those in panel 1, suggesting that English reading standards expected by lecturers and teaching staff participants were higher than those expected by English language and academic skills specialist staff. Results from both panels, however, yielded similar differences in mean theta values between 'at risk' and 'borderline' undergraduate students: a 1.33 logit difference between ability levels in the ESL/Academic Skills group and a 1.30 logit difference in the Lecturers/teaching staff group. For postgraduate students, mean theta values and the differences between 'at risk' and 'borderline' students were very similar across the two panels, suggesting a better alignment of expected standards between the two groups. Furthermore, mean theta values for both groups corresponded to the same DELA band scores for each of the four categories of students.

As shown in table 4, below, the mean theta values for the combined panel group also corresponded to the same DELA band scores derived from each individual panel. Reading cut scores were DELA band 3 for 'at risk' undergraduate students (those in need of compulsory English language support) and DELA band 4 for 'borderline' undergraduate students (English language support recommended but not compulsory). For postgraduate students, these cut scores were DELA band 4 and DELA band 5, respectively.

Table 4. Combined group

n=21	Undergraduate 1*	Undergraduate 2#	Postgraduate 1*	Postgraduate 2#
Mean theta (ability)	-0.01	1.14	0.79	1.75
SD	0.65	.60	0.77	0.85
Theta value range	-0.55 to 2.12	0.02 to 3.37	-0.637 to 2.72	-0.24 to 3.37
DELA band score	3	4	4	5

*'Undergraduate 1/Postgraduate 1' = 'at risk': in need of compulsory language support.
'Undergraduate 2/Postgraduate 2' = 'borderline': support is recommended but not compulsory.

DELA listening

Tables 5 – 7 below show findings for panel 1, panel 2 and the combined group, respectively.

Table 5. Panel 1 – ESL and Academic Skills group

n=9	Undergraduate 1*	Undergraduate 2#	Postgraduate 1*	Postgraduate 2#
Mean theta (ability)	-0.31	1.03	0.09	1.67
SD	0.85	0.57	0.91	0.72
Difficulty range	-1.61 to 1.36	0.44 to 2.09	-1.00 to 2.09	0.79 to 2.61
DELA band score	3	4	4	5

*'Undergraduate 1/Postgraduate 1' = in need of compulsory language support.

'Undergraduate 2/Postgraduate 2' = support is recommended but not compulsory.

Table 6. Panel 2 – Lecturers and teaching staff group

n=9	Undergraduate 1*	Undergraduate 2#	Postgraduate 1*	Postgraduate 2#
Mean theta (ability)	-0.21	1.03	0.62	2.23
SD	0.73	1.25	1.47	1.23
Difficulty range	-1.00 to 1.11	-0.31 to 2.73	-0.88 to 3.28	-0.31 to 3.28
DELA band score	3	4	4	5

*'Undergraduate 1/Postgraduate 1' = in need of compulsory language support.

'Undergraduate 2/Postgraduate 2' = support is recommended but not compulsory.

Standard setting outcomes for the listening sub-test were similar to those for the reading sub-test, described above. Tables 5 and 6 show that mean theta values for both the ESL and Academic Skills group (panel 1) and the lecturers and teaching staff group (panel 2) were higher for 'borderline' undergraduate and postgraduate students than for 'at risk' students. Again, mean theta values in the lecturers and teaching staff group were higher for 'at risk' undergraduate and postgraduate students than corresponding mean ability levels in the ESL and academic skills staff panel. As was the case for reading, this possibly indicates that expected minimum listening skill standards are higher in the former group. For 'borderline' undergraduate students, however, mean theta values were the same in both panels. By contrast, mean ability levels were higher in the lecturers and teaching staff panel than the ESL and academic skills group for both 'at risk' and 'borderline' postgraduate students. The range of ability levels was wider in the former panel, however, at three out of four decision points ('borderline' undergraduate students and 'at risk' and 'borderline postgraduate students). This suggests that although expectations are higher on average in the lecturers and teaching staff panel, judgments varied significantly.

As was the case in the reading sub-test, between panel differences in mean theta values did not correspond to differences in DELA band scores. As shown in table 7, below, the mean theta values for the combined panel group corresponded to the same DELA band scores derived from each individual panel, and yielded cut scores consistent with those for reading. Listening cut scores were DELA band

3 for 'at risk' undergraduate students (those in need of compulsory English language support) and DELA band 4 for 'borderline' undergraduate students (English language support recommended but not compulsory). For postgraduate students, these cut scores were DELA band 4 and DELA band 5, respectively.

Table 7. Combined group

n=18	Undergraduate 1*	Undergraduate 2#	Postgraduate 1*	Postgraduate 2#
Mean theta (ability)	-0.26	1.03	0.36	1.95
SD	0.77	0.94	1.22	1.02
Difficulty range	-1.61 to 1.36	-0.31 to 2.73	-1.00 to 3.28	-0.31 to 3.28
DELA band score	3	4	4	5

*'Undergraduate 1/Postgraduate 1' = in need of compulsory language support.
'Undergraduate 2/Postgraduate 2' = support is recommended but not compulsory.

DELA writing

Undergraduate

Table 8 below shows the means of the DELA writing scores for each of the three groups for undergraduate students. The same information for postgraduates can be found in Table Y. To arrive at this mean, each time a participant grouped a script into one of these groups, the DELA writing score was recorded against that group. Based on these judgements, the mean was calculated. It can be seen that the means increase as the groups advance and that lecturers chose the mean scripts to be slightly higher than the language support staff in the lower groups, but the opposite was the case for the proficient group.

Table 8: Mean DELA scores within each group – postgraduate

	Language support	**Lecturers**	**Combined**
Compulsory support	3.11	3.17	3.14
Borderline	3.78	3.95	3.88
Proficient	4.50	4.25	4.38

Table X: Mean DELA scores within each group – undergraduate

	Language support	**Lecturers**	**Combined**
Compulsory support	3.21	3.27	3.25
Borderline	3.97	4.06	4.02
Proficient	4.62	4.17	4.46

Based on the means presented in Table X above, the cut-scores between the three groups were calculated by simply calculating the means between the group means. Based on this analysis, we arrived at the following cut-scores for undergraduates:

- Language support cut-scores: compulsory – 3.45 – borderline – 4.14 – proficient
- Lecturer cut-scores: compulsory – 3.56 – borderline – 4.10 – proficient
- Combined: compulsory – 3.51 – borderline – 4.13 - proficient

The cut-scores for postgraduates can be found below:

- Language support cut-scores: compulsory – 3.59 – borderline – 4.30 – proficient
- Lecturer cut-scores: compulsory – 3.67 – borderline – 4.10 – proficient
- Combined: compulsory – 3.64 – borderline – 4.24 - proficient

5. Discussion and Recommendations

In the reading and listening standard setting sessions, the cut-scores arrived at by each of the two panel groups were consistent with existing DELA cut-scores for undergraduate students. That is, students scoring below DELA band 3 in these skills were deemed in need of compulsory support. For postgraduate students, however, the English level expected was higher in both panels, who each set a cut score (DELA band 4) that is higher than the score currently used to categorise students as in need of support.

According to existing policy, a student achieving DELA band 4 is deemed proficient and not in need of support. Overall judgments from both panels also suggest that a score at or above DELA band 5, rather than DELA band 4, would be a more appropriate indicator of proficient and not in need of support in the case of incoming postgraduate students.

The combined cut-scores arrived from the writing standard-setting panels show that, when compared to the existing DELA writing cut-scores, the combined panels set the cut-scores higher, meaning that more students would be identified as needing support. This was even more pronounced for post-graduate students, where the cut-scores from the combined were even higher than those for undergraduate students, reflecting higher language expectations.

The two panels were fairly similar in terms of where they placed the writing cut-scores between the three groups for the undergraduates, although the lecturers were slightly more severe, that is including more students into the compulsory support group. For postgraduate students, the cut-score between compulsory and the borderline groups was almost identical between the two groups of participants, while for the cut-score between the borderline and proficient group the language support staff was more severe, that is including more students into the borderline group.

Given that both groups are important stakeholders in the university context, with different but equally valid perspectives on the English-related skills needed by students to succeed in their studies, we are of the view that adopting the combined cut-score, rather than prioritizing one panel's perspective over the other's, is the most appropriate course of action.

Based on this study, we make the following recommendations:

1. that students are grouped using separate cut-scores for undergraduate and postgraduate students
2. that the cut-scores based on the combined panels are adopted
3. that the new cut-scores are monitored to see how they serve the different key stakeholder groups, including students, lecturers, language development staff
4. that the new cut-scores are monitored to explore what impact they have on resources

6. References

Angoff, W. H. (1971). *Educational measurement*. R. L. Thorndike (Ed.). Washington, DC: American Council on Education.

Brunfaut, T. & Harding, L. (2014). Linking the GEPT listening test to the Common European Framework of References. Taipei: The Language Training and Testing Center.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. *Setting performance standards: Concepts, methods, and perspectives*, 3-17. New Jersey: Lawrence Erlbaum Associates.

Cizek, G. J. (2012). *Setting performance standards: foundations, methods and innovations, 2nd edition*. New York: Routledge.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. *Educational measurement*, 4, 433-470.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis*, 461-475.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. *Setting performance standards: Concepts, methods, and perspectives*, 19-51.