



CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 59

February 2015

ISSN 1756-509X



CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Research Notes

Issue 59 / February 2015

A quarterly publication reporting on learning, teaching and assessment

Senior Editor and Editor

Dr Hanan Khalifa, *Head of Research and International Development*, Cambridge English Language Assessment

Dr Ivana Vidaković, *Senior Research and Validation Manager*, Cambridge English Language Assessment

Editorial Board

Dr Ardeshir Geranpayeh, *Head of Psychometrics and Data Services*, Cambridge English Language Assessment

Dr Fred Davidson, *Professor Emeritus of Linguistics*, University of Illinois at Urbana-Champaign

Dr Andrew Q Somers, *Principal Research and Validation Manager*, Cambridge English Language Assessment

Production Team

Karolina Fraczak, *Marketing Project Co-ordinator*, Cambridge English Language Assessment

John Savage, *Publications Assistant*, Cambridge English Language Assessment

Printed in the United Kingdom by Canon Business Services

Research Notes

Issue 59

February 2015

Contents

Editorial	2
Introduction Ardeshir Geranpayeh	3
Item analysis Michael Corrigan and Paul Crump	4
Test construction: The Cambridge English approach Coreen Docherty and David Corkill	10
Grading and test equating Mark Elliott and Lynne Stevenson	14
Determining standards in assessing language ability Gad S Lim	20
Reporting test scores and the Cambridge English Scale Andrew Q Somers	23
A modern perspective on statistical malpractice detection Chris Bell	31
Computer-adaptive testing Agnieszka Walczak	35
The role of Quality Management in ensuring accurate and meaningful test scores Nick Beresford-Knox	40
Glossary	45
The Cambridge English Scale	47
Studies in Language Testing	47



Editorial

Welcome to issue 59 of *Research Notes*, our quarterly publication reporting on matters relating to learning, teaching and assessment within Cambridge English Language Assessment. This issue sheds light on a life cycle of Cambridge English language tests, focusing on analyses and processes which ensure a delivery of accurate and meaningful results.

Corrigan and Crump discuss how item analysis is used to inform and enhance test construction. They explain the process of item analysis (pretest or post-live exam) that ensures the quality of exam materials. This kind of analysis includes determining item facility, discrimination and differential item functioning. The authors provide practical examples of real items from the live exams to demonstrate how statistical information should be interpreted during test construction and administration.

Docherty and Corkill discuss the principles of test construction in Cambridge English Language Assessment, which provide a platform for enabling the test equating procedures, and for ensuring that tests are set at the right level. They explain the various stages of test construction from defining test specifications and the construct of language proficiency to ensuring comparability of tests across versions using an item banking approach.

The grading process pulls together a number of activities to ensure consistent application of standards across sessions and test versions. Elliott and Stevenson discuss the test equating procedures which are used to achieve a consistent approach to grading exams to the same standard and arrive at component and overall test scores for reporting. They also highlight how the training and monitoring of examiners helps to remove the impact of any variation in marker performance that might otherwise be introduced in the examinations. They conclude that there is no 'best' approach for grading for all contexts but the best practice is to ensure that the processes for test scoring should be controlled, fit for purpose and adhere to a sound statistical approach.

Lim examines how standard setting and comparability/alignment studies have informed and shaped the Cambridge English Scale. He argues that exams require standards to adhere to but there are various ways by which standards in tests can be determined, concluding that one has to adopt multiple methods for standard setting and validate them through repeated studies. Lim makes references to the embodiment of the Common European Framework of

Reference for Languages (CEFR) by Cambridge English and reports on empirical evidence to support it.

Somers' article on reporting and interpreting test scores considers how test scores should be reported to candidates to ensure that accurate and meaningful results are communicated to all stakeholders. In doing so, he discusses the key considerations in reporting scores such as the range of stakeholders and their requirements, types of scores, what they mean and the degree to which one can rely on their reliability. He then discusses the introduction of the new Cambridge English Scale and the evolution that had led to it. It is important to note that the new Cambridge English Scale is a scale for reporting results of different examinations on the same metric so that the relationship between scores in different examinations can be demonstrated in a meaningful and transparent way. It would also allow reasonable score comparisons between performance of candidates in one exam against others.

Malpractice, or cheating, is said to be the inevitable consequence and a by-product of high-stakes testing. Bell describes the modern context for this increasing phenomenon in high-stakes testing and explains various detection techniques employed by Cambridge English to address it.

Walczak provides a brief review of how computer-adaptive testing (CAT) has been implemented in Cambridge English since the late 1990s. Most processes discussed in the previous articles are applicable to all forms of assessment, but the process of test construction is rather more elaborate in the world of CAT. The automated test assembly is progressively based on a candidate's performance during the test, which makes CAT a popular test for contexts where speed in reporting results to candidates is paramount.

Finally, all processes underlying a life cycle of a test need to be maintained and continually evaluated. Beresford-Knox echoes Milanovic and Saville's (1992) *Principles of Good Practice for UCLES Examinations* by outlining the various aspects of the Quality Management system that remain in place in Cambridge English. He explains how Quality Management contributes to producing accurate and meaningful test scores for our stakeholders.

We finish this issue with details of the Studies in Language Testing volumes that link to the theme of the issue, and a diagram representing the Cambridge English Scale which was launched earlier this year.

Introduction

ARDESHIR GERANPAYEH RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

This issue of *Research Notes* is intended to demystify various statistical analyses that are applied to Cambridge English exams throughout the whole test development process. This process can largely be broken down into three main areas: test construction, grading and reporting results.

Since the early 1990s, statistical analyses have been used in Cambridge English for constructing test items and interpreting results, but their role in the examination process was relatively limited until the late 1990s. The role of measurement came under lens in 1987, when the University of Cambridge Local Examinations Syndicate (UCLES, now Cambridge English Language Assessment) commissioned a 3-year project, the Cambridge-TOEFL Comparability Study (Bachman, Davidson, Ryan and Choi 1995). The goals of the project were to examine various aspects of test content and administration of the then biggest Cambridge English test, i.e. *First Certificate in English (FCE)*. The findings of the study highlighted areas in the testing process which could benefit from more measurement rigour.

Around the same time, in 1989, Cambridge English set up an Evaluation department to review test construction, grading, administration and implementation of any necessary changes to the examinations. One of the first reports of the Evaluation unit was the publication of a document entitled *Principles of Good Practice for UCLES Examinations* by Milanovic and Saville (1992). Drawing on Bachman (1990), Milanovic and Saville argued that every test needs to possess four main qualities: Validity, Reliability, Impact and Practicality (VRIP). Even though these concepts were not new in Cambridge English, the report was seminal because it formalised VRIP in 1992 and introduced it as a framework for test development and validation in Cambridge English. One of the significant aspects of introducing VRIP to Cambridge English examinations was to emphasise the importance of providing evidence to support the claims made by the organisation. Hawkey and Milanovic (2013:126) argue that 'regardless of the context in which examinations are used, it is essential that users have access to the information they need to verify that the examinations are fit for purpose'. This, of course, required research evidence both on the construct/content of a test and test reliability if it were to be used in a high-stakes decision-making context. It goes without saying that achieving VRIP in examinations could not have been possible without a thorough revision of the test development cycle. Therefore, the 1990s was a time for applying VRIP in every aspect of test development, which resulted in the revision of many Cambridge English exams, such as *FCE* and *IELTS*.

Two important measurement developments were also introduced in the 1990s: item banking and a common Measurement Scale. The need for a tool to automate the Item Banking procedure was paramount. At the same time, there was a need for a measurement scale to link the exams to one another and to align them across a language proficiency scale. This required a robust pre-testing system

to allow item calibration using a measurement test theory. The Evaluation department, whose name had changed to Validation and Test development department, was tasked to develop a Local Item Banking System (LIBS), as well as a measurement scale; see Beeston 2000. Several measurement projects were initiated to develop a language ability scale underlying LIBS. At this point in its history, Cambridge English had become one of the first UK examination boards to adopt item response theory (IRT) as part of its scientific approach to calibrate language tests and to create a measurement scale for assessing language ability across all its exams; see Jones (2014) for the historic development of the Cambridge English underlying measurement scale. This underlying measurement scale underpinned all Cambridge English test development activities and was later used for the development of the ALTE 5-level reporting scale. The development of the latter reporting scale had a significant contribution to the emergence of the Common European Framework of Reference (CEFR) (Council of Europe 2001) 6-level scale in 2001. Soon after 2001, Cambridge English decided to abandon the ALTE 5-level scale in favour of the CEFR 6-level scale for reporting results to its candidates. This latter development had no impact on the underlying measurement scale that Cambridge was using for constructing its tests.

The Validation and Test Development department was also tasked to introduce quality assurance not only to test development, but also to training and monitoring examiners' using measurement techniques. A Performance Testing unit was established to manage examiners training and monitoring. By 2000, the Validation and Test Development department had grown significantly covering a Performance testing unit, Centre Inspections, Analysis and Interpretation of Results, Commissioning, Pretesting, Item Banking and Question Paper Preparation, Test Development and Institutional Testing, in addition to hosting the ALTE coordinator. With expertise in IRT and other psychometric techniques, Cambridge English were able to develop their first computer-adaptive test (CAT) in language proficiency back in 1999, which was the winner of the 2000 European Academic Software Award.

Hawkey and Milanovic (2013) refer to the time between 1987 and 1999 as the era of 'revolution and evolution' when measurement techniques became an integrated part of what Cambridge English did. Between 1995 and 1999, *IELTS* was revised several times. The use of measurement techniques and item banking led to a much more coherent way of aligning exams and improving their qualities after the millennium, resulting in several revisions of exams: *Certificate of Proficiency in English (CPE)* in 2002, 2008 and 2013-14, and *FCE* and *Certificate in Advanced English (CAE)* in 2004 and 2015. There is also a host of new revisions forthcoming for *Cambridge English: Business Certificates (BEC)*, *Cambridge English: Preliminary (PET)*, *Cambridge English: Key (KET)* and *Cambridge English: Young Learners (YLE)*.

Measurement theory and practice are now so embedded in what Cambridge English does that it is hard to think of a single process within which is not impacted by them. The articles in this issue explain to a general audience how measurement issues are addressed in Cambridge English.

References

- Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.
- Bachman, L, Davidson, F, Ryan K and Choi, I (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.
- Beeston, S (2000) The EFL Local Item Banking System, *Research Notes* 1, 5–7.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Hawkey, R and Milanovic, M (2013) *Cambridge English Exams: The First Hundred Years: A History of English Language Assessment from the University of Cambridge 1913–2013*, Studies in Language Testing volume 38, Cambridge: UCLES/Cambridge University Press.
- Jones, N (2014) *Multilingual Frameworks: The Construction and Use of Multilingual Proficiency Frameworks*, Studies in Language Testing volume 40, Cambridge: UCLES/Cambridge University Press.
- Milanovic, M and Saville, N (1992) *Principles of Good Practice for UCLES Examinations*, Cambridge: Cambridge ESOL internal report.

Item analysis

MICHAEL CORRIGAN RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

PAUL CRUMP ADMISSIONS TESTING SERVICE, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

This chapter will discuss the statistical analysis of objectively marked items. Items comprise the questions to which candidates respond; the instructions, or rubric, which accompany items; and the text (written or audio) which is used to derive the response.

For an objectively marked item, the marker is not required to use any judgement, as the correct answer has been identified in advance by the item writer. Such item types typically require selection of a correct response (e.g. multiple choice, matching tasks) or a limited amount of writing (one word or a few words, e.g. cloze, sentence completion). This way of recording candidate test performance contrasts with the writing of essays, where a high degree of expert judgement is required. When responses from a sufficient number of candidates are available, analysis can be very informative, as it can show how well an item works and help to diagnose problems with the item, hence the interest in *item analysis*.

Item analysis plays a role in the larger process of quality assurance for language tests (see Beresford-Knox, this issue). The aim of this process is to ensure that the measurement of candidate ability is as appropriate and accurate as possible. In order for this to be the case, high-quality items are required. Item analysis, together with its interpretation, is a cornerstone in appraising item quality and making decisions concerning the use of items in future live tests (see also Docherty and Corkill, this issue). The collection of appropriate data for item analysis to take place, therefore, must also be considered.

This article will describe the types of statistical analysis conducted and provide a worked example of item appraisal. Following this, data requirements will be discussed and some examples of practical considerations of the way

in which item analysis is implemented within a larger process of test production and administration will be given.

Item analysis

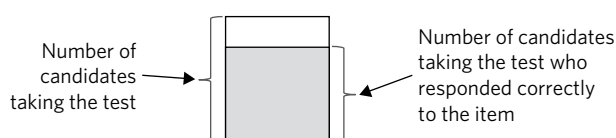
Varieties of item analysis and statistics

More than one approach to item analysis exists but they can, for many practical uses, be considered complementary. Two common approaches are classical test theory (CTT), which is based on parametric descriptive statistics (Traub 1997) and a more modern form of analysis, item response theory (IRT), which is derived from an understanding of the probability of success of individual candidates on individual items (Bock 1997). IRT is described in Elliott and Stevenson (this issue) and will not be discussed in depth here. It is more useful than CTT for comparing items across test forms, as the frame of reference is not limited to the particular group of candidates taking the test. IRT allows supportable judgements to be made about whether an item is within a suitable difficulty range for a particular test. This range is established in relation to the range of abilities of candidates who take the test. These two ranges should broadly match, as this provides maximum information about candidate ability.

Despite the advantages of IRT over CTT, the classical approach is still of great use, especially for diagnosing the performance of items. The results it produces are conceptually straightforward to interpret and, for item types such as multiple choice, CTT option-level statistics greatly facilitate understanding of item functioning. Two of the most important CTT statistics for appraising items will be described in this section: the *facility value* and the *discrimination index*. The facility value, illustrated in Figure 1

for a dichotomous item (scored wrong or right), is an index which represents how easy the item is for the candidates taking the test. It is calculated by dividing the number of candidates selecting the correct response for an item by the number of candidates taking the test. The result is a figure between 0 and 1, where high numbers indicate easy items, and low numbers indicate difficult items. As pointed out above, the interpretation of facility, and other CTT statistics, has a frame of reference limited to the particular group of candidates taking the test. For example, if two groups take the same test, the facility of the test, i.e. the proportion of candidates responding correctly to an item, may be quite different if the ability of each group varies significantly. The two different facility values which result are equally correct but only in relation to the group they were calculated for. This means that, although suitable for diagnosing problematic items, CTT is less suited to inform important decisions regarding test construction, where IRT is preferred. The facility value is useful, however, to determine how well the item is matched to a group of candidates. If that group is typical of the test-taking population, and the item is within an acceptable facility range, the item is more likely to be suitable for inclusion in a test.

Figure 1: The facility value for a dichotomous item



Item discrimination is the extent to which an item distinguishes between strong and weak candidates in the score they obtain. This is an important quality for items, as it is also the aim of most tests. Candidates responding correctly are grouped according to their score on the whole test. The proportion in the low-scoring group is then subtracted from the proportion in the high-scoring group. The result is a value between -1 and 1 , where -1 would indicate that all of the low-scoring group and none of the high-scoring group had selected the correct answer. This index is sometimes termed the *discrimination index*. Table 1 shows a range of values ('Example index') and what they imply for items ('Diagnosis' and 'Interpretation'). Since a figure of 0 indicates no discrimination between the low- and high-scoring groups, good items must have a higher number than this. Exactly how much higher the value must be before the item is considered acceptable is the subject of some debate. A figure around 0.3 is generally considered acceptable, however, relative comparisons can also be made. If two items have the same facility, the one with a higher discrimination index is likely to be selected. Another commonly used measure of discrimination is called the *point biserial correlation coefficient*. It is a correlation between the score each candidate receives for an item, and their test score. Unlike the discrimination index, data from all candidates is used in its calculation. As with the discrimination index previously described, figures range from -1 to $+1$. Figures for the discrimination index and the point biserial correlation coefficient are generally close to each other and are interpreted in the same way when

diagnosing an item. However, it is not advised that they are used interchangeably when working with a particular test, as some differences are expected.

Table 1: Example diagnosis and interpretation of the discrimination index

Example index	Diagnosis		Interpretation
	Selecting the correct response	Selecting an incorrect response	
-0.3	More low-scoring than high-scoring	More high-scoring than low-scoring	A poor item
0	Same for both groups	Same for both groups	A poor item
0.3	More high-scoring than low-scoring	More low-scoring than high-scoring	A good item

When considering how well an item functions, the answer options for multiple-choice or matching items, or the categories of correct and incorrect for other items are interesting to consider in more detail. If a particular option is found to be unattractive by candidates, this may be because it is clearly wrong. Option statistics, or *distractor analysis*, can therefore flag problems with the text of the item. The indices for answer options are based on the same principle as those described above for the item. In fact, the facility value, discrimination index and point biserial correlation coefficient for the item are identical to those for the key. For the other options, 'number of candidates who responded correctly to the item' need only be substituted with 'number of candidates selecting the option'. An example is provided below to show how the statistics discussed here can be used to appraise and edit items.

An example: Appraising an item

The following example is intended to illustrate the quality assurance processes performed in Cambridge English Language Assessment during test production to ensure that test items are fit for purpose. The description refers to stages in the process which correspond to those outlined in a generic description in the Appendix.

The example is taken from the Reading and Writing paper of the *Cambridge English: Key* exam, which measures from A1 to B1 on the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001). The task under analysis is a reading comprehension task, designed to test candidates' understanding of detailed information.

Table 2 shows the statistics for the task at the pretest review meeting. Values which are outside of typically accepted tolerance ranges at this level have been circled.

Table 2: Statistics for multi-item Cambridge English: Key task at pretest review

Question	Difficulty (IRT)	Facility (CTT)	Discrimination (CTT)
1	61.43	0.26	0.25
2	56.15	0.36	0.27
3	45.59	0.61	0.31
4	43.96	0.64	0.32
5	67.80	0.16	0.01
6	40.22	0.72	0.53
7	44.68	0.63	0.32

The extract below (Excerpt 1) shows the first item from Table 2. Although the discrimination is adequate, the item is above the difficulty level for this exam.

Excerpt 1: The first iteration of Item 1 and accompanying text

We Meet Twice

My name is Anna King and I was born in a town called Madison in Wyoming in the centre of the USA.

In my 20s, I moved to work on the east coast, in a town just south of New York. I got a job in a department store there. One day when I was working in the store, a young man with short brown hair looked at me and asked, 'Are you Michelle Golden?'

1 Anna found work in a department store in New York.

A Right B Wrong C Doesn't say



The statistics for answer options in Table 3 below reveal that 73% of all candidates have opted for A, and have clearly not understood the referencing to 'a town *just south of* New York', confusing it with '*in* New York'. The discrimination is adequate, because significantly more better candidates are choosing the key than those performing less well on the test as a whole; the 'Endorsing high' and 'Endorsing low' columns show the percentage of candidates in the low and high groups choosing each option, when the total candidature is divided into three groups, based on their performance on the whole test. However, it is important to note that 16% more of the better candidates are selecting the distractor A (57%) than are choosing the key (41%).

Table 3: Option-level item statistics for the first iteration of Item 1

Seq. No.	Scale-Item	Item Statistics				Alternative Statistics				
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing		Point Biser.	Key
							Low	High		
23	1-23	.26	.28	.25	A	.73	.84	.57	-.23	
					B	.26	.13	.41	.25	*
					C	.01	.03	.02	-.06	
					Other	.00	.00	.00		

At the review meeting, a decision was made to simplify the item by removing the testing reference to New York and targeting something else (Excerpt 2).

Excerpt 2: The second iteration of Item 1 and accompanying text

We Meet Twice

My name is Anna King and I was born in a town called Madison in Wyoming in the centre of the USA. When I was twenty, I moved to the east coast, to a town just south of New York, to start a job in a department store. One day, a young man with short brown hair who was shopping in the store looked at me and asked, 'Are you Michelle Golden?'

1 Anna left home and began working in a department store

A Right B Wrong C Doesn't say



This resulted in much better statistics when the task was re-pretested (see Table 4 below).

Table 4: Item statistics for the second iteration of Item 1

Question	Difficulty	Facility	Discrimination
1	30.58	0.81	0.37

The difficulty has now reduced significantly, and is right at the bottom of the acceptable difficulty range for *Cambridge English: Key* (possibly because of the repetition or 'wordspot' of 'department store' in text and item). The discrimination has also improved, with a greater division between the number of stronger and weaker candidates selecting the key (see Table 5 below).

Table 5: Option-level statistics for the second iteration of Item 1 and accompanying text

Seq. No.	Scale-Item	Item Statistics				Alternative Statistics				
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing		Point Biser.	Key
							Low	High		
18	1-18	.80	.33	.37	A	.80	.59	.92	.37	*
					B	.12	.23	.06	-.27	
					C	.08	.18	.02	-.23	
					Other	.00	.00	.00		

The second item in the task which failed to perform well in its first pretest outing is shown below. It was far too difficult at 67.80 in scaled IRT logits (see Elliott and Stevenson (this issue) for an explanation of IRT values and their interpretation) and discriminated very poorly (0.01). Excerpt 3 below shows the first iteration of this item, and the accompanying text.

Excerpt 3: First iteration of Item 5

Six months later, I got a better job and moved to the west coast to work at the company's San Francisco store. One day, on my home from work a young man with short brown hair passed me in the street and asked, 'Are you Michelle Golden?'

5 Anna got a job with a different company in San Francisco.

A Right B Wrong C Doesn't say

The option-level statistics are again informative (see Table 6), and indicate that many candidates failed to grasp the difference between '*the* company' (meaning the same one) and '*a* company'.

Table 6: Option-level statistics for first iteration of Item 5

Seq. No.	Scale-Item	Item Statistics				Alternative Statistics				
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing		Point Biser.	Key
							Low	High		
27	1-27	.16	.02	.01	A	.67	.68	.57	-.10	
CHECK THE KEY					B	.16	.16	.18	.01	*
B was specified, C works better					C	.17	.16	.25	.12	?
					Other	.00	.00	.00		

Although the key is B, A is proving very popular for both the low and the high group. Notice that the 'Check the Key' warning (generated for the attention of Assessment Managers) says that 'C works better', not A. This is because it is a warning about the discrimination, rather than the facility. C is the only option which shows a marked difference between the low and the high group. The fix for this, agreed at pretest review, was to change the text and the item, so that Anna's new job was now with a *different* company. Excerpt 4 below shows the revised version.

Excerpt 4: Second iteration of Item 5

Six months later, I got a better job with another department store and moved to the west coast to work at their San Francisco store. One day on my way home from work a young man with short brown hair passed me in the street and asked, 'Are you Michelle Golden?'

- 5 Anna got a job with the same company in San Francisco.
 A Right B Wrong C Doesn't say

In its second incarnation at pretest, the statistics for the item were as follows:

Table 7: Statistics for the second iteration of Item 5

Question	Difficulty	Facility	Discrimination
1	49.32	0.42	0.25

And the option-level statistics were these:

Table 8: Option-level statistics for the second iteration of Item 5

Item Statistics					Alternative Statistics					
Seq. No.	Scale-Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing		Point Biser.	Key
							Low	High		
22	1-22	.42	.24	.25	A	.39	.49	.24	-.25	
					B	.42	.31	.56	.25	*
					C	.19	.20	.21	-.01	
					Other	.00	.00	.00		

The difficulty level has come down considerably, but it is still at the top end of what is acceptable for *Cambridge English: Key*. A lot of the low group are going for A – perhaps because it's the linking across the sentence of 'another department store' and 'their' that makes it clear that it is another company and this is just too sophisticated for those at the bottom of the ability range.

Finally, Table 9 below shows how these two items performed in the live administration of the test:

Table 9: Post-live statistics for Items 1 and 5

Question	Difficulty	Facility	Discrimination
1	33.64	0.86	0.27
5	47.62	0.79	0.50

It is interesting to note that the live values are very similar to those from the second round of pretesting, with only slight shifts in the difficulty and discrimination values. Both items are now within target ranges for the test.

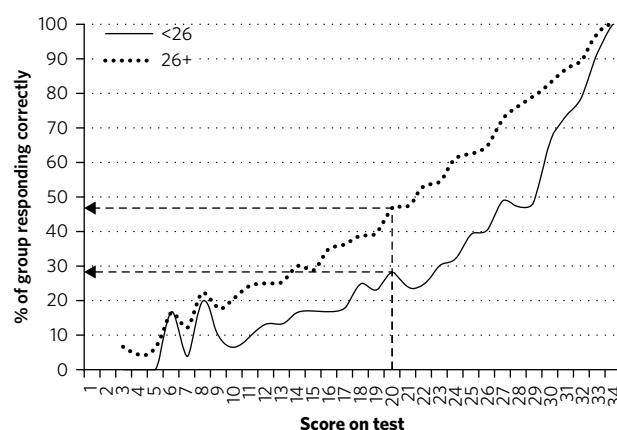
Advanced item analysis

The description of item analyses in this article is by no means exhaustive but it does attempt to outline the most important types of analysis carried out in Cambridge English Language Assessment. One final area of analysis which will be discussed here is that of item bias. In the context of item analysis, this may be defined as the propensity of an item to favour one identifiable subgroup of candidates (e.g. females vs. males, older vs. younger candidates) over another *not in an area which the test is intended to measure*. The last part of the preceding sentence is italicised because this is sometimes overlooked. English language tests are expected to distinguish between candidates based on their ability in English. If candidates taking a test are broken into subgroups based on their first language, differences between subgroups would be expected due to first language interference. This would not be considered bias however, as the challenge of learning English will vary for candidates with different first languages. Difference between subgroups based on gender, however, is not a legitimate aspect of English language ability, and items which discriminate between males and females would raise serious questions about fairness. For this reason, the analysis of items for potential bias is important.

A type of analysis called *Differential Item Functioning* (DIF) is often a starting point for identifying potential bias. It involves defining subgroups of candidates based on their background characteristics (e.g. age, gender) and determining whether there is a significant difference in performance on individual items whilst controlling for ability level. If DIF is flagged for any items, they must be investigated qualitatively to determine the likely cause of the performance differences.

An example of DIF is given in Figure 2, which shows the performance of two groups (candidates aged under 26 and candidates aged 26 and over) on an item. The selection of the groups being compared should be made with an understanding of the test population, and focus on aspects where bias is most likely to occur. In this example, it was considered that older candidates might have world knowledge which could help them to select the correct response. The percentage of candidates responding correctly to the item for each group (y axis) is plotted against test score (x axis). The gap between the two lines at most score points (an example for those scoring 20 is inscribed with a dashed line) shows that candidates at the same ability level perform differently

Figure 2: Chart showing DIF for an item when comparing those aged 26 and over (broken trace) with those aged below 26 (solid trace)



depending on group membership. In general, those in the 26 and over group score higher on this item. A plot of an item which did not exhibit DIF would show the two lines very close together, with each group performing approximately the same.

Data collection

Item analysis and the test construction and administration process

To ensure that high-quality items are used to measure candidate performance in live tests, they must be appraised beforehand. To this end, they can be pretested (included in a special test set up for the purpose of obtaining performance information on test items). The data obtained from the pretest should not be unduly influenced by spurious effects such as candidate fatigue, lack of motivation or lack of preparedness, as these will lead to misleading findings of the analysis. Furthermore, candidates should be neither too able, nor too weak for the items, as very limited item information will be yielded. If all candidates respond correctly to all items, for example, the only thing which can be concluded is that the candidates were more able than the items were difficult.

An additional consideration when administering items is to ensure that there is a data link between the items in the test and other items with which they will later be compared. This allows comparison using IRT measures, described above, and is important for later test construction (see Docherty and Corkill, this issue), as well as initial item appraisal.

Data may also be gathered from live tests, either to improve on pretest statistics, or, in some cases, because it is possible to include and pretest items in live tests without using them to evaluate candidate ability. This is made possible by IRT, which provides ability estimates, rather than summed scores (see Elliott and Stevenson, this issue). Whether data is gathered from a specially arranged pretest, or a live test, requirements for suitable data do not change.

Collecting suitable data for item analysis

In order to collect suitable data from a pretest, sampling requirements may be set. These can relate to considerations such as the preparation status of candidates and their first language. The former concerns the familiarity of candidates with the exam and their motivation to complete the pretest. Both are important, as lack of familiarity or motivation can lead to significantly poorer performance. In order to counter such effects, therefore, pretest candidates are recruited from amongst candidates who are preparing to take the live test in the near future and would welcome the chance to complete a mock test as part of their preparation. Candidates are also provided with rudimentary diagnostic feedback to help them with this. Involving candidates preparing for a specific test will also help to ensure the ability level is appropriate.

The frequency of candidate first language in the data is an important consideration. The differences and similarities between a particular language and English can mean that some items would be significantly easier or more difficult for some candidates. In addition to effects directly attributable to first language, educational or cultural practices associated with particular languages can also impact responses. If any particular

first languages or language families were allowed to dominate, the calculation of item difficulty would be unduly influenced.

One solution to the problem of constructing a suitable sample is, in pretests, to attempt to include as many first languages and language families as possible, up to the target number of candidates (usually between 150 and 250). In addition, there is a limit on the number of candidates who speak a first language from one particular family. If too many candidates are found to speak a single first language, or languages belonging to a single language family, the responses of some candidates are removed from the data to equalise the proportions of candidates in each group. If necessary, the number of candidates who have taken the pretest can be increased if too few remain at this point.

Live tests differ from pretests in two important ways, however. First, candidates cannot be selected to sit live tests as they are for pretests. For this reason, it is not possible to pre-specify the balance of L1s amongst the candidates. Second, the number of candidates taking live tests is normally much higher than for pretests, so, for the purposes of grading only, responses from some candidates can be removed from the data if necessary, without affecting the quality of the analysis. Balancing L1s within the data is therefore done after test administration, rather than before, as with pretests.

Linking data

As discussed in detail by Docherty and Corkill (this issue) test construction may be done using an item bank, which contains a large number of items of known quality and characteristics. As it is a collection of items from many administrations, a way must be found to compare key aspects of all banked items. Among these characteristics, item difficulty is particularly important and is also useful to know before banking an item, as it should be within a suitable difficulty range for a particular test before it is accepted into the bank. IRT can be used to measure and compare difficulty but the additional analytic step of *calibration* is required before difficulty is directly comparable between items. This is described more fully in Elliott and Stevenson (this issue). However, in order to calibrate items, a data link between calibrated and new items is required, and this will be discussed in this article.

A data link between calibrated and new items is provided by administering some items which are already calibrated with the new items. The items which are already calibrated are termed *anchor items* because they link the new items to the rest of the items in the bank. The way in which anchor items are combined with other items can vary and, in practice, the choice of the approach depends heavily on the context in which the calibration is done.

An anchor can be either external or internal. The former involves candidates completing a standalone anchor test after the main test, whereas, for the latter, items are embedded in the main test and candidates cannot distinguish between the anchor and non-anchor items. An external anchor has the disadvantages that, the addition of an anchor test may cause fatigue effects, or lack of motivation among the candidates, as their test result will come from the main test. An external anchor is most likely to be used for pretests, rather than for live tests for these reasons.

An internal anchor avoids the issues associated with an external anchor. For this reason, it may be possible to use

one for live tests and conduct pretesting at that stage. It may not be possible, however, to include items being pretested among those items which count towards the candidates' test result. If a test contains more items than are required and the anchor items alone are sufficient to provide reliable and comprehensive results, it is possible to use an internal anchor to pretest in a live test.

Conclusion

The focus of this article was on item analysis, one element of the quality assurance process in Cambridge English Language Assessment. The role item analysis plays in quality assurance can only be understood as a stage in a process, however. The purpose of the process is to ensure test forms are of high quality by measuring appropriate candidate abilities and the required level. To do this, information about the items must be gathered, processed and interpreted to facilitate decisions about the items. The quality of this information is therefore important and this is determined by the way in which it is obtained.

There is no single way in which the process of test construction and administration should be realised, however. Analytic choices are governed by the context and nature of the test concerned. The most important conclusion is, therefore, that the construction of systems of quality assurance for language tests must be carried out in a reasoned way, with item selection informed by robust statistical analysis in combination with expert judgement, so that high-quality measurement of candidates' language ability is achieved regardless of context or the nature of the test.

References

- Bock, R D (1997) A brief history of item theory response, *Educational Measurement: Issues and Practice* 16 (4), 21–33.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Traub, R E (1997) Classical test theory in historical perspective, *Educational Measurement: Issues and Practice* 16 (4), 8–14.

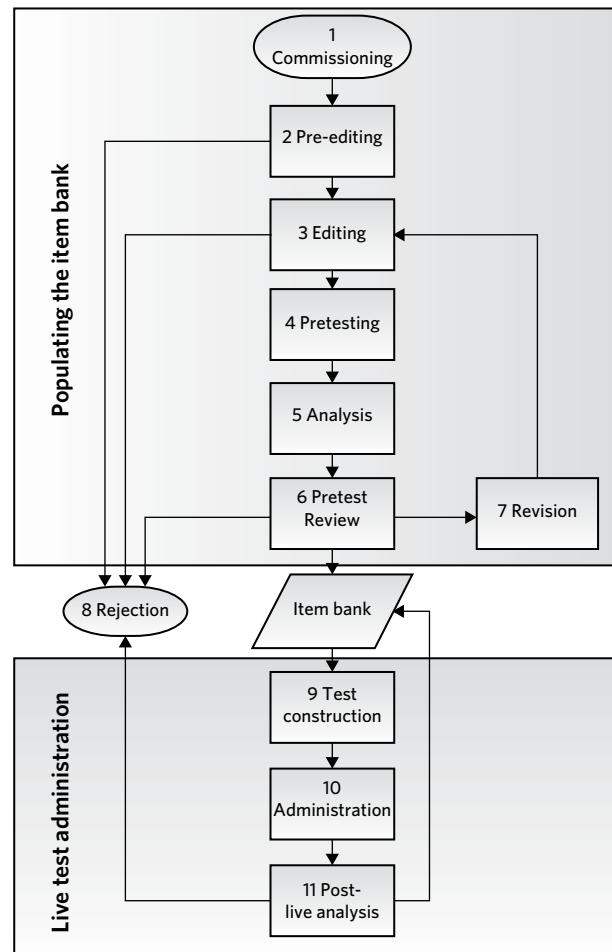
Appendix – A process of test construction and administration

An outline of a generic test production and administration process is provided in Figure 3 for reference. It contains two macro stages:

- populating the item bank
- live test administration.

The item bank is central to this model – it is the source of items for test construction in the live test administration stage. However, before test construction, the item bank must be populated with suitable items. Populating the item bank involves a pretesting stage. In this model tests are specially arranged for this purpose. Stages of editing take place before pretesting, so items with obvious problems do not undergo the expensive and logistically challenging stage of pretesting. After the results of the pretest are analysed, decisions over whether to bank items are made. If items are banked, their characteristics, such as difficulty and discrimination, are recorded together with them in the bank. This information facilitates test construction in preparation for live test administration. Further analysis follows live test administration, and the record of characteristics stored in the bank is usually updated, as live tests are expected to yield more accurate information about items. After a suitable delay to avoid security concerns over item exposure, the items may be used again in later tests.

Figure 3: The test production and administration process in Cambridge English Language Assessment



Test construction: The Cambridge English approach

COREEN DOCHERTY RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

DAVID CORKILL ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

Tests are designed for particular purposes and to meet specific identifiable needs such as to measure achievement, make selection decisions for further study or employment or provide diagnostic feedback to support lesson planning. Test scores are interpreted and have meaning only in relation to a test's purpose. The main aim of an exam board, such as Cambridge English Language Assessment, is to develop tests that measure the skills and abilities associated with a test's purpose in a reliable and valid manner. This involves tapping into the underlying trait or construct of interest through the development of tasks and their associated items, which can then be used to provide evidence of this ability while, at the same time, differentiating learners according to their mastery of these skills and abilities. A level of complexity is introduced by the fact that most exam boards typically produce multiple test versions, allowing test takers to sit a unique exam at different times and places. Without different test versions, there is a risk of malpractice as the test content may become familiar or overexposed to the candidate population prior to the test date and thus reduce the trustworthiness of the results (see Bell (this issue) for more information on malpractice). These different versions need to be parallel in terms of the constructs covered and their statistical properties in order to facilitate test equating protocols and ensure a comparable and fair test-taking experience for candidates. Construct comparability across test versions is derived from the test specifications and practically realised by following an explicit and standardised approach to test paper production while the psychometric comparability is underpinned by the generation of statistical measurement indices and the use of an item banking approach.

In this article, we will discuss how Cambridge English assembles tests in order to generate scores that can be used to make inferences about test takers' language ability. In particular, it will be discussed how test specifications, an item banking approach and statistical modelling facilitate the construction of parallel test versions. As this article will focus on test construction from an operational and psychometric perspective, readers who are interested in finding out more about the content and underlying traits covered in Cambridge English exams are invited to read the construct volumes published in the Studies in Language Testing series for each skill (Listening: Geranpayeh and Taylor (Eds) (2013); Reading: Khalifa and Weir (2009); Writing: Shaw and Weir (2007); Speaking: Taylor (Ed) (2011)).

Ensuring construct comparability: Test specifications

Test specifications are guidelines which define the constructs covered in an exam and how they are tested. These documents are, as Davidson (2012:201) points out, meant to be 'generative' in that they support the production of multiple, standardised items/tasks and thus comparable test versions. At Cambridge English, test specification documents, which take the form of Item Writer Guidelines and Handbooks for Teachers, are produced for each exam. These documents outline the structure of the exam: the number of tasks in a test, the range of task types used, and the number of items in each task. They also describe the content of the exam, such as the testing focus for each section and the range of testing focuses that should be targeted in each task/test to ensure adequate construct coverage. For example, a reading task may be designed to provide evidence of a candidate's ability to read carefully and, in order to do this, a range of items with different testing focuses are produced to elicit different aspects of this ability. If the range of testing focuses is not specified, there is a danger that a task or test may only cover a small subset of the underlying traits of interest and result in construct underrepresentation. The specifications also contain information germane to the construct of the test at the relevant Common European Framework of Reference (CEFR) levels (Council of Europe 2001), such as Reference Level Descriptors which specify the grammar, vocabulary and functions indicative of a particular level (see the English Profile at www.englishprofile.org for more information) or the amount of support or scaffolding included in the test depending on the level of learners. For example, tests at A1 and A2 tend to include more visuals to provide support and there are empirically based wordlists at these levels to guide learners' vocabulary development in preparation for the exam; similarly there are minimum and maximum word lengths specified for an input reading passage at each level for each task.

The test specifications for item writers are stable and practical documents used by item writers when commissioned to produce tasks. They include information about the test construct, task requirements, example items and guidelines about topic choice (Weyant and Chisholm 2014). The guidelines also include information, based on past experience, which can assist item writers in the writing process; for example advice on the features to look for when searching for a source text or those to avoid. More information on the process involved in developing new tests and defining the associated test specifications can be found in Ingham and Thighe (2006) and Corkill and Robinson (2006).

Ensuring psychometric comparability: Item banking

An item bank is a database used to manage test content. The data held in the item bank will include information on the content and structure of tasks and associated item statistics. At Cambridge English, we use an item banking approach to facilitate test comparability through the systematic linking of items across levels using a single measurement scale (see Elliott and Stevenson, this issue). Thus, the item bank plays a pivotal role in the test construction process.

Any item bank is only as good as the quality and richness of the data it contains: the less rich the data, the more manual certain processes will have to be, while conversely the richer the data, the more automation is possible. The item bank software will generate some task data automatically, while other data is entered by the user at certain stages in the task lifecycle. Examples of data that is generated and captured automatically are the task history – which test versions a task has been used in – and any amendments made to the task document, which enables the user to see how a task has evolved. Some data is encoded automatically in the task ID when it is authored such as the particular product it was written for, the task type and the target age group. The data that is manually entered for each item and task includes the topic, the word count and the testing focus of the items. For a listening task, further information such as the name of the actors and the accents used can be added following the recording of the audio.

Statistical information, produced after a test administration from candidate responses, is also uploaded to the relevant item. This data includes the difficulty of the item, the facility, which is the proportion of candidates who answered the item correctly, and the discrimination index, which indicates how well the item discriminates between the strong and weak candidates (see Corrigan and Crump (this issue) and Elliott and Stevenson (this issue) for more information on how these indices are calculated). As Cambridge English trials or pretests items before including them into a live test, items are also classified according to their calibration status: uncalibrated, part-calibrated or fully calibrated. An uncalibrated item has no item statistics because 1) it has not been included in a pretest or 2) it has been pretested but some changes, which could affect its statistical properties, were made to the item in response to how it performed during pretesting (see Corrigan and Crump (this issue) for more information on pretesting). An item that is part-calibrated has been pretested and the item statistics are indicative of how the item is expected to perform when included in a live test, and given to test takers within a similar ability range as those who sat the pretest. After items have been included in a live test session and taken by a sufficiently high number of candidates with different first language backgrounds, they become fully calibrated. The calibration process establishes the links between items at different levels through the use of common items or anchor tests (see Corrigan and Crump, this issue).

The information on task content and the statistical properties of items contained in the item bank facilitates test construction by providing the information that can be used to ensure that test versions are comparable on these dimensions. Both of these are necessary to ensure that a test presents

good evidence of validity and reliability across sessions. A test which fulfilled all of the content criteria but the items included did not fit the design of the test statistically, or vice versa, would not be fit for the intended purpose or fair. Such an outcome might signal instability of the test results across one administration to another.

Test construction

Cambridge English is known for producing a range of standardised international tests, aligned to the CEFR, and reported on a common measurement scale through the application of item response theory (IRT) (see Elliott and Stevenson, this issue). Some of these exams such as *IELTS* (*International English Language Testing System*), *BULATS* (*Business Language Testing Service*) and the Cambridge English Placement Test are designed to measure at several different CEFR levels, while others target a particular level such as *Cambridge English: Preliminary* or *Cambridge English: First*, which measure at B1 and B2 respectively. In addition to these standard products, we also develop bespoke exams, which are used for projects such as to benchmark learners or designed for a specific set of learners. The term ‘benchmarking’ refers to providing a snapshot of a population’s abilities, which is typically used to make policy or educational decisions. Bespoke tests may report on one or more CEFR levels as the particular project dictates and are linked to the same standard as the rest of our Cambridge English exams through our item banking approach. For the purposes of this article, we will illustrate how tests are constructed to target one CEFR level and then how this applies to multi-level bespoke products.

Test construction criteria

Just as test specifications provide information about what a candidate should be able to do at a particular level, test construction criteria offer guidance on the statistical properties of a test. These criteria help in ensuring that the items/tasks included in a test provide enough evidence to draw conclusions about a candidate’s language ability, while minimising any measurement error. The test construction criteria consist of the mean difficulty targeted by a particular test and the standard deviation, which describes how much, on average, the item difficulties vary from the test mean. These two statistics together will determine the shape of the item distribution for the test. Statistical modelling techniques are used to identify the appropriate values for these statistics depending on the nature and purpose of the test as well as the associated Standard Error of Measurement (SEM). The SEM is a statistic related to the reliability of the scores and derived from the standard deviation and the number of items in the test. It defines the extent to which a candidate’s score may fluctuate if they took the test again (see Somers, this issue).

An illustration of the type of item distribution desired for a cohort with a known ability on items of increasing difficulty is seen in Figure 1. As the items become more difficult, the proportion of the candidates getting them correct (i.e. facility) diminishes. We can see that there is an optimum range of difficulties for testing these candidates. Between the two vertical lines, an increase in difficulty produces a greater

decrease in facility than on either side of the vertical lines, where the curve becomes flatter. This is because, beyond a certain level of difficulty, all of the candidates are tending to get the answer wrong and, up to a certain level of difficulty, all of the candidates are tending to get them right. None of these items, in the lighter portion of the graph, would usefully help in differentiating the strong candidates from the weak ones and their inclusion would simply make the test unnecessarily long. Therefore, the test specifications for a Cambridge English test positioned at a particular CEFR level stipulates that a test should have items with a normal distribution in the range between the two vertical lines. Figure 2 illustrates a normal distribution for a 40-item test, formed as a bell curve around the mean. If this item distribution is well targeted at the language ability of the test takers, then it will differentiate between learners, who are more or less proficient, quite well. That is, the test will have maximum precision around the mean so these items will be most useful for distinguishing between those who pass and those who fail. The easier items in Figure 2 will help to differentiate the weaker candidates from one another, while the harder items will do the same for the more proficient candidates. It is these easier and more difficult items which allow for enhanced certification in that they provide the information to identify learners who are performing above the level or just below it (see Somers, this issue). In this example, the standard deviation is small to reduce the SEM. If the standard deviation was not small, and we wanted to keep the SEM small, we would need to increase the number of items.

Figure 1: Model of the relationship between facility and difficulty

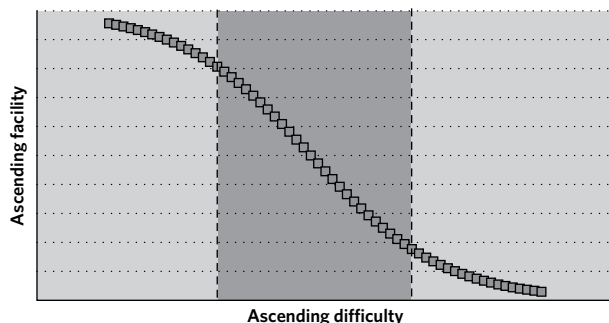
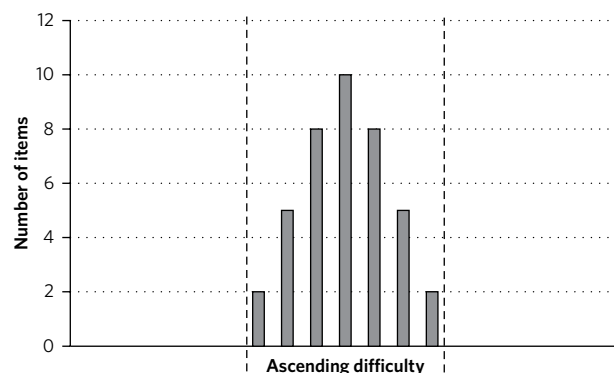


Figure 2: Normal distribution for a 40-item test



Having considered how tests are constructed in general, we shall now look at how test construction is modelled for exams with different purposes. Figure 3 shows a 40-item test with a standard deviation of 5 and with a normal distribution

around the mean difficulty of the test as seen in Figure 4. The CEFR threshold boundaries are represented by the crosses in Figure 3. A threshold boundary reflects the score required to achieve a particular CEFR level. For example, a candidate would need to get 20 items correct on this test to achieve B1 and 29 items correct to achieve B2.

Figure 3: Item curve for a test measuring at A2, B1 and B2

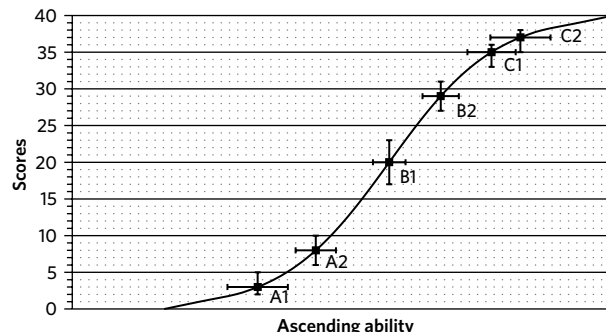
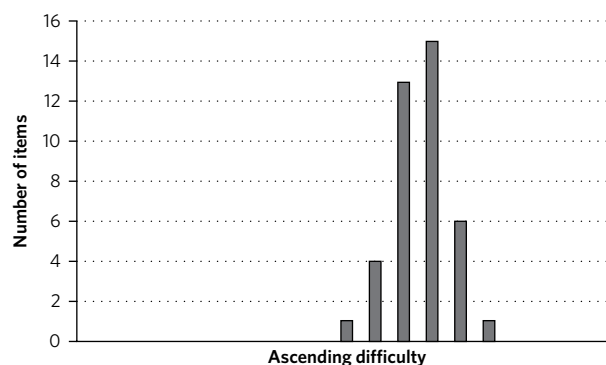


Figure 4: Item distribution for a test measuring at A2, B1 and B2



This test would be suitable for reporting at A2, B1 and B2 as we have maximised the reporting scores at these levels: there are more items at these levels and the curve is the steepest between A2 and B2, which is an indication that the items are discriminating well for these levels. However, there is an insufficient number of items at A1, C1 or C2 to report on those levels with any precision, which is indicated in Figure 3 by the fact that the thresholds are bunched together at the extremes and the curve has become flatter. The separation between the C1 and C2 threshold is too small to make a meaningful distinction between these levels. The SEM would indicate that candidates' scores could randomly fluctuate between C1 and C2 if they took the test again. A test with this item distribution would be most appropriate for a group of candidates who are expected to have an expected ability between A2 and B2. The items outside this range could provide additional information to stakeholders who may want to have an idea of how many candidates are performing above or below these targeted levels. This test would also be suitable for making comparisons between candidates within the levels reported on (i.e. A2, B1 and B2) although, as mentioned in Somers (this issue), care has to be taken when drawing conclusions based on similar performances on a single test event with only a few items used to measure a particular level.

The test modelled in Figure 5 has the same constraint of only containing 40 items and has the same mean difficulty as

the test in Figure 3. However it has a standard deviation of 25, and the items form a uniform distribution around the mean (see Figure 6). A uniform item distribution, unlike a normal item distribution, is most appropriate for tests targeting several CEFR levels. Items are not clustered around the targeted mean difficulty but equally spread across a range of difficulty targets.

This test, therefore, would not be as good a tool as the one in Figure 3 for reporting A2, B1 and B2 performance as there are fewer reporting scores between each grade boundary in this range. However, it would be a better tool for providing a broad description of how a given population lay across the whole range of CEFR levels because it covers a wider range of items. This would be useful for example, for benchmarking purposes. Naturally, the more widely we try to report with the same number of items, the less accurate that reporting becomes because the SEM increases. If we compare the item curve in Figure 5 with the one in Figure 3, it can be seen that the former is flatter, which indicates that it does not discriminate between CEFR levels as well as the latter.

Figure 5: Item curve for a test measuring across all CEFR levels

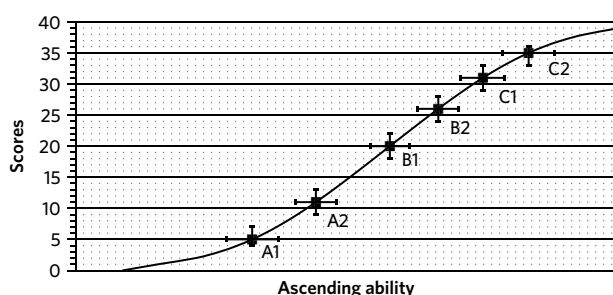
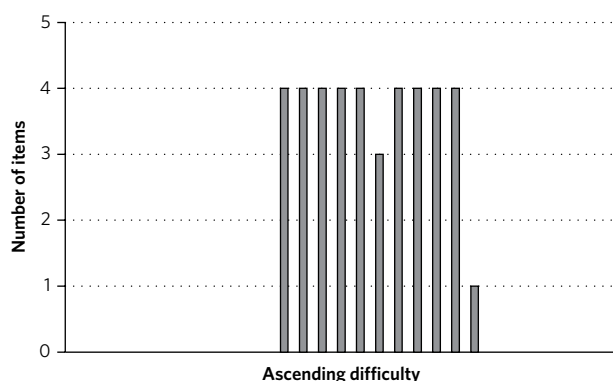


Figure 6: Item distribution for a test measuring across all CEFR levels



By adjusting the mean, standard deviation or length, one can target a test to report on the levels of interest more or less precisely. At Cambridge English, we have identified the optimal values for the mean, standard deviation and SEM for each of our products based on large datasets of candidate performance over decades, trialling and standard setting exercises (see Lim, this issue). During the design stage of a new or bespoke test, the ability to model the item distribution facilitates the production of tests which are fit for purpose in each instance as different test designs are appropriate for different purposes. The rich data held in the item bank is essential for designing well-targeted tests, both on a psychometric basis and in terms of content. As we will have a very clear idea of the candidature, such as their age range, we can select tasks which will have engaging topics. We can also select tasks which were calibrated on a

similar population. Although both *Cambridge English: Business Preliminary* and *Cambridge English: Preliminary for Schools* are B1 level exams, we could not expect the items to perform according to their statistical profile if we put tasks calibrated on one population into a test intended for the other because the target language use of the former exam is the workplace whereas the latter is a General English exam for teenagers. It is important to ensure that we can have confidence that the items will perform in line with their statistical profile.

Ensuring quality

As mentioned earlier, test construction can involve either manual, automated or a combination of manual and automated processes. Regardless of which is used, a series of checks and balances should be embedded into the test paper production process to ensure that tests are constructed according to the test specifications and test construction criteria. As Weir (1993) points out, testing should be considered a group activity as the review and discussion of items, tasks or tests helps maintain a consistent standard. To this end, at Cambridge English, test construction involves many different people at different stages of the process and for different purposes. For example, after tests are constructed, a meeting is held with those responsible for managing the exam and key external consultants who are involved in writing for the exam to check that statistical criteria and content specifications are met. During this meeting the participants also ensure that key or topic overlap is avoided. Test papers are then reviewed by content vetters as a further check that the papers are adhering to test specifications and construction parameters. These additional checks also contribute to the quality of the tests produced and the comparability of different test versions (see Beresford-Knox (this issue) for more information about quality assurance processes).

Conclusion

The main aim in the test construction process is to provide candidates with a valid and reliable test which contains a suitable range of items/tasks which allow them to demonstrate their ability. Clear guidelines and criteria that define the test content and the statistical properties of a test are important in ensuring that different test versions are comparable, test scores are a reliable indication of performance and that candidates are being given a similar test experience regardless of when they take the test. Statistical modelling can inform test design to ensure that the test is effective and the rich data in the item bank allows suitable content to be selected. The more sophisticated the item bank, the more automation can be applied to processes but, to ensure the highest levels of quality and fairness, expert analysis of test versions is also necessary.

References

Corkill, D and Robinson, M (2006) Using the global legal community in the development of ILEC, *Research Notes* 25, 10–11.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.

Davidson, F (2012) Test specifications and criterion referenced assessment, in Fulcher, G and Davidson, G (Eds) *The Routledge Handbook of Language Testing*, London: Routledge, 197–207.

Geranpayeh, A and Taylor, L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Ingham, K and Thighe, D (2006) Issues with developing a test in LSP: The International Certificate in Financial English, *Research Notes* 25, 5–9.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice*

in Assessing Second Language Reading, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Shaw, S D and Weir, C J (2007) *Examining Writing: Research and Practice in Assessing Second Language Writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.

Weir, C J (1993) *Understanding and Developing Language Tests*, New York: Prentice Hall.

Weyant, K and Chisholm, A (2014) Safeguarding fairness principles through the test development process: A tale of two organisations, *Research Notes* 55, 3–6.

Grading and test equating

MARK ELLIOTT RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

LYNNE STEVENSON ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

Grading decisions involve applying interpretations to candidates' marks and converting them into meaningful scores and grade bands, allowing certificate end users, for example potential employers or universities, to draw appropriate inferences and make decisions where appropriate. This is a central aspect of the testing process and one on which its usefulness hinges.

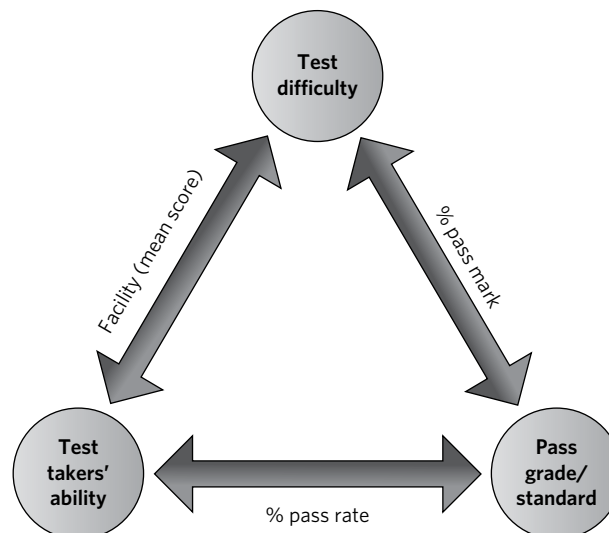
There are various possible approaches to grading, which may be appropriate for different contexts, but they can typically be classified into types:

- *Norm-referenced* grading, which awards grades by dividing the candidature into percentiles, for example 60% of candidates passed each administration and 10% achieved Grade A. This approach can be useful where a test is used to select candidates from a particular cohort, but involves problems of test equating (comparability of scores between test versions) – the passing standard changes according to the ability profile of the candidature – and of interpreting the scores in terms of their meaning beyond comparing candidates.
- *Criterion-referenced* grading, which explicitly links grading decisions to fixed levels of performance which do not vary from administration to administration. These fixed levels of performance may take the form of links to an external frame of reference such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), or they may relate to a reference scale specific to a particular examination, as is the case for example with *IELTS*. Proportions of candidates achieving each grade may vary from session to session according to the ability profile of the candidature but performance on different test versions is directly comparable.

The grading of norm-referenced exams is a relatively straightforward affair since the primary task is to link

percentiles of candidate scores to grades, which is statistically simple. The grading of criterion-referenced exams, however, is a more complex affair which requires establishing and maintaining the grading standards which have been chosen. Within a test, there is a triangular relationship between three statistical concepts: mean score, pass mark (the minimum mark on the test required to achieve a pass) and pass rate (the proportion of candidates achieving a pass). These three concepts are each determined by two of three factors: test difficulty, test takers' ability and pass grade/standard, as illustrated in Figure 1. For example, the mean score is function of test difficulty (mean scores will decrease as difficulty increases) and ability (mean scores will increase as ability increases). Knowing the difficulty of a test and the standard allows us to determine the ability of the candidates, which is of course the ultimate goal of the test.

Figure 1: The relationship between three key testing concepts (after Jones and Saville 2007)



Within this paradigm, grading is the means by which test equating is achieved, since the purpose of grading is to ensure that performance standards determined during a standards setting process (Lim, this issue) are applied uniformly across different test versions in order to determine candidates' ability – in other words that the tests are equated. This is not to be confused with the use of the term 'grading' in the US assessment context to refer to the marking process.

Cambridge English Language Assessment exams are all criterion referenced, with standards of performance related to the CEFR reference scales used for the majority of exams, although some exams such as *IELTS* and the *Cambridge English: Young Learners (YLE)* exams use their own standards to determine grading thresholds. With this in mind, this article discusses some of the issues involved in grading criterion-referenced exams at component and overall level and details Cambridge English Language Assessment's approach to producing robust, reliable and consistent grading decisions within such a framework. We will discuss, in turn:

- grading productive components (Writing and Speaking)
- grading objectively marked components (Reading, Listening and Use of English)
- producing overall results.

Productive components are treated separately since the different nature of these components necessitates different marking processes and different methods for interpreting results. The same underlying principles outlined in Figure 1 still apply to both productive and objectively marked components.

Grading productive components

Productive components can be graded in many different ways, depending on the response format and marking method employed. For example, Writing scripts can be computer marked using algorithms based on latent regression of particular linguistic features of the text in a manner which correlates highly with human markers; the mechanics of this, however, fall outside the scope of this article. Computerised marking of Speaking is also possible, although at present it is not possible to test a broad Speaking construct in a reliable manner. For the purposes of this article, we will restrict ourselves to considering Writing and Speaking tasks which are marked by humans according to a set of criterion-based assessment scales which require the use of expert examiners, and which are the most common task types in Cambridge English Language Assessment Writing and Speaking tests. Such tasks require the use of experienced language teaching professionals who are capable of interpreting a criterion-based mark scheme which is tied to an external frame of reference – in the case of Cambridge English Language Assessment exams, the CEFR (Council of Europe 2001) is used. Since such open mark schemes are liable to be interpreted differently by different examiners unless their interpretation is carefully standardised, it is necessary to establish rigorous systems, firstly of training, standardisation and co-ordination and secondly of monitoring. These will now be discussed in turn.

Examiner training, standardisation and monitoring

Cambridge English Language Assessment employs a highly structured Team Leader System, which for Speaking features a hierarchy of Speaking Examiners, Team Leaders, Regional Team Leaders and Professional Support Leaders. This system is critically important to ensure quality and consistency of marking across the 20,000+ Speaking examiners operating around the world. New examiners, who must meet a set of minimum professional requirements, first undergo initial training to familiarise themselves with all aspects of examining, including marking, materials handling and security issues. They then undertake annual standardisation sessions. Firstly, a series of exemplar Writing scripts or videos of Speaking tests are marked by a group of senior examiners. These marks are combined to produce standardised marks using Multi-Faceted Rasch (MFR) analysis, which is a variant of Rasch analysis as described below. The MFR analysis focuses on different facets of the testing situation (e.g. raters' severity/leniency, candidates' ability and assessment criteria) and results in fair average marks for each candidate, which are the marks adjusted for harshness/leniency of the raters and the ability of the candidate. This process of establishing a standardised mark independent of marker severity can be seen as analogous to the process of determining item difficulty independent of candidate ability for objectively marked components. Secondly, all examiners are given a selection of exemplar scripts/videos to mark, with the marks they award compared to the standardised marks; examiners must meet a specified level of accuracy before they are certified to mark in live sessions. All examiners up to and including Professional Support Leaders, must undergo the standardisation process.

Taylor and Galaczi (2011) provide a detailed description of the Team Leader System for Speaking. A similar system exists for Writing, although it is less elaborate due to the smaller number of examiners involved. One difference is that there is a co-ordination process for Writing examiners before each session to familiarise them with the task-specific marking criteria for a given paper. This does not apply to Speaking, as there are no task-specific criteria.

Examiner monitoring

Examiners' live marking performance is subjected to two kinds of monitoring to provide support and ensure consistency. Firstly, examiners are periodically subjected to live monitoring, in which a senior examiner observes their performance and provides judgements against a set of criteria which cover all aspects of test administration and marking. The results of live monitoring are fed back to the examiner and may form the basis for further training. Secondly, statistical monitoring of marker performance is carried out to ensure two aspects of marker reliability:

- *intra-rater reliability*, to identify erratic examiners whose marks are inconsistent
- *inter-rater reliability*, to identify examiners who are marking overly strictly or leniently.

Statistical monitoring can take the form of comparing the marks awarded by two examiners for the same task or comparing marks awarded to those in other components in

order to identify unusually large differences in performance between components (so-called 'jagged profiles'). Unusual marking patterns may lead to tasks being re-marked in order to confirm whether scores were appropriate (statistical monitoring alone cannot determine this) and may form the basis for further examiner training. A further means of statistical monitoring involves the dissemination to raters of seeded scripts (scripts for which standardised marks have been determined against which the consistency of an examiner's performance can be judged).

In terms of grading, a key feature of the use of criterion-based assessment scales is that it is the scales which need to be validated and aligned to an external framework rather than the tasks, which function as vehicles to allow candidates to express their ability. Tasks need to be trialled to ensure that they fulfil this function appropriately, eliciting the required language functions at the target proficiency level. Lim (2012) provides a full account of the process of validating a set of Writing assessment scales.

Since the Writing and Speaking assessment scales Cambridge English Language Assessment uses for examiner-marked tasks are aligned directly to the CEFR, grading is relatively straightforward provided that rater consistency is maintained: the raw marks total which corresponds to each CEFR threshold is known through the alignment of the assessment scales to the CEFR (Lim, this issue) and can be converted to the Cambridge English Scale score for the relevant threshold, while raw mark totals between thresholds can be aligned by a simple process of linear interpolation.

Grading objectively marked components

The output of a productive test component is a directly observable product – a written text or oral performance – against which the aspects of language ability being assessed may be judged directly. Objectively marked components, on the other hand, typically assess comprehension, which is an internal process. This 'black box paradox' means that comprehension needs to be tested indirectly, by having a candidate respond to a series of items based on input material (texts for Reading and audio recordings for Listening), which are designed to elicit the relevant aspects of language ability being assessed.

Candidate performance on objectively marked tasks is marked according to a fixed mark scheme on the basis of correct/incorrect responses, which should be free from judgement error (subject to appropriate checking processes), hence the term *objectively marked*. This distinguishes them from the productive skills, which require examiner judgement and are sometimes referred to as *subjectively marked*.

Candidates' marks for objectively marked components are derived from their correct/incorrect responses on test items rather than being a result of awarding a rating against defined criteria (i.e. the assessment scales for Speaking and Writing). Therefore, we need a means of relating what a correct response, or a certain number of correct responses on a test, mean in terms of the reporting framework (the CEFR and the Cambridge English Scale). This requires a process of *standards setting*, which is discussed by Lim (this issue).

There are issues with the use of raw scores from objectively marked components as determinants of candidate ability. While it is reasonable to conclude that a score of 21 on a test represents a better level of performance than a score of 20, it is not clear how much better, or whether the difference is the same as the difference between scores of 19 and 20. It is also not clear whether scores of 20 on two different test versions represent the same level of performance. In fact, using raw test scores to measure performance can be seen as analogous to counting the number of stops on a train journey as a means of measuring the distance travelled. Counting stops is certainly not an unreasonable method of judging distance travelled, and has the advantage of being simple and convenient, but it does have limitations:

- while we know that someone who has travelled one more stop has travelled further, we cannot quantify the extra distance without further information
- the distance between stops can vary along a given railway line (comparing the difference in underlying ability corresponding to an increase of one mark in a test at various points)
- a journey of five stops can be of very different lengths on two different lines (comparing the same score on two test versions).

The solution for comparing train journeys is to measure the distance in kilometres or miles; to provide a solution for comparing test scores, what is needed is a means of converting raw scores into a meaningful measurement metric analogous to kilometres or miles.

Rasch and item response theory (IRT)

The most common method of establishing a measurement metric for test items and candidates is to use statistical models from two closely related families of *probabilistic* models known as *Rasch* models, named after the Danish statistician Georg Rasch (Rasch 1960, Wright and Stone 1979) and *item response theory* (IRT) models (Lord 1980). In a probabilistic model, each candidate has a probability of obtaining a correct response to a given item, which differs from deterministic models which state that the candidate will definitely obtain a correct or incorrect response for a given item. In this sense, estimating a candidate's ability (which is what the whole testing process is designed to achieve) can be seen as being analogous to estimating the proportions of black and white balls in a large sack by pulling out balls at random – if we draw 10 balls and find eight black balls and two white balls, we might estimate the proportion of black balls to be 80% (i.e. the probability of drawing a black ball is 80%). This is an estimate, since it is only based on partial data (we haven't seen all the balls), but the more balls we draw, the more confidence we can have in our estimate. Similarly, we can estimate the ability of the candidate from the number of correct and incorrect responses to items in the test; the more items we have, the more confidence we can have in our estimate. The situation is naturally somewhat more complicated for a test, since the probability of a correct response to each item will vary depending on the difficulty of the item – what is needed is a statistical model

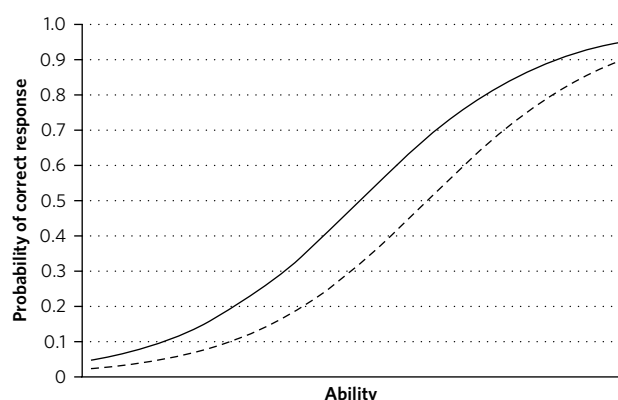
of the probability of a candidate of a known ability obtaining a correct response to an item which has given properties (known in IRT and Rasch as *parameters*).

Cambridge English Language Assessment mainly uses Rasch models, which are the simplest form of IRT models. For the case of a dichotomous item (i.e. possible scores of 0 or 1), the probability of a candidate achieving a correct response in the Rasch model is a function of the difference between the candidate's ability and the difficulty of the item in question; the ability and difficulty are measurements placed on the same unidimensional scale. Candidates of higher ability are able to answer the harder questions correctly (or, in probabilistic terms, they have a high probability of doing so).

Rasch/IRT ability estimates are presented in units called *logits* on an *interval scale*, which means that a difference of one logit corresponds to the same difference in ability anywhere along the scale (unlike a difference of one raw mark or one train stop). One important feature of Rasch difficulty estimates is that they are sample independent – in theory (and subject to certain limitations) they apply equally to any group of test takers, not only the group from which the calibration was drawn.

Figure 2 shows the relationship between the ability of a candidate and their probability of obtaining a correct response to two items with different difficulties under the Rasch model, called *item characteristic curves*. The probability is almost zero for candidates with low ability, then increases and approaches one as ability reaches high levels relative to the difficulty of the items. The curve to the left represents an easier item since the probability of a correct response is higher for any ability (the curve is higher). The difficulty parameter of the item is defined as the ability where the probability of a candidate responding correctly is 0.5.

Figure 2: Item characteristic curves for two items under the Rasch model



Rasch models form the simplest type of IRT model, featuring only one item parameter (difficulty); as well as the Rasch model family, there are two more commonly used IRT model families:

- Two-parameter (2PL) IRT models, which add an item discrimination parameter – item response curves may have slopes of different steepness and cross as a result, meaning that for high-ability candidates, the item with the steeper curve is easier than the item with the shallow curve, but for low-ability candidates the situation is reversed.

- Three-parameter (3PL) IRT models, which introduce a lower asymptote – a minimum probability which candidates with extremely low ability approach. This is also referred to as a guessing parameter since it models the probability of guessing a correct answer randomly in, for example, a multiple-choice item.

Figure 3 shows item characteristic curves for two items in the 2PL model with the same difficulty but different discrimination parameters, with correspondingly different slopes. The steeper item with the steeper slope has higher discrimination, i.e. it shows greater difference in performance between two candidates of different abilities. The two curves cross at the point where person ability equals item difficulty, which is why the difficulties are the same for both items; high-ability candidates will find the item with the steeper curve easier than the item with the shallow curve, but for low-ability candidates the situation is reversed. The discrimination parameter corresponds to the steepness of the item characteristic curve.

Figure 3: Item characteristic curves for two items under the 2PL IRT model

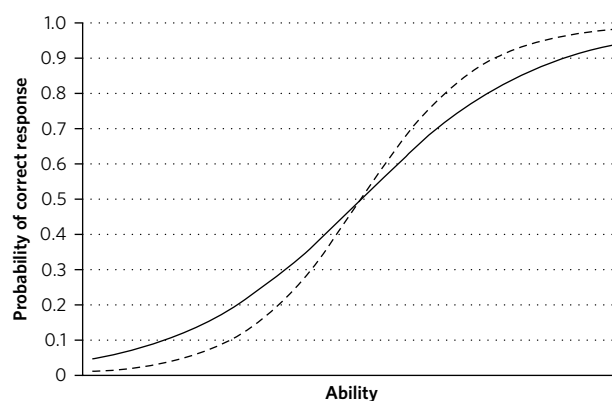
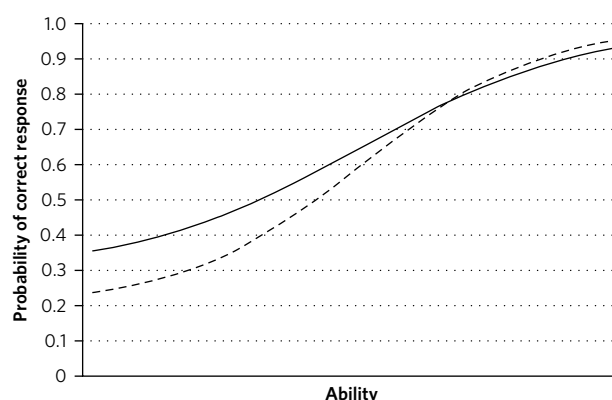


Figure 4 shows item characteristic curves for two items in the 3PL model. Note that both the slope (the discrimination parameter) and the lower asymptote (the guessing parameter) are different. The guessing parameter corresponds to the lower asymptote – the level below which the probability of a correct response never falls, no matter how low a candidate's ability; for example, for a 4-option multiple choice-item, this is likely to be around 0.25 since this is the probability of obtaining a correct response by selecting an option at random.

Figure 4: Item characteristic curves for two items under the 3PL IRT model



Given a set of response data, the relative difficulties of the items in a test can be calculated, and by employing shared items of known difficulty across different test versions (known as *anchor items*), items from different test versions can be calibrated on the same scale. It is this critical point which is central to the usefulness of Rasch/IRT – it provides a means of ensuring that the same standard is applied at each testing session, whether at the live stage of the test production cycle to facilitate grading decisions or at the pretesting stage to facilitate quality and consistency in test construction (see Corrigan and Crump (this issue) for a discussion of pretesting).

Each model has its advantages and disadvantages. On the one hand, models with more parameters will always provide a better fit to the data since they can account for more causes of variation, but on the other hand they are less robust and require significantly larger sample sizes to produce sufficiently accurate calibrations for pretesting purposes, which presents logistical issues: while the Rasch model requires a sample size of around 250, the 2PL and 3PL models require sample sizes of around 500 and 1,000 respectively for the same level of accuracy (Hulin, Lissak and Drasgow 1982). See Corrigan and Crump (this issue) for a further discussion of pretest sample selection.

There are ongoing conceptual debates about many aspects of the models and which is preferable: for example, there is the question of whether it is acceptable that two items with the same difficulty have different discriminations and crossing curves (see Figures 3 and 4) as can happen under the 2PL and 3PL models. Items with high discrimination are considered better since they differentiate more sharply, as are items with difficulty around the target ability levels, since they provide more information and result in a lower *Standard Error of Measurement* (SEM), which is a means of evaluating the precision of a test.

Under the Rasch model, a candidate's total score is a sufficient statistic (i.e. one that contains all the information needed) to produce an ability estimate – it is the number of correct responses which determines a candidate's ability estimate rather than which items the candidate answered correctly; this feature of the Rasch model, where ability estimates preserve the same order as the raw score, is one reason Cambridge usually uses the Rasch model rather than other IRT models. Unlike the Rasch model, under the 2PL and 3PL models, two candidates with the same total raw score can achieve different ability estimates since the item discrimination and guessing parameters are also used to calculate ability estimates (Lord 1980). This added (and counter-intuitive) complication constitutes a practical detriment compared to the Rasch model, which relies on simple raw scores, for little real practical benefit.

When employing any statistical model, it is always important to remember that, as Rasch himself notes (1973), 'every model is basically wrong, i.e. it is bound to fail, given enough data'. The point is that all models – and this extends to models used in other disciplines such as physics and weather forecasting – are simplifications of a complex reality. When using a model, the key question to ask is not whether a model is right – it is not – but whether a model is useful in that it is accurate enough to produce reliable results. Fortunately, it is possible to test the extent to which the data fits the model,

and whether this degree of fit is acceptable for the purposes for which it is being used – this will vary, for example, between a relatively low-stakes test like a school's placement test and a high-stakes test like *Cambridge English: Advanced* being used as a university entrance requirement.

In sum, Cambridge English Language Assessment employs the Rasch model on the following grounds:

- the model is practical to employ, particularly for pretesting purposes
- although there will be more items which do not fit the Rasch model than with the 2PL and 3PL models, these items can be discarded for live test construction.

All models carry with them certain assumptions which may be violated to a greater or lesser extent in practice, with consequences of greater or lesser significance. In the case of the Rasch model, key assumptions include:

- unidimensionality of the trait under investigation
- that the difficulty of an item is the same for all candidates
- local item independence (a candidate's response to one item does not affect their chances on subsequent items)
- that the probability of responding correctly to an item approaches zero at very low ability levels (i.e. no guessing parameter).

In practice, the first assumption (which also applies to the 2PL and 3PL models) is never met, but it needs to be interpreted as a matter of degree rather than an absolute – while there are different factors involved in Reading proficiency, for example, there is enough similarity between them for the Rasch model to provide accurate enough results. Indeed, if the trait was truly unidimensional, there would be no need to sample broadly from across the construct since candidates could be expected to exhibit the same level of performance on, for example, both tasks focusing on reading for specific information and those focusing on reading for gist. This, of course, is not fully true for a process as complex as reading comprehension; however, we can reasonably claim that there is an underlying general Reading proficiency trait and that the minor variations around this trait represented in different test parts do not constitute a gross violation of the assumption of unidimensionality, as demonstrated by Geranpayeh (2007), for example, using a statistical method known as *Structural Equation Modelling*.

Tests can be carried out to investigate whether any test items perform differently for different candidates, or rather different groups of candidates sharing a trait – for example, whether they are harder for male candidates than for female candidates; this phenomenon is known as *Differential Item Functioning* (DIF). Cambridge English Language Assessment carries out regular analyses to ensure that live test items are functioning to within acceptable parameters (Corrigan and Crump, this issue).

In a similar vein, the assumption of independence does not always fully hold – multiple-matching tasks, for example, violate this assumption to an extent, which will inevitably have an effect on the accuracy of the test, although the effect should not be large unless there are a large number of such items in a test; this is one reason why Cambridge English Language Assessment employs a range of

response formats (see Elliott and Wilson (2013) for a fuller discussion on the use of different response formats). There is broader consideration here, which is that psychometric aspects need to be balanced against other aspects related to construct coverage – the multiple-matching gapped-sentence task in the *Cambridge English: First Reading* component, for example, provides an excellent means of testing a candidate's awareness of textual coherence and cohesion which it is difficult to achieve under other response formats; rejecting such a task on the grounds of relatively minor violations of statistical principles would be a case of putting the narrow psychometric cart before the broader language testing horse and could eventually lead to an impoverished test which only covered a narrow range of the Reading proficiency construct at CEFR Level B2. As Davidson (2000:605) notes, 'statistical determinism is a dangerous epistemological force'.

The absence of a guessing parameter in the Rasch model is a subject of debate; while the 3PL model formally models random guessing, proponents of the Rasch model take the opposite viewpoint and argue that in fact it is the 3PL model which does not take account of test taker behaviour in high-stakes exams, where guessing is rarely random but rather typically informed by partial understanding, while in any case item fit criteria will reject items which exhibit a large non-zero asymptote. Additionally, item statistics can become sample-dependent under the 3PL model, since a sample with a higher propensity to guess is likely to produce higher guessing parameters. In any case, as already stated, the critical question is whether the model fits well enough for the purpose for which it is being used, and appropriate item selection criteria can ensure that this is the case for the Rasch model, although this does not mean that the 2PL or 3PL models should not be used – what is important is choosing an appropriate and adequate model for a given testing context.

In the past, the Rasch model provided a useful means for Cambridge English Language Assessment to set consistent cut scores but with the introduction of the Cambridge English Scale it provides a means to convert raw scores onto one metric – for objectively marked components, the Cambridge English Scale is a transformation of the underlying Rasch scale, presented in an easily interpretable format.

Producing overall results

Once the individual components of an exam have been scored and graded, the results need to be combined to produce a single overall result and grade. There are two main approaches to this problem, both of which are theoretically defensible:

- A *non-compensatory* approach, which stipulates that a candidate must achieve a given threshold in every component in order to achieve that threshold as an overall result. The argument for a non-compensatory approach is that to be B1 level it is necessary to demonstrate B1 competence in each skill independently – a particular case where non-compensatory aggregation may be appropriate could be a test for air traffic controllers, where it is absolutely necessary to reach the standard in all the

competencies tested since they are all essential to ensure the safety of air traffic.

- A *compensatory* approach, which allows for performance below a given threshold in one component to be compensated for by stronger performance in other components. The argument for a compensatory approach recognises that learners have jagged profiles (differences in terms of level across skills) and that even within a particular skill there may be some aspects, which a learner who is classified at B1 level still operates at A2 level; a compensatory approach extends this logic across the full range of skills.

Cambridge English Language Assessment exams follow a compensatory approach where a candidate's overall Cambridge English Scale score is the average of the component scores, on the same reference scale, with the overall CEFR level determined from the overall Cambridge English Scale score.

Concluding remarks

Grading and test equating are complex and intertwined issues. There is no 'best' approach, but rather a range of options which may be more or less fit for purpose in a given context. The responsibility of a test provider is to make a principled (and practical) choice, and to back up this choice with both a rationale and evidence of fitness for purpose. This article has attempted to sketch out the approach chosen by Cambridge English Language Assessment and to highlight the use of controlled procedures and statistical analysis to ensure that grading decisions are consistent and that standards are maintained across test versions.

References

- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Davidson, F (2000) The language tester's statistical toolbox, *System* 28, 605–617.
- Elliott, M and Wilson, J (2013) Context validity, in Geranpayeh, A and Taylor, L (Eds) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press, 152–241.
- Geranpayeh, A (2007) Using Structural Equation Modelling to facilitate the revision of high-stakes testing: the case of CAE, *Research Notes* 30, 8–12.
- Hulin, C L, Lissak, R I and Drasgow, F (1982) Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study, *Applied Psychological Measurement* 6 (3), 249–260.
- Jones, N and Saville, N (2007) Scales and frameworks, in Spolsky, B and Hult, F M (Eds) *The Handbook of Educational Linguistics*, London: Wiley-Blackwell, 495–509.
- Lim, G (2012) Developing and validating a mark scheme for Writing, *Research Notes* 49, 6–10.
- Lord, F (1980) *Applications of Item Response Theory to Practical Testing Problems*, New York: Routledge.
- Rasch, G (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Nielsen and Lydiche.

Rasch, G (1973) *The notion of redundancy and its use as a quantitative measure of the deviation between a statistical hypothesis and a set of observational data*, paper presented by Per Martin-Löf, at the Conference on Foundational Questions in Statistical Inference, Aarhus, Denmark, 7–12 May 1973.

Taylor, L and Galaczi, E (2011) Scoring validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Cambridge: UCLES/Cambridge University Press, 171–233.

Wright, B D and Stone, M H (1979) *Best Test Design: A Handbook for Rasch Measurement*, Chicago: MESA.

Determining standards in assessing language ability

GAD S LIM RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

Tests are often taken to inform decision making, for example, whether a person knows enough about a subject to proceed to the next level or whether a person has the skill to perform a job. In some cases, this activity is quite straightforward. In order to determine whether someone knows how to jumpstart a car, a person can be given the leads and told to perform the task, and if the engine roars to life, we know that the person indeed knows how to jumpstart a car. Furthermore, the task is basically the same whatever the conditions (e.g. a large car or a small car), so it is not difficult to say that the person will be able to do it under other conditions. In other cases, however, testing can be more complicated. The ability being tested may not be a tangible, physically observable thing, or it may be affected under different conditions.

An example of this more complicated kind of testing is the assessment of language ability. Listening and reading ability cannot be directly observed, as they involve things happening inside the human brain. Writing and speaking ability can be directly observed, but performance can be affected by context – people might find it easier talking to their friends than to their boss, for example. A test cannot cover every possible context where language might be used. Thus, testing the language skills necessarily involves inference, and necessarily requires making determinations on whether a person is at a given level or not, able to do something with language or not. How these determinations are made in ways that are defensibly valid is the subject of this article, and is discussed with specific reference to Cambridge English exams and practice.

Scores that facilitate setting and maintaining of standards

To begin with, let us consider tests which contain items that are marked correct or incorrect, leaving aside for now tests that are marked on a scale by examiners. In many testing contexts, the number of items answered correctly is added up to produce a total score, or the total divided by the number of items in the test and multiplied by 100 to arrive at a percentage. It may be tempting to simply set a particular total score or a particular percentage correct as the standard. This, however, can be problematic. Imagine for example that the

next time around, the test questions were harder. If the same candidates retake the test, they would have a lower total score and a lower percentage correct. Their ability would not have changed, but this time around they might be deemed as not meeting the standard where previously they were considered as having met the standard.

Number of percentage correct is demonstrably inadequate for determining comparable performance. In the interest of fairness and validity, if different forms of a test are used, the tester needs to determine how much harder or easier the items in one test form are compared to another and then make adjustments accordingly. Two approaches to making the adjustment come to mind. One, a new total score or percentage correct that correctly represents the standard can be determined for each new test form. Two, the tester can determine what performance on one set of items equates to in terms of a standardised score scale which does not change. With an unchanging score scale, a location on the standardised scale can be identified once as representing the standard, and the activity of setting standards does not need to be repeated any more thereafter.

The first approach is obviously quite cumbersome, and can be very confusing for test users, who might incorrectly conclude that the standard had changed. The second approach has the advantage of simplicity. With a standardised score scale, the standard does not need to be determined anew every time, and the same number/score always represents the standard.

The latter approach is what Cambridge English uses. In particular, there is a statistical approach called Rasch measurement which can account for the relative easiness/difficulty of test items in determining a candidate's ability and place the test taker on an underlying scale. The underlying scale is sometimes transformed in some manner to facilitate reporting – in the case of *IELTS*, for example, turned into a 9 band system – but nevertheless retains the properties and advantages of a standardised scale. This statistical approach and its use within Cambridge English to ensure score consistency is discussed further by Elliott and Stevenson (this issue). The way underlying scales are transformed to facilitate reporting of scores in a meaningful way and across Cambridge English exams is discussed further by Somers, (this issue).

In any event, raw score totals and percentage correct do not facilitate the setting of standards, whereas standardised

score scales provide a basis for doing this that is both sound and efficient. With this in hand, we turn now to the process of determining what score or level of performance represents attainment of a particular standard. Standards that are properly set are crucial, as they render meaningful the consistent practice described by Elliott and Stevenson (this issue) and underpin the reporting of outcomes discussed by Somers (also this issue). We will consider three ways of setting standards, which Cambridge English uses in integrated fashion to set standards 1) by building them explicitly into test materials; 2) through a process of expert judgement; and 3) by comparing the test to another measure of the ability we are interested in. We now go through each of these in turn.

Standards determined a priori in test materials

Standards, in the more general sense of the word, need to be built into test materials first of all. If the test does not cover the level desired and the abilities we are interested in, or if it does cover them but using items that do not work well, then any score set as the standard will not be valid. For this reason, specifying the level of a test, what a test covers, ensuring that those specifications are consistently followed, and trialling to confirm that test items actually work, are crucial. Cambridge English practice in these regards is discussed by Docherty and Corkill and by Corrigan and Crump (this issue).

Beyond ensuring that test materials are of good quality, a relationship to external benchmarks can also be built into certain kinds of tests. For example, many people want to know how exam outcomes relate to the levels of the Common European Framework of Reference (CEFR) (Council of Europe 2001). (As an aside, it should be noted that the CEFR is a framework of reference and not a standard. For more on this, see Jones and Saville 2009.) Because this reference framework is so widely used, it has been explicitly built into certain Cambridge English papers. Historically, some Cambridge English exams were actually the basis for defining some of the CEFR levels (North 2014). Nevertheless, in order to further tighten the relationship, the assessment criteria for most Cambridge English Writing and Speaking papers were redeveloped to explicitly reference the CEFR. In particular, the performance descriptors were developed following an iterative process to match the levels of the CEFR. (For more detail, see Galaczi, French, Hubbard and Green (2011) for Speaking and Lim (2012) for Writing.) Thus, because the descriptor for a particular score matches a particular CEFR level, anyone who obtains that score is by definition at that level. The passing standard has been built into the assessment criteria, and does not need to be separately determined.

It might be worth pointing out here that the assessment scale for each exam includes descriptors covering three CEFR levels (Figure 1). For example, *Cambridge English: First* is aimed at the B2 level, and a candidate whose performance matches that of the B2 descriptor for a particular sub-scale gets a 3. Note however that the assessment scale for that exam has a score point 5 which relates to the C1 descriptor. Thus, a candidate who obtains 5s on this test across the different criteria (and papers) has shown evidence of C1

level performance. This is in part the basis for the Cambridge English reporting of scores on the Cambridge English Scale

Figure 1: Stacked design of assessment scales for some Cambridge English exams

	Preliminary	First	Advanced	Proficiency
				5
C2			5	3
C1		5	3	1
B2	5	3	1	
B1	3	1		
A2	1			

above and below the level of the exam they sat.

It can be seen that particular performance levels can be built directly into test materials, and the predefined standard can simply be applied (rather than having to discern the standard afterwards each and every time).

Standards determined by expert judgement

Another way of determining what level of performance reflects a particular standard is to engage a panel of experts who make judgements about where the 'cut score' is. Quite often, when people speak of 'standard setting', this approach is what they have in mind.

Because this approach depends primarily on human judgement, steps need to be taken to ensure the validity of standard setting outcomes. For example, many different standard setting methods have been developed, each with their own strengths and weaknesses. Different test types also lend themselves to different standard setting methods. If one is not aware of a particular method's limitations, or about their appropriateness for the test type, they might draw wrong conclusions about what the correct cut score is.

It has also been observed that when standard setting exercises are repeated, different outcomes are sometimes obtained. Standard setting theory says that this is not a problem, explaining that it is a policy exercise rather than an exercise in identifying an objectively existing cut score. We would argue that it depends on the type of standard setting one is engaged in. For instance, an organisation might want to give scholarships to the five best applicants. The pool of applicants may well be very poor, so the standard will need to be relatively low in order to identify five winners. On the other hand, the pool of applicants may be very strong, so the standard will correspondingly be higher. In these instances, the standard may indeed change on different occasions, because the organisation has a particular policy goal (giving five scholarships, no more, no less).

On the other hand, there are (many more) instances where standard setting is an empirical exercise. For example, the CEFR is built around things that individuals can do with language. In determining cut scores vis-à-vis the CEFR therefore, there is a purported objective score at and above which someone can do said activities using language.

Therefore, different and repeated standard setting exercises should arrive at the same answer. Jones (2009) and Lim, Geranpayeh, Khalifa and Buckendahl (2013) argue this topic in greater detail. In any case, it is important to remember which type of standard setting one is engaged in.

Cambridge English exams generally involve objective standards, awarding candidates certificates indicating that they are at a particular level. In the article by Lim et al (2013), the authors report on a standard setting exercise that has been conducted with regard to one Cambridge English test – *IELTS* – and the CEFR. They give detailed consideration to standard setting method effects, and triangulated results with those obtained using other approaches, making the resulting cut scores more robust. Expert judgement-based standard setting has also been used by Cambridge English to determine standards relating to other language ability frameworks such as the Canadian Language Benchmarks (Buckendahl, Foley and Rodeck 2005) and for particular professions such as nursing (O'Neill, Buckendahl, Plake and Taylor 2007). It has also been used to reconfirm the cut scores for different exams in preparation for the introduction of the Cambridge English Scale.

Standards determined by comparisons to another criterion

While the judgement-based approach is what most people tend to think of when 'standard setting' is mentioned, it is by no means the only approach. Indeed, the Council of Europe (2009), in the manual for relating examinations to the CEFR, outlines a 5-step process, and recommends that the results of judgement-based approaches be validated. Validation can be done, for example, by comparing them with the results of another standard setting exercise or by comparing them with some other external criterion measuring the same ability and the same level.

Cambridge English has always sought to verify standards set by repeated validation activities. Standard setting studies relating *IELTS* to the Canadian Language Benchmarks have been conducted more than once – in 2005 and in 2013 – with consistent results being produced for the vast majority of levels considered (Lim 2014). As mentioned earlier, the exercise relating *IELTS* to the CEFR was triangulated with empirical data, where *IELTS* candidates also sat another exam with known cut points for several CEFR levels (*Cambridge English: Advanced*), and vice versa. They were further triangulated with other external studies relating *IELTS* and various other exams with the CEFR (Lim et al 2013).

Cambridge English has also noted that a good number of test takers sit more than one Cambridge English exam within days of each other, thus providing a natural way of confirming the relationship between scores obtained on different Cambridge English exams across levels. In preparation for the introduction of the Cambridge English Scale, groups of candidates were more formally invited to sit more than one exam (e.g. *Cambridge English: Preliminary* and *Cambridge English: First*).

In conducting studies where candidates take more than one test, some do not properly account for preparation and motivation effects. For example, someone who has taken a test is also asked to try the field test for another exam. The

person is likely to take the former more seriously than the latter, as they had reason to take it and indeed paid for the opportunity to take it, whereas the field test they do not prepare for and has no bearing for them. Thus, when people's results on the two tests are compared, a false picture emerges of the scores on one exam that relate to scores on the other exam.

In the studies conducted by Cambridge English, care has always been taken to employ counterbalanced samples. That is, a number of candidates who have already taken exam A are invited to also take exam B, and an equal number of candidates who have already taken exam B are invited to also take exam A. In this way, preparation and motivation effects are properly accounted for, and the resulting cut scores are accurate and dependable.

Conclusion

In this article, a brief overview has been presented of various ways by which standards in tests of language ability can be determined. Good testing practice requires that these methods be used in combination at different points of the assessment process, from initial test design to final score reporting. Standard setting is not always straightforward and involves many possible pitfalls. For this reason, it is also necessary to validate these standards by repeated validation activities, using a variety of approaches as much as possible, and ensuring that the resulting evidence is properly appreciated.

Conducting repeated studies using a composite of different methods is exactly what Cambridge English has done. This body of work, accumulated over many years, constitutes a strong evidence base for claims made about the relationship between different Cambridge English exams, between scores on Cambridge English exams and external frameworks, and about levels of performance appropriate for various contexts where language is used. These standards have been built into Cambridge English exams and processes, even as they are continuously revalidated, underpinning the test outcomes reported on the Cambridge English Scale.

References

- Buckendahl, C W, Foley, B P and Rodeck, E (2005) *Canadian Language Benchmarks International English Language Testing System Standard Setting Study*, Lincoln: Buros Institute.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*, Strasbourg: Council of Europe.
- Galaczi, E D, French, A, Hubbard, C and Green, A (2011) Developing assessment scales for large-scale speaking tests: A multiple-method approach, *Assessment in Education: Principles, Policy & Practice* 18 (3), 217–237.
- Jones, N (2009) A comparative approach to constructing a multilingual proficiency framework: Constraining the role of standard setting, in Figueras, N and Noijons, N (Eds) *Linking to the CEFR Levels: Research Perspectives*, Arnhem: CITO, 35–43.

- Jones, N and Saville, N (2009) European language policy: Assessment, learning and the CEFR, *Annual Review of Applied Linguistics* 29, 51–63.
- Lim, G S (2012) Developing and validating a mark scheme for writing, *Research Notes* 49, 6–11.
- Lim, G S (2014) *On standards and multilingual frameworks: Evidence from multiple standard setting exercises*, paper presented at the Language Testing Research Colloquium, Amsterdam, Netherlands, June 2014.

- Lim, G S, Geranpayeh A, Khalifa, H and Buckendahl C, W (2013) Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing* 13 (1), 32–49.
- North, B (2014) *The CEFR in Practice*, English Profile Studies volume 4, Cambridge: UCLES/Cambridge University Press.
- O'Neill, TR, Buckendahl, C W, Plake, B S and Taylor L (2007) Recommending a nursing-specific passing standard for the IELTS examination, *Language Assessment Quarterly* 4 (4), 295–317.

Reporting test scores and the Cambridge English Scale

ANDREW Q SOMERS RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

The processes through which a candidate's exam responses are translated into their final result are expanded on in the various articles in this issue. We will show that a wide range of processes are in place to ensure that each result is a fair indication of that candidate's language ability and is not influenced by irrelevant factors. However, the efforts of arriving at fair and reliable scores would ultimately be wasted if the reporting of the results was not also given sufficient attention. As suggested by Petersen, Kolen and Hoover (1989:222) the usefulness of the reported result 'depends on its fulfilling two important goals: facilitating meaningful inferences and minimizing misinterpretations and unwarranted inferences'. If results are presented in a confusing or inappropriate way, they can be misused and misunderstood, which in turn negates the efforts undertaken to ensure their accuracy and reliability in the first place.

Reporting test results is a context-dependent exercise which can be viewed from two perspectives: '1) that of the test developer who needs to report meaningful information on test results and to provide guidance for their use, and 2) that of the test user who needs to be able to interpret and use scores from tests' (Bachman 2004:294). Bachman (2004) also notes that as many test users may be far removed from the actual testing situation, the various stakeholders (i.e. test users) will vary in their familiarity with the purpose and content of a specific test. Therefore, there is no single universally accepted way of presenting test results. Test providers need to find the way which is the most appropriate for a given context. To establish an approach to reporting results in a clear and meaningful way requires an evaluation of the intended use of test results and the relevant stakeholders' needs. This article outlines a number of key concepts and considerations necessary for identifying the most suitable approach to reporting test results, and the creation of a meaningful reporting scale.

The use of test results will evolve over time, as new stakeholders emerge and requirements change. Similarly, the way test results are presented will change over time. The second half of this article will show how reporting scores for Cambridge English exams has evolved over time to reflect changing stakeholder requirements, culminating in the most

recent development – the launch of the Cambridge English Scale in January 2015.

Considerations in reporting results

A brief review of a range of English language tests across awarding bodies reveals the variety in approaches to solve the challenge of reporting test scores. The approaches differ from one awarding body to another, across products within an institution, test components (e.g. Reading, Writing) and also by level of reporting (overall performance or individual skills). There is yet further variation in the choice of format for their presentation (e.g. words, numbers, tables or graphics) as well as the extent of the explanation and supporting information (which may include verbal descriptors and statistical properties). A plethora of terms exists to describe what is reported, but perhaps the most common are marks, scores, grades and levels. For the purposes of this article, we will refer to these four terms throughout. Their meaning and differences are explained next.

Whilst 'result' is used as a general term for any information provided to explain exam performance, we refer to four more specific types of result throughout this article. These are defined as follows:

- Marks are the most basic measure of performance as awarded by the person or system marking the test, without any further manipulation. Thus, they may be dependent on the difficulty of the question, and/or the severity of an examiner. Whilst useful for comparing candidates on the same test, marks are not suitable for comparing performance in a wider context – comparing across different test versions, or from one exam to another.
- Scores are a standardised form of marks. They are transformed to take account of the variations that may affect marks. This means that a score should be a consistent measure of performance whenever and wherever it was issued. Whilst maintaining a degree of detail like marks, they are more generalised, facilitating comparison between different test versions. However, the degree of detail means it is hard to describe what each distinct score says about a candidates' language ability.

- Grades are essentially broad categories spanning a range of marks or scores. They are labels often represented by words (Pass/Fail) or letters (A, B, C, D and E) with an implied order or hierarchy. Like scores, they are standardised so that they are comparable across test versions. Being broader categories, they can be more easily described qualitatively and real-world meaning can be attached to them, such as, what a candidate at a certain grade can do in a language. However, their interpretation is still restricted to those who are familiar with the specific exam.
- Levels are similar to grades, but refer to a wider framework which is not restricted to a particular exam. This gives results a context beyond a specific test and allows comparison across different exams. The levels of the Common European Framework of Reference (CEFR) and the National Qualifications Framework (NQF) are two such examples.

Stakeholder requirements

When a new test is to be developed or an existing test should be revised, the planning phase begins with a definition of the intended context, test takers and use (Weir and Milanovic (Eds) 2003). The prospective stakeholders are consulted to determine the appropriate test specification to meet their needs. The same steps apply when developing an appropriate mechanism for reporting results.

There are numerous groups to whom we may be reporting test scores. Different tests will inevitably be aimed at different groups of test takers, though others may also make use of the results (e.g. employers and university admission officers). Each user will doubtless wish to draw their own information from test results, and for a number of different purposes. Bachman (2004) outlines a range of potential stakeholders, some of which are reproduced below with examples of the possible ways they may use test results:

- individual test takers may want to compare performance with their peers, measure themselves against a required standard, or receive feedback on their strengths and weaknesses in different areas or skills
- parents and guardians who are interested in seeing their child's progress over time
- teachers may be looking for feedback and diagnosis of performance to aid their teaching for individuals or whole classes, or to make decisions about setting or streaming students
- administrators, ministries and government organisations may be looking at curriculum development
- employers, universities and immigration authorities require certain minimum standards to consider people for jobs, access to courses or visas.

In view of the above, the key dimensions to reporting results are the following: the context, the level of detail and the grouping of results. As far as the context is concerned, some test users may have an interest confined to a specific set of questions and how many they got right, whilst others may have a more external focus and may need to relate results to a framework of reference or to performance on another test. The level of detail may also vary as some stakeholders may

seek detailed feedback on individual parts of a test, while others may be content with a broader overall assessment of proficiency. Finally, for the majority of users, it is their individual results that are of importance. However, a number of stakeholders are more interested in the performance of groups, classes, schools, regions and countries. It is conceivable that given the range of differing viewpoints, results for a given test may need to be presented in different ways for different purposes.

Criterion-referenced or norm-referenced results?

The ultimate outcome of a test is to provide a statement of a candidates' ability in the area of interest. However, no matter how much rigour and objectivity is brought into the process to ensure consistency, there is still an underlying judgement at its heart. That judgemental aspect leads us to comparisons because 'there is no absolute judgement. All judgements are comparisons of one thing with another' (Laming 2004). Extracting any meaning from a test result requires comparison: solely based on the mark of 56 awarded to a test taker, for example, we can draw no useful information about that test taker; the mark does not reveal how able that candidate is compared to others, and if that candidate is able to do certain things we require.

Two approaches have been developed to reporting test results that facilitate the comparisons required to interpret test results. The criterion-referenced approach compares test takers to a defined standard: has the candidate mastered certain skills that are deemed necessary or valuable? Alternatively a norm-referenced approach draws comparisons between individuals: who is the better candidate, how does this candidate compare against a defined population of other test takers? Different contexts may lead to the choice of one approach over the other. For example, a driving test requires a certain minimum standard for success – the focus is on being able to do certain tasks to a clearly defined, described level of competence; who the best drivers are or how many will ultimately pass the test is irrelevant. Conversely, a job interview or audition is more concerned with selecting the best candidate(s) from those available.

Such dichotomies are perhaps misleading. They should not be seen as mutually exclusive possibilities, but rather as representing the extremes of a continuum. More likely, they will go hand in hand: job interview panellists may look for the best candidate, but may still require the best candidate to possess some minimum competence in an area of interest. There is no right or wrong way to approach the issue, and reasoned decisions should be made to provide sufficiently clear and meaningful information.

Overall test results or individual skills?

Whilst summative assessments look backwards, identifying what a learner has achieved to date, formative assessments look forward to help learners learn and develop. As the design of tests for these two purposes will most likely differ, so will the way in which results are fed back. An overall result is useful for the backward-looking summary of achievement, but serves little by way of advice for future learning. Conversely, detailed diagnostic feedback on strengths and weaknesses of various subsections of a test will be more beneficial going forward, but may not be appropriate for the results of an

achievement test. Such fine-tuned reporting on specific skills and competences may not be easily measurable in the same way as a more general trait, particularly since it will only be based on a few items. Indeed, it may be preferable to adopt a more qualitative approach. Thus, the level at which information is provided, overall by skill or by specific areas of curriculum, will vary as will the manner in which it is reported.

Accordingly a number of decisions need to be taken around reporting results for tests which comprise a number of clearly defined subsections. A language test, for example, based on individually assessing each of the four skills would need to consider the following questions:

- Should performance in each skill be reported individually and/or as a single overall score?
- Where overall scores are provided should they be derived in a compensatory manner (a weak performance in one skill can be made up by strong performance elsewhere) or a non-compensatory manner (where a minimum criterion in each element is required in addition to any overall performance)?
- If scores from different sections are combined, how important is each to the overall performance and therefore how should they be weighted?

Discussions in relation to the questions above need to relate back to the test construct and purpose, as they influence the manner in which we produce results, the format and interpretations of results. Decisions should not be made in an 'either/or' manner, as an optimal solution in many scenarios is a balance of overall performance and performance in individual test elements, with some quantitative and qualitative supporting information.

Individual or group results?

Hitherto we have mostly focused on reporting results to individual stakeholders, whether to the candidate awaiting their result, or someone making a decision about that candidate (e.g. an employer or an immigration officer). Ministries of education are becoming increasingly more interested in the wider performance of a population, such as the performance of schools, regions and whole countries. Large-scale surveys such as the Programme for International Student Assessment (PISA) (Organisation for Economic Co-operation and Development 2012) and the European Survey on Language Competences (ESLC) (European Commission 2012) are examples of this.

Such aggregate information, on the face of it, would just seem to be a summary of individual performances answering questions for each group: how many candidates passed/failed, what was their average score? However, test providers also need to consider the psychometric properties of results to determine if certain score reporting is appropriate, e.g. what the reported results mean and how may they be used and combined. It may be that some reporting is not appropriate at the level of an individual candidate, e.g. it may not be possible to reliably quantify performance on a specific testing focus that only comprises three or four questions. However, it may be possible to provide such information as an overall summary for a suitably sized cohort.

Precision and Standard Error of Measurement (SEM)

The issues discussed above around measurement properties lead into a wider discussion of the nature of score scales and precision. As suggested by Flanagan (1951), score scale units 'should be of an order of magnitude most appropriate to express accuracy of measurement'. If too few distinct points are used, for example, measuring the size of your feet to the nearest metre, precision may be lost. Alternatively the use of too many score points can lead to misuse, with users attaching more significance to differences in scores than is appropriate (for example measuring the distance between cities to the nearest millimetre). At the very least it is important to provide information about the Standard Error of Measurement (SEM) to accompany explanations of numerical results.

The idea of SEM is a particularly challenging one. For statisticians, the concept of error is not problematic and the SEM actually provides guidance in interpreting scores. On the other hand, SEM is more likely ignored by test users, or misinterpreted. Indeed, the concept of there being any error in test scores may be concerning. In everyday measurement, we often take error for granted, but in high-stakes situations we take multiple measurements before acting, e.g. a carpenter will 'measure twice, cut once' to avoid expensive mistakes, or we may measure a sick child's temperature two or three times to be certain. In these situations we are aware that there is a natural degree of error in the measurement and so we take multiple readings before committing to a decision. Testing is similar in this regard, though it is clearly impractical to administer a test multiple times. The SEM enables us to describe the degree of variation we would expect with repeated measurement.

To illustrate the meaning of the SEM further we will use an example of a simple multiple-choice test of reading, comprising 10 questions each worth one mark. When the SEM was calculated for this test, it was found that it was one mark. On this test candidate A had six items correct whilst candidate B got seven correct (including the six that candidate A answered correctly). Which candidate is better? It would seem that candidate B performed better on the test, as they got more items right.

Given that the candidates differ by one mark, does the knowledge that the SEM is also one mark change our view of who is the better candidate? Assuming there were no mistakes in the marking, we can still be certain that candidate B performed better, as they got more correct responses. So, what does an SEM of one mark mean in this context? It relates to the inferences one draws about the two candidates on the underlying construct of the test – which candidate is more proficient in reading? There is no doubt that on that specific test, at that moment in time, Candidate B performed better. Their mark of 7 is our only evidence of their ability and that mark is only a snapshot, at a given time, and based on a limited sample of performance. We have assessed their reading ability with a limited number of questions, which only covers part of that construct. Saying which candidate is more able at reading is only an inference based on the available data, and is, thus, subject to some error.

If the two candidates took the same test again, their scores may be reversed. If they were to take a different test, we may get stronger evidence to support or reject our initial view that B is stronger. In other words, each time we administer

the test, whilst the explicit number of marks awarded is not in question, there is uncertainty as to how those marks best reflect a candidate's true ability. SEM attempts to quantify this uncertainty. In other words, candidate B is probably the better candidate given they got one more question right, but the SEM of one gives an indication of our confidence about that conclusion. If the difference in scores is greater than the SEM, it is highly likely that the higher scorer is the stronger candidate more generally, and will likely be the higher scorer on repeated testing. If the score difference is much smaller than the SEM, this is much less likely and for high-stakes decisions additional evidence should be considered.

One way of avoiding unwarranted inferences of who is more able on a limited sample of performance, is devising score scales where there are sufficiently few points, which are around one SEM apart. In this scenario, it is more than likely that the higher scoring of any two test takers is the more able one. Practically, this may not provide the degree of granularity some users want and they may need to separate people further. Test providers may then need to provide more precise scores, but with appropriate guidance about interpreting them – they will show who scored better on a given day, on a given test, and to a limited extent, they can also show who is likely to be the more able test taker in general terms.

The SEM is a property of a given test that contributes to the interpretation of its results. A number of factors can influence the SEM, some of which will guide the test design and development. Longer tests comprising more questions are likely to produce a smaller SEM. Similarly, tests which focus on a narrower range of ability or content domain also tend to have a smaller SEM. Accordingly a quick placement test with few items and a wide range of coverage will have a relatively high SEM and would typically only report results in broad categories such as a CEFR level. Conversely, a test providing an in-depth assessment at a specific level would likely have a lower SEM and report scores in more detail. However, lengthening a test may come at the expense of practicality and may cause fatigue in test takers, while narrowing the focus of the test may come at the expense of validity (too narrow a focus may under-represent the domain or ability of interest). Whilst consideration of the intended use and precision of scores is important in designing a test this needs to be balanced with several other key assessment principles.

Maintaining score scales

One of the challenges with score scales is their maintenance and ensuring that the information originally incorporated into them remains relevant over time. For instance, where norm-referenced approaches are used, the norm-group may become less relevant over time and may warrant updating. Alternatively, the content of a test may change considerably, and scores may need to change accordingly. This can be hugely problematic for communication. As Angoff (1962:32) comments: 'a scale has reasonable chance of being meaningful to a user if it does not change'. This is because some exams may have myriad stakeholders, a large and varied audience to whom changes need to be articulated. The change should be explained to all test users, and the new and the old scales should be related to each other, while avoiding confusion between the two. Therefore, rescaling a test is a process that should not be undertaken too frequently. When rescaling a

test is necessary, long-lasting scales should be designed, so that widespread familiarity with them can be established.

Choosing a score scale

Devising and maintaining a score scale is not a straightforward business if it is to be effective in providing useful information to the necessary range of users. We have discussed some of the factors that need consideration when establishing an appropriate approach, but, ultimately, deciding on a score scale is a matter of choice informed by context. The context in which results are used defines how to interpret a result – in relation to a standard, or to other individuals. With these considerations in mind we will now look at the application of some of these principles with Cambridge English exams. We will look at how the reporting approach has evolved over time to meet stakeholders' needs, and outline the most recent change, the introduction of the Cambridge English Scale in January 2015.

Evolution of Cambridge English exam results

Cambridge English exams have been administered for over 100 years, dating back to 1913. New exams have been added to the portfolio as required for specific purposes. Additionally, over the years, regular revisions to all exams have been made to reflect both developments in language assessment, and changes in their uptake, use and recognition. In conjunction with the modifications to the assessments themselves, the way in which results are reported has also evolved. The rest of this article presents the major developments to reporting results for Cambridge English exams in relation to the considerations discussed earlier.

Use of overall grades

Hawkey and Milanovic (2013) detail the history of Cambridge English exams from the very first administration of the *Certificate of Proficiency in English (CPE)* in 1913. The test was designed 'for Foreign Students who desire a satisfactory proof of their knowledge of the language with a view to teaching it in schools' (UCLES 1913:5). *CPE* was designed to certificate achievement of a particular standard, with a particular purpose in mind. Further details expand on the content of the exam across a number of areas. The *University Class List and Supplementary Tables*, provides a summary report on the performance of the exam – highlighting the candidates' strengths and weaknesses – an early sign of the potential for diagnostic feedback, and of skill profiling, although no such feedback was given directly to candidates. The final result was, nonetheless, a statement of achievement in the overall qualification aggregated across all skills. One final remark highlights what we would today call the criterion-referenced nature of the exam – 'the records of the day indicate that none of the three 1913 CPE candidates were actually awarded a CPE Certificate' (Hawkey and Milanovic 2013:22) – nobody reached the desired standard.

From the very first edition of CPE Cambridge English exams have reported grades to candidates. The nature of these has evolved through different labels over time (examples include Pass with Honours/Pass/Fail, Grade I,II,III and more recent A,B,C,D,E) but all represent the same notion: a fixed standard specific to that exam. The meaning of the result was very much limited to an understanding of the exam in question; in that respect the exams defined their own standard.

Alignment to the CEFR and NQF Levels

Following *CPE* in 1913, the *First Certificate in English (FCE)* was introduced in 1939 (initially as the Lower Certificate which was renamed into *FCE* in 1975), the *Preliminary English Test (PET)* in 1976, the *Certificate in Advanced English (CAE)* in 1991 (responding to the large difference in ability between *FCE* and *CPE* candidates) and the *Key English Test (KET)* in 1994. Over time this group of Cambridge English exams began to form a ladder of exams assessing across a range of different levels. Passing any one exam still had its own intrinsic meaning, but there was now a progression – passing one exam representing a higher standard than passing a lower one. The five levels thus established helped provide a basis (North 2014) for the Common European Framework of Reference (CEFR), first published in 2001 (Council of Europe 2001). Relating the Cambridge English exams to the CEFR, and reporting CEFR levels on certificates, further extended the meaning and interpretability of results, taking them beyond a specific exam and enabling comparisons between the various levels, and across other exams. Such comparisons have become increasingly necessary as the uptake of the exams has broadened, e.g. due to increased migration within Europe. The above-mentioned Cambridge English exams were also accredited within the National Qualifications Framework (NQF) in the UK as part of the government's Skills for Life strategy.

Skill-based reporting: Graphical profiles

Up until the development of the CEFR, Cambridge English exams only reported the overall result as a final grade. They were assessments of proficiency that explicitly tested all four skills, each of which was given equal weighting in a compensatory model, to provide an overall assessment of communicative language ability. In the early 2000s, the Graphical Profile was introduced in response to requests for more information on candidate performance. Whilst maintaining the overall emphasis on proficiency, a new system was introduced to report performance in the separate test components (Reading, Use of English, Listening, Writing and Speaking). At its inception, the profile gave a visual illustration of relative performance across the skills. However, rather than explicitly relating performance to defined criteria, for practical reasons it was designed to provide a norm-referenced profile compared to the rest of the population in a given test sitting, although the overall exam results were still criterion referenced.

The introduction of these profiles is an example of how results reporting evolved to satisfy a number of stakeholder requirements using a range of approaches. Criterion-based overall achievement was enhanced with an element of norm-referenced skill-based diagnostic feedback. A new scale of results (Exceptional, Good, Borderline and Weak descriptors) and introduction of skills-based reporting addressed a number of issues, but these profiles have also been subject to a number of enquiries. In response to feedback Cambridge English standardised the approach to generating the graphical profiles making use of the existing item response theory (IRT) based approach to test equating (see Elliott and Stevenson (this issue) for more information). Whilst using the same graphical image, the various thresholds were moved following a criterion-referenced approach.

More detailed information: Standardised scores

The next change to results reporting for Cambridge English exams was the introduction of standardised scores in 2008. A numeric score was introduced to supplement the overall grade, giving candidates the advantage of receiving more detailed reporting of their performance. It enabled them to see where in a particular level or grade they fall, or how close they are to the next level. Furthermore, the numeric scores facilitate comparisons with other exams (Lim, Geranpayeh, Khalifa and Buckendahl 2013), unlike wider categories such as grades. The standardised scores were chosen to span a range from zero to 100. This was partly due to practical considerations in terms of the implementation, but also to aid understanding. Although not percentage scores, the standardised scores were broadly comparable with the approximate percentages required to achieve the respective grades.

Enhanced certification

The introduction of enhanced certification in 2009 further cements the relationships between the level-based Cambridge English exams. Each of the Cambridge English examinations is targeted at a particular CEFR level (see Docherty and Corkill (this issue) for more detail on test construction) so that those who successfully pass the examination have demonstrated ability at the target CEFR level. However, the way the tests are constructed and candidates are assessed means that they also cover a wider range of ability around that target CEFR level; they enable candidates to demonstrate ability at the CEFR levels above and below the target level. This performance is reflected on the certificate which is issued to such candidates.

The assessment of the productive skills – writing and speaking – is carried out by comparing the sample of language produced by the candidate to a set of rating scales with performance descriptors which are linked to the CEFR as described by Lim (this issue). Unlike the productive skills, it is not possible to observe performance in the receptive skills – reading and listening – or use of English directly. These components are objectively marked with 'right/wrong' items, so a different approach is required. Items for these papers are designed to primarily elicit evidence of the candidates' skills at the targeted CEFR levels (B1 for *Cambridge English: Preliminary*). However some task types are appropriate for assessing at adjacent levels. For example, particularly strong performance on the harder items at the targeted level allows us to gather evidence of ability at a higher level. More information on this is provided in Elliott and Stevenson and Docherty and Corkill (both this issue).

For example, *Cambridge English: Preliminary* is targeted at the B1 level, and candidates who pass have achieved B1 level. Exceptional performance on *Cambridge English: Preliminary*, shown by achieving a Pass with Distinction, is recognised with a certificate showing that the candidate demonstrated ability at the higher B2 level. Candidates who do not pass, but demonstrate an appropriate level of ability, receive an A2 level certificate. Whilst the exams in question have always covered a range of performance they previously only stated achievement (or otherwise) of one particular level. Enhanced certification recognises the achievements of those candidates.

It is important to note that candidates who take, say, *Cambridge English: Preliminary* and are awarded a certificate at

CEFR Level B2 have demonstrated that they have the language knowledge and skills at this level, but this does not mean that they have demonstrated the full range of skills required for *Cambridge English: First* which covers a comprehensive range of the language, functions and cognitive processes at B2 level. In contrast, *Cambridge English: Preliminary* covers a narrower range since it is primarily targeted at B1 level. We can, therefore, have greater confidence that a candidate with a *Cambridge English: First* B2 certificate is able to perform the range of language activities associated with B2 level. In view of this, it is not recommended to 'jump' a level in a course of language study, as a result of achieving the upward certification level. A B2 certificate in *Cambridge English: Preliminary* is not a *Cambridge English: First* certificate, nor is it evidence of sufficient ability to follow a B2 course of study.

Thus, the way in which the examination results are used is important. Where a certain level of ability and evidence of ability in specific language functions is required it becomes important for decision makers to consider the examination in which a particular level was achieved. Such a decision should be based on the content of each exam and the range of language, functions and cognitive processes it covers as well as the overall level of achievement.

IELTS band scores

In 1989, the *International English Language Testing System* (*IELTS*) was launched. In contrast to the other Cambridge English exams discussed above, the results format in *IELTS* has remained virtually unchanged since its inception. Whilst other exams built on their existing output and extended and refined it to keep up with the needs of test users, *IELTS* was able to incorporate a range of requirements from the outset. Given the use of the test for immigration and work purposes, where performance in particular skills was often required, *IELTS* reported both overall and skill level results. Results are reported as Band Scores (ranging from 1 to 9), each accompanied by a descriptor, articulating the candidate's language skills at that level. The Band scores are comparable to grades in other exams, in that they represent a standardised result defined for that particular exam, but can also be viewed as levels given each is described in terms of the wider context of English language skills. The only change in the past 15 years was to introduce half bands in order to provide more detail as to a candidate's performance within a given level. By averaging the component Band Scores to yield the overall result, *IELTS* only has one single reporting metric ensuring a clear and simple format for end users. Indeed, the global standing of *IELTS* is such that other awarding bodies seek to align their exams to the *IELTS* scores to it. Such is the relative simplicity and understanding of *IELTS* results that they are widely known and recognised in their own right, which is why the results format has remained virtually unchanged since the start.

The Cambridge English Scale

In January 2015, the Cambridge English Scale was introduced to provide a single reporting scale for a range of exams. Each of the earlier developments, graphical profiles, standardised scores and enhanced certification in particular, paved the way for the introduction of this scale. This scale harmonises the

results across a range of products and facilitates comparisons between them. Furthermore, it simplifies the information provided to candidates whilst reinforcing the need for a range of exams to meet different needs.

It is worth bearing in mind the similarities (and differences) between measuring language proficiency and measurement in the physical sciences. Perhaps the closest analogy is the measurement of temperature. Bond and Fox (2007) provide an extensive description of the similarities between the two, from which we will draw when illustrating the Cambridge English Scale.

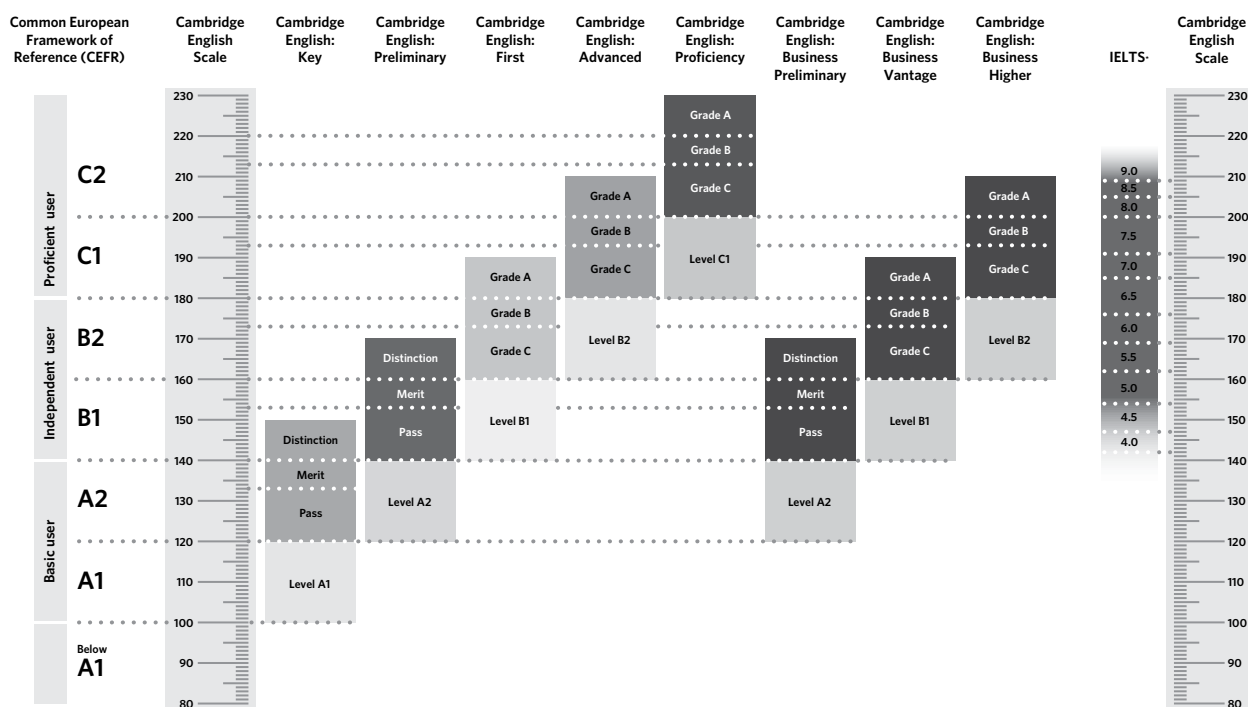
Firstly, temperature cannot be measured directly, but the effect of temperature on other materials can be observed, e.g. through expansion of alcohol or liquid in a glass thermometer, the bending of a bimetallic strip in a thermostat. In the same way, tests cannot directly measure proficiency in isolation. Language ability is complex and multidimensional where tests can only take a snapshot of performance on a sample of language tasks. One single test cannot adequately measure all aspects of language ability validly and reliably across the whole spectrum of performance. Thus the context is a key aspect meaning that the measurement of temperature or proficiency is dependent on the choice of instrument.

Thermometers have a defined range at which they will operate reliably and are used for different purposes: checking a baby's temperature, controlling the temperature in your house, or setting the oven to bake a cake. Each has a different range, and, accordingly, different degree of precision depending on its purpose. In the same way, tests can cover wider or narrower ranges of language ability with varying degrees of precision. Docherty and Corkill (this issue) discuss the way tests are constructed to reflect this and Walczak (this issue) discusses how adaptive testing gives some benefit in this area. This is a key feature of the range of Cambridge English exams across a number of levels and contexts. We provide exams to meet different needs, just as we would use a range of thermometers for different applications.

The development of a range of exams for varying needs led to a number of exams each reporting results on their own scales. Similarly, different types of thermometers measure temperature by measuring other physical properties (the volume of liquid in a tube, the degree to which a bimetallic strip bends, or the electrical resistance of a probe). In both cases we need a way of comparing across instruments. In this regard the temperature analogy is far simpler – there is one single dimension of interest: the temperature or the amount of energy in a system. Each instrument can be calibrated to measure against a single temperature scale. With language testing it is not so simple – construction of a single underlying unidimensional scale is not possible. However, we can relate the measurement from each test to a common reference scale to facilitate understanding and comparisons.

This forms the basis of the Cambridge English Scale: a single reference scale to which each of our exams can be related (see Figure 1, at the very right). The scale provides a means of comparing performance between tests. Each product reports on an appropriate range of the scale, over which it is designed to assess ability in the same way as a thermometer has a defined range of use (see Figure 1 for an overall view of Cambridge English exams in relation to one another and to the Cambridge English Scale).

Figure 1: Alignment of exams with the Cambridge English Scale



*IELTS is mapped to, but will not be reported on the Cambridge English Scale

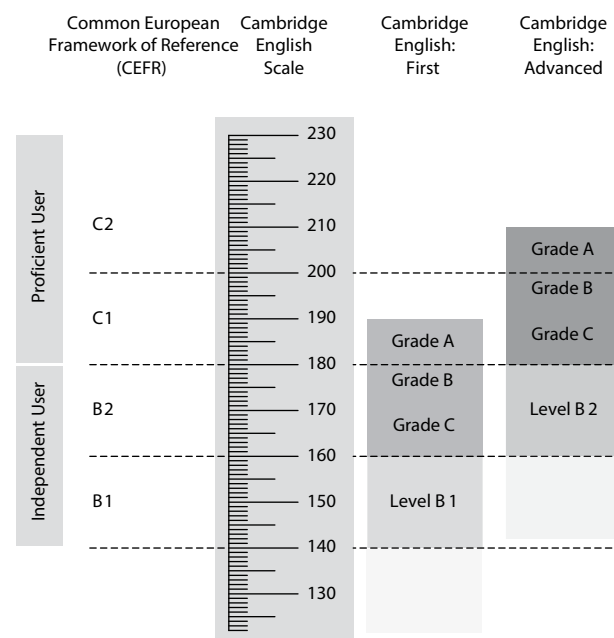
It is possible for thermometers to measure beyond the defined range to a degree – the liquid in a glass tube can keep expanding above the printed scale as the temperature increases. However, due to design and physical properties of a thermometer, such expansion can no longer be reliably related to the overall scale. Tests are similar in that the range of marks on a test can span the full range of zero to 100%. However, performance at the extremes cannot reliably be related to performance on other tests, or to a wider general reference scale like the CEFR.

For example, if a test (like *Cambridge English: First*) is designed to determine if a candidate is at B2 level or not we cannot be sure how able someone who scores 100% is. We know they are at least B2 level but they could be much stronger than that. Similarly someone who gets every question wrong is definitely not B2 level – but how weak are they? Measuring the language ability of both these candidates is beyond the scope of that instrument. Even though both can get a score on that specific test we can't reliably interpret this further. We know that the candidate getting full marks has a higher ability than our test is designed to measure, but not by how much.

As scores get closer to the extremes, it becomes harder to reliably ascertain language ability. This is why the Cambridge English Scale scores cover a defined range which does not cover every possible mark achieved in the exam. The maximum score that candidates can achieve on the exam will be less than 100% of the marks available. Similarly candidates with very few marks in the exam will receive a 'not reported' result. Although we can sum up how many marks they received, we cannot relate such low performance to the CEFR or to other exams. In such cases, a different exam would be a better instrument for assessing their level, which is why we have different exams set at different ability levels. Figure 2

highlights how the Cambridge English exams, *Cambridge English: First* and *Cambridge English: Advanced* are aligned with each other and the Cambridge English Scale.

Figure 2: Alignment of Cambridge English: First, Cambridge English: Advanced and the Cambridge English Scale



Cambridge English: First is targeted at the B2 level, such that candidates who achieve a passing grade (Grade C or higher) are certificated at B2 level, and will achieve a Cambridge English score of 160 or higher. Candidates who do particularly well, and demonstrate ability at the C1 level receive an A Grade and a score of 180 or higher. The maximum score awarded in *Cambridge English: First* will be 190. Candidates

may well be of a higher level ability than this, but *Cambridge English: First* is not designed to measure higher than this. At this level *Cambridge English: Advanced* would be more appropriate to cover the full range of C1 level performance – where the highest score achievable is 210, representing the lower end of C2 performance.

If we look at the lower end of *Cambridge English: Advanced*, those that do not reach C1 level may still receive a result. If their performance is sufficiently high to demonstrate B2 level performance then they will receive a general Cambridge English certificate (it is not a *Cambridge English: Advanced* certificate as they have not passed the exam) showing their B2 level achievement included a score from 160 to 179. If performance is weaker than this, no grade or certificate is awarded, although they may receive a score if their performance is close to B2 level. Below a minimum performance (a score of 142) candidates will not receive a score at all. This is the point at which *Cambridge English: Advanced* is not able to reliably or validly assess performance, nor relate this performance to the overall scale. This is where candidates should take a lower level exam, *Cambridge English: First* or even *Cambridge English: Preliminary*.

Just as each exam covers its own defined range, so too does the overall suite of products. The suite focuses on the range of measurement required to assess those learners taking these exams, and also covers the CEFR levels. Hence there is no zero point shown in the scale: we are not trying to measure or define zero language ability much like absolute zero in temperature cannot physically be achieved or measured. The important features of the scale are the meanings of the scores – how they relate to real-world performances, to a performance on other exams, or particular skills, cognitive processes and Can Do statements.

The use of this scale facilitates the determination of requirements for immigration and university entrance, for example. A single defined level of achievement can be stated across a range of tests, rather than having different criteria for each recognised exam. The organisation should also evaluate if each test measuring at that level has the desired content and construct to be appropriate for their needs. For example it may be appropriate to require a C1 level, or Cambridge English score of 180, but only accept this in *Cambridge English: Advanced* or *Cambridge English: Proficiency* but not in *Cambridge English: First*. Similar to the introduction of enhanced certification previously – the level of achievement is comparable, but is measured in different ways and this whole picture must be considered when making decisions. Furthermore, these scores can readily be aligned with *IELTS* scores, and thereby recognise an institution's existing requirements. The decision was taken not to report *IELTS* on this same scale, but to maintain its current scale aligned to the Cambridge English Scale. This is testament to the widely recognised and understood *IELTS* scale – it has become embedded around the world as a standard people implicitly understand.

The second advantage of this new scale is the clarification of the component level reporting. The old graphical profiles have been updated with a new image, and the reporting scale has been brought into line with the overall scale. Thus, candidates now have a clearer criterion-based report on their performance in the various skills aligned to a single reference

scale, allowing them to better understand their strengths and weaknesses. Furthermore, such information allows institutions to specify minimum scores in individual skills if desired. It should be noted that achievement in the overall qualification is unchanged – candidates must achieve an overall level across all skills, with no minimum requirements in any one skill – weak performance can still be compensated by strength in other areas, but we are allowing additional inferences to be drawn about their level – independent of their achievement of the qualification. As always it is necessary to evaluate the suitability of tests for these purposes based on the content and measurement properties they show.

Thus we can see how the Cambridge English Scale brings together several years of developments and enhancements into one concept for reporting exam results across our suite. The scale facilitates the use and interpretation of our exam results for various stakeholders and was established through a number of methods. Firstly the alignment between products is built into our test development processes. As Lim (this issue) shows, the assessment of productive skills implicitly relates performance across the levels through the application of the aligned mark schemes which share appropriate descriptors between adjacent levels. Although it is not appropriate to make one single scale that can span and be applied across the whole range of performance for each exam, a ladder of overlapping scales can be built for adjacent exams which enable them to relate back to a common reference framework.

In a similar way, the assessment of the receptive skills is also based around a ladder. The IRT methods outlined in Elliott and Stevenson (this issue) ensure consistency of decisions between sessions of the same test. However, the way in which the items are calibrated in this model also allows linking adjacent exams together – defining cut-offs that relate to a number of levels within each test, demonstrably linked to those in other tests.

It is still not sufficient to have this in-built alignment and assume that everything else falls into place. First we must ensure such alignments are robust and maintained. The various processes outlined throughout this issue explain how that consistency is maintained year on year. Furthermore, the final article, by Beresford-Knox, outlines how the various Quality Management processes and mechanisms ensure that such processes are regularly monitored and improved. Even so, the in-built alignment makes certain assumptions in the way it is applied, and we must therefore also carry out *post hoc* analysis to provide the empirical data to show the outcomes are in accord with the design. Thus an ongoing programme of alignment studies based on candidates taking adjacent exams is in place to monitor this.

Conclusion

A number of issues and considerations in defining the most appropriate approach to reporting test scores in a given situation were discussed in this paper. Decisions on score reporting should always be made in view of the use and purpose of an examination, and in view of who the key stakeholders are. Determining how to best balance those considerations to ensure that a clear, well-communicated mechanism is in place and minimise the likelihood of any

misuse or misinterpretation is a matter of choice. The Cambridge English Scale is the latest development that brings reporting of results across a range of exams into a single unified approach. This approach provides greater clarity in reporting results, both overall and for individual skills using a single scale clearly aligned with meaningful descriptors of performance. Furthermore, by reporting all exams against the same scale, the relative performance on different levels of exams is made more explicit, both facilitating comparisons between them, and enabling the tracking of learning and progression over time. This single scale also simplifies requirements that may be set by recognising institutions such as universities or immigration bodies – one score can be established which can be used across all recognised qualifications. The use of one scale showing a number of different exams aligned to it also reiterates that whilst the level of performance may be comparable, the context and manner in which it is assessed varies depending on the specific context required – there is no one-size-fits-all product that can assess English language proficiency across the full range of proficiency and in all contexts and situations.

References

- Angoff, W H (1962) Scales with non-meaningful origins and units of measurement, *Educational and Psychological Measurement* 22 (1), 27–34.
- Bachman, L F (2004) *Statistical Analyses for Language Assessment*, Cambridge: Cambridge University Press.
- Bond, T G and Fox, C M (2007) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, London: Routledge, 2nd edition.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- European Commission (2012) *First European Survey on Language Competences*, available online: ec.europa.eu/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf
- Flanagan, J C (1951) *Units, Scores, and Norms: Educational Measurement*, Washington DC: American Council on Education.
- Hawkey, R and Milanovic, M (2013) *Cambridge English Exams: The First Hundred Years: A History of English Language Assessment from the University of Cambridge 1913–2013*, Studies in Language Testing volume 38, Cambridge: UCLES/Cambridge University Press.
- Jones, N (2014) *Multilingual Frameworks: The Construction and Use of Multilingual Proficiency Frameworks*, Studies in Language Testing Volume 40, Cambridge: UCLES/Cambridge University Press.
- Laming, D (2004) *Human Judgement: The Eye of the Beholder*, London: Thomson.
- Lim, G S, Geranpayeh, A, Khalifa, H and Buckendahl, CW (2013) Standard setting to an international reference framework: Implications for theory and practice, *International Journal of Testing* 13 (1), 32–49.
- North, B (2014) *The CEFR in Practice*, English Profile Studies volume 4, Cambridge: UCLES/Cambridge University Press.
- Organisation for Economic Co-operation and Development (2012) *Programme for International Student Assessment (PISA)*, available online: www.oecd.org/pisa/
- Petersen, N S, Kolen, M J and Hoover, H D (1989) Scaling, norming, and equating, *Educational Measurement*, New York: Macmillan, 3rd edition.
- Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.

A modern perspective on statistical malpractice detection

CHRIS BELL RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

Within Cambridge English, it is the responsibility of the Research and Validation department to research and apply appropriate statistical malpractice detection techniques for use on Cambridge English exams. This article provides a brief summary of current trends in statistical malpractice detection from a Cambridge English perspective. It begins with an overview of the current context of malpractice in high-stakes examinations – how malpractice is occurring and why it is an important issue. This is followed by a discussion of several techniques for detecting malpractice statistically, assessing the advantages and disadvantages of each and providing insight into how theoretical concepts for detecting malpractice can be practically developed into effective and rigorous processes. The discussed detection methods have been created in entirety by members of the Research and Validation team or adapted from the available studies. The

article concludes with a short discussion of how modern technology can aid statistical malpractice detection.

For the purpose of this article, malpractice can be broadly defined as any action which allows a taker or takers of an examination to gain an unfair advantage. Malpractice and cheating can be used interchangeably in this context.

The modern context of malpractice

A quick internet search uncovers pens able to scan entire documents, erasers with hidden electronic screens and miniscule earpieces that can fit inside the human ear completely undetected. However, rather than being props for a spy film, these instead are all openly advertised online as 'examination equipment'. Innovations in technology have created new challenges in combating malpractice,

and as malpractice prevention through invigilation becomes an increasingly complex proposition, greater emphasis than ever is placed on detecting malpractice through statistical techniques, processed once exams have been completed.

Research regarding the extent of, and attitudes towards, malpractice can make for equally worrying reading. Cizek (1999:3) writes that 'nearly every research report on cheating . . . has concluded that cheating is rampant'. He reports that 40% of US sixth graders copy and that about 60% of undergraduates do so at some point during their college careers. Fifteen years on from Cizek's research, the easier accessibility of online and electronic resources means these figures will have significantly increased (Geranpayeh 2013:982). Indeed, a recent study regarding approaches and attitudes towards malpractice reports that 35% of US teens surveyed admitted to cheating with mobile phones, and more than half admitted to using the internet to cheat (Common Sense Media 2009).

In many ways this attitude towards malpractice reflects the high-stakes nature of modern examinations. Performance on language tests can be crucial to immigration, employment and further education prospects. This is particularly true of exams with government recognition such as the *International English Language Testing System (IELTS)*. Many takers will be aware that their prospects and hopes for the future are on the line, and this can create incredible pressure to succeed. It could be argued that malpractice is simply an unavoidable consequence of this pressure being placed, often on young shoulders. Cambridge English is obliged by the Office of Qualifications and Examinations Regulation (Ofqual) to have measures in place to manage the prevention and detection of malpractice, reflecting the significance of modern English-language qualifications.

Another explanation of the amenable attitude many seem to have towards malpractice may be the oft-heard argument that cheating is a 'victimless crime'. However, in reality, the impact of malpractice can be widespread and significant. While cheating on an examination may improve the cheaters' future prospects, it will inevitably deny these same opportunities to others, as immigration, education and employment opportunities tend to be limited. Furthermore, qualifications can lose credibility and importance if issues of malpractice are believed to be a widespread and significant problem, even for candidates who have obtained their results fairly (Geranpayeh 2013:980).

Malpractice can also have a deeper underlying impact on the validity and reliability of tests. Cambridge English examinations and corresponding performance thresholds are constructed based on detailed statistical analysis of candidate performance, both past and present. Any undetected cases of malpractice therefore represent a problem of data quality; the candidate's performance on the test is neither a fair reflection of their own ability, nor of the true difficulty of the test. Invalid data can bias and distort the results of statistical analysis, which can potentially lead to performance indicators being set inaccurately and to future tests being constructed to unsuitable levels of difficulty.

Detecting malpractice through statistical analysis

Given the modern, innovative approaches towards malpractice, and a wide-ranging impact of malpractice, there is an increased focus on the use of statistical analysis as a tool to combat the problem. It is important to note that statistical malpractice detection should only be used in conjunction with other, observation-based methods and never instead of them. Statistical malpractice analysis cannot prove anything on its own. The core idea of statistical malpractice detection is to look for patterns of unusual performance and assess the likelihood of any given pattern of results or scores being found by chance. In doing so the possibility that unusual performance can be completely genuine must be acknowledged. For example, some cohorts or regions will typically perform better in certain skills than others, and so compared to a general population their performance in a strong skill may appear unusual whilst being totally genuine. Similarly, a weak performance on an individual paper can exacerbate a slightly high score on one paper and make it seem more suspicious than it actually is (Somers, Calver and Bell 2009:2). It is vital that additional evidence such as seating plans, handwriting analysis and invigilator reports be gathered and assessed for any malpractice case identified statistically. Furthermore, the best, most effective and most reliable method for controlling malpractice remains prevention through invigilation. Modern preventative methods have included banning gadgets in test centres, installing CCTV, and introducing identification measures such as passport checks and biometrics before candidates are allowed to enter examination halls (Geranpayeh 2013:991).

The relationship between these observational measures and statistical malpractice detection techniques is multidirectional. The statistical techniques outlined later in this article are considered to be standardised, routine analyses which are performed on all Cambridge English candidates within an examination sitting. Candidates or centres identified from these techniques as possible cases of malpractice are then referred to a committee who will assess them in conjunction with any additional evidence gained through observational methods. However, malpractice detection in Cambridge English does not necessarily begin with statistical analyses. Often, a concern will have been raised at centre level, for example in invigilation reports, which will then be referred to the Research and Validation department to see if there is any statistical evidence that can be provided as further support. In these cases, the Research and Validation department will have a greater idea of the nature of the suspected malpractice and specialised analyses, tailored to focus on the specific areas of concern, can be performed. The Cambridge English approach to malpractice can, therefore, be seen as an iterative process; observation on the day of the exam can be used both as evidence in itself and to tighten the scope of post-exam statistical techniques. Committee decisions on what action to take based on all this evidence can further inform centre training, inspection and approval measures, important themselves in preventing malpractice in future exams.

The first stage of designing relevant malpractice detection techniques is to understand exactly how malpractice can

occur. Some of the most common malpractice techniques seen in examinations are:

- obtaining unauthorised access to secure exam materials
- educators or students engaging others to take an exam on a student's behalf
- students receiving unauthorised assistance from teachers or other students during an examination (which can encompass both copying and collusion)
- students accessing non-allowable resources (notes, textbooks, the internet) during an examination
- altering of exam scores post-examination.

(The Council of Chief State School Officers 2013)

It should be noted that the above examples of test security breaches are generalised techniques which can occur irrespective of delivery (paper-based or computer-based) or test type (linear or adaptive). Other types of malpractice, specific to one delivery method or type of exam, such as computer hacking of computer-based and/or adaptive tests, are also possible.

When considering how the aforementioned types of malpractice manifest themselves, certain patterns can appear when comparing candidate performance, such as:

- similarity of response patterns between two or more candidates
- likelihood of certain responses and scores given the overall ability of the candidate
- unusual candidate performance on one test section compared to performance on the other sections
- unusual performance of a group (such as a centre) compared to historic trends
- unusual performance of repeater candidates on one test date compared to their previous performance.

Each will be examined next in more detail, but it is worth noting that the outlined statistical detection techniques are complementary; in most cases several checks are combined. The following sections do not cover the full range of statistical detection techniques used by Cambridge English Language Assessment, but present the most widely used ones. For reasons of security and commercial sensitivity, it is necessary to exclude some details pertaining to the specific analyses performed by the Research and Validation department from the following discussion.

Similarity of response patterns between two or more candidates

Most instances of cheating take place through copying responses or collusion of some kind (Geranpayeh 2013:994). An inevitable consequence of this is that candidates will share a number of identical responses to test items. Analysis of identical responses between candidates can uncover unusual patterns between two or more candidates. This is an effective detection method as it is able to cover a wide range of malpractice activities. It can also work on both multiple-choice items and items requiring a short text response. However, there are limitations to this method. Analysis of similar response patterns is, by its nature, more effective at identifying patterns of similar incorrect responses than similar

correct responses. Candidates with suspicious patterns of incorrect responses are unlikely to have performed strongly overall on the test. In view of this, it can be argued that while it is an effective method at identifying breaches of test security, it is less effective in identifying breaches of test security which have allowed candidates to pass an exam they would have otherwise failed. Furthermore, there can be legitimate reasons why candidates have common patterns of incorrect responses – many items are intentionally written to include strong distractor elements and candidates are more likely to have incorrect responses for more difficult questions.

The method evaluated next can tackle some of the above-mentioned limitations.

Likelihood of certain responses and scores given the overall ability of the candidate

The difference between this and the previous method is subtle, but significant. The method described here helps identify unusual performance by examining the overall pattern of correct/incorrect responses to calculate the probability of certain responses. As such, it accounts both for the relative difficulties of items and the varying abilities of the candidates sitting the test, unlike the analysis of identical responses. A candidate who has performed strongly on certain items which other candidates of a similar ability have struggled on is one example of the type of pattern that could be identified from this analysis. It is, therefore, considered an effective method for indicating cases of copying and collusion.

Unusual candidate performance on one test section compared to performance on the other sections

Certain breaches of test security (such as unauthorised access to a single test paper) may impact one test section, but not the others in the same examination. As such, a comparison of candidate performance in one section with concurrent performance in others within the same administration can be an effective method of identifying malpractice. This is particularly true when comparing receptive skills (reading, use of English and listening papers) with productive skills (writing and speaking). For example, the leaking of an entire test is likely to advantage candidates to a greater extent where they can simply memorise the appropriate multiple-choice answers than for example, a Speaking examination. This is because in a Speaking examination, even if the questions are known beforehand a certain level of speaking ability will still be required to be able to engage in discourse with examiners. Essentially, productive language skills are harder to fake because they are directly observed.

While it is possible to cheat across all type of examination paper and skills, the most effective method of cheating will vary depending on the type of exam and how it is marked. One exception to this generalisation would be the use of imposters sitting all exam papers on a candidate's behalf, which might result in a cheating candidate having a flat profile (i.e. similar performance across all skills) and would not be flagged from this type of analysis. Again, this highlights the importance of using statistical detection techniques in conjunction with preventative measures such as biometric photo identification.

Unusual performance of a group (such as a centre) compared to historic trends

Overall performance of centres and cohorts tends to be fairly consistent over long time periods and therefore, it is possible to statistically identify sharp peaks and drops in performance. This may suggest a range of malpractice activities such as widespread collusion, centre assistance and unauthorised access to test materials. It is clear, however, that there may be genuine explanations for improved performance, including changing demographics and improved teaching and educational programmes. Therefore, when deciding on whether to take action at a centre level, it is important to consider other factors, too, such as the outcomes of recent centre inspections. There are other practical limitations with the analysis of centre performance. Identifying suspicious centres is heavily dependent on having a significant historical record of the centres' performance; a centre which has only taken a few administrations of an exam will not be flagged from this approach because there will be no baseline to show whether the centre's current performance is actually atypical or not. One possible solution to this would be to compare the centre's current performance to the historical performance of *all* candidates across all centres, rather than to the historical performance of that centre. This would further increase the chance of identifying suspicious centres, but the downside is that it may simply detect genuinely strong-performing centres. As such, it is not a particularly discerning method in and of itself.

Unusual performance of repeater candidates on one test date compared to their previous performance

A comparison of individual candidate performances with their historical performances in the same exam can identify candidates with high score gains over short time periods. This can be used to identify almost all forms of individual candidate malpractice. In particular, this is one of the few statistical detection techniques which can highlight instances of educators or students engaging others to take an exam on a student's behalf. This technique is only useful when candidates have taken the same test more than once. This tends to be more common with exams such as *IELTS*, which is frequently used by governments for immigration purposes.

Applying the theory – thresholds and practical considerations

A recurring problem common to all types of statistical malpractice detection is where to draw the line to separate unusual behaviour that is a result of chance, with those cases that constitute actual malpractice. In reality, it is impossible to completely eliminate the highlighting of false positives, but the establishment of a defined threshold or tolerance above which behaviour is thought to be unusual can help to minimise this. Thresholds can be derived from a purely statistical perspective by highlighting anything which deviates a certain amount away from average performance, but are most effective and rigorous when a review of previously identified cases is taken into account. For example, an analysis identifying candidates at a statistically significant level of 99.9% will identify any cases in which there is a 0.1% or less chance of the particular behaviour having occurred by chance. This may seem a very robust threshold to set, but on an exam sat by 100,000 candidates (far from unheard of for Cambridge English; for

example, 120,000 Portuguese students sat a recent *Cambridge English: Key for Schools* administration), can highlight 100 candidates whose performances are simply a result of chance. Further analysis, in terms of analysis of handwriting, seating plans and invigilation reports, may indicate that this was actually too broad a threshold; it may be that additional evidence was only uncovered for the 10 most extreme outliers (although in reality it is very rare for things to work as neatly as this). In that case there is a certain level of justification to raise the threshold to a level of 99.99%; now only one in 10,000 candidates will be flagged. As with this example, initial thresholds often only act as guides in the absence of any other information.

There are also practical considerations to take into account; resources to investigate malpractice are not endless, and full investigations can be time-consuming. Most research projects will use a 90%, 95% or 99% level of statistical significance to justify the validity of their findings. In relation to Cambridge English exams, statistical significance of 90% could potentially highlight hundreds of thousands of cases a year. Investigating each and every case would create huge delays in the release of results, which would be extremely unpopular with candidates and schools, especially in the context of high-stakes examinations. It would also likely prove fruitless for the vast majority of highlighted cases. Setting an appropriate threshold can therefore be seen as a balancing act; taking into account the statistics underpinning the malpractice detection method, feedback from previously flagged cases and practical implications. It is only by considering all of these factors that rigorous, well-defined and effective tolerances can be set.

The modern context of malpractice detection

While new technology can produce new challenges in combating malpractice, it does also offer certain advantages. Investments in computing infrastructure and software at Cambridge English have hugely increased the potential for data processing and analysis. Key to this is the idea of 'big data' – data sets so large that they become difficult to process with traditional applications. The need for more powerful solutions is clear when one considers that for much of its existence Microsoft Excel had a limit of 65,000 rows of data, and even some single administrations of Cambridge English examinations are sat by more candidates than this. By analysing huge amounts of data across many years of examinations, trends, patterns and anomalies in performance become clearer and can be calculated with greater precision and rigour. In general, malpractice detection analyses (including the derivation of thresholds and tolerances) are considered to be most effective when the amount of information available is as large as possible. However, this process can often be extremely computer intensive and could even be considered unfeasible unless underpinned by a powerful computing infrastructure which can effectively and reliably process, transfer, store and protect large amounts of data.

An interesting consideration to make at this point is that the size and scope of Cambridge English is fundamental

to its capability to detect malpractice; large amounts of information are able to be analysed only because large amounts of information are available in the first place. A new organisation launching today for example, would not be able to draw on the same level of historical data in order to determine what would actually constitute statistically significant unusual performance. There are however, practical concerns which can negatively affect the size of data being analysed. A desire for more frequent examinations, though understandable from a business and customer perspective, lowers the numbers of candidates taking each individual sitting and as such restricts the capability of statistical methods to detect malpractice. The aim to release results as quickly as possible to candidates can also impact this as it may be that malpractice analysis needs to be performed on some candidates in order to release their results before data from others taking the same test has been fully processed and made available for analysis, again decreasing the size of the population being compared.

The adoption of specialist statistical software that is both powerful and flexible also facilitates effective and efficient statistical malpractice detections. It allows complex statistical analyses to be developed and automated (beyond the scope and capability of more basic software) in-house by trained staff with the appropriate knowledge of statistics and language testing, which can then be adapted where necessary to fit a specific scenario. This is often preferable to the use of proprietary (or closed-source) software developed either externally or by a specialist information technology department, as the restricted nature of this software limits the capacity for the type of modification and flexibility required for an in-depth malpractice investigation, often within short time-spans. It is also possible now to program analysis routines to run automatically at a certain point in the day without the need for further human involvement, which can be extremely useful when tackling periods of busy activity or tight deadlines for the release of results.

Conclusion

By examining the modern context of both malpractice and malpractice detection, one thing that becomes clear is that the current situation will not represent the endpoint of the malpractice battle. At times the field resembles an arms war – for every new detection technique developed, a new method of evading detection materialises. New challenges will continue to appear, and, at the same time, new detection methods will continue to be developed or adapted from the existing research and old ones will be refined or discarded if no longer considered useful. Geranpayeh (2013) elaborates on some other statistical techniques left unexplored by this article. For example, one potential area of increased focus is the growing popularity of computer-based delivery models. As well as opening up new risks such as the possibility of computer hacking and remote assistance, computer-based exams allow new types of data to be recorded; such as the time taken to answer specific questions and the pattern of individual keystrokes, which in turn provide the opportunity for new methods of detection to be developed. Therefore, it seems clear that in the foreseeable future, at least, the malpractice arms war will continue unabated.

References

- Cizek, G J (1999) *Cheating on Tests: How to Do It, Detect It, and Prevent It*, Mahwah: Lawrence Erlbaum Associates.
- Common Sense Media (2009) *35% of Teens Admit to Using Cell Phones to Cheat*, available online: www.common sense media.org/about-us/news/press-releases/35-of-teens-admit-to-using-cell-phones-to-cheat
- The Council of Chief State School Officers (2013) *TILSA Test Security Guidebook*, available online: www.ccsso.org/Documents/TILSA Test Security Guidebook.pdf
- Geranpayeh, A (2013) Detecting plagiarism and cheating, in Kunnan, A J (Ed) *The Companion to Language Assessment*, London: Wiley Blackwell, 980–993.
- Somers, A, Calver, L and Bell, C (2009) *Statistical Malpractice Detection*, Cambridge: Cambridge ESOL internal report.

Computer-adaptive testing

AGNIESZKA WALCZAK RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

Computer technology has been present in educational testing since the 1970s. Initially, computers were only used for scoring tests and processing score reports. Since the 1980s computers have been employed as delivery platforms to administer exams. It was only in the early 2000s when the new technology became more widespread in educational assessment, mostly due to the increased availability of inexpensive and high-powered computing (Thompson and Weiss 2011, Thurlow, Lazarus, Albus and Hodgeson

2010, Wainer 2000). Nowadays, they are widely used in educational and language assessment.

Computer-based testing is a broad category which spans *linear* and *adaptive testing* (e.g. Suvarov and Hegelheimer 2014). In a linear test, the same number of test questions in the same order is administered to all test takers. In this scenario, a computer-based linear test is similar to a standard paper-based test, but the medium of administration differs. In comparison, in computer-adaptive tests (CAT) candidates receive different test questions; in such a test each

item is selected by the computer based on the test taker's performance on the previous task. Here, the test adapts to the response pattern of the candidates and hence to their ability level (Suvarov and Hegelheimer 2014). Computer-adaptive testing offers several advantages to examination boards and test takers, ranging from shortening the time for test delivery and immediate score reporting to candidates.

At Cambridge English Language Assessment, computer-based testing has a long tradition. Computer-based linear tests started in 2002 with the development of computer-based *IELTS*. This was followed by delivering other Cambridge English tests on the same platform from 2005 onwards. Most Cambridge English linear exams are now offered on both the traditional paper-based platform as well as on computers. The development of the first computer-adaptive test at Cambridge English dates back even earlier – in the early 1990s, Cambridge English developed a CAT test for Singapore Telecom and a Linguaskill test (Jones 2014) to be used by an employment agency. Subsequently, other CAT tests were created – among others, CommuniCAT (winner of the 2000 European Academic Software Award), computer-based *BULATS* Online and the Cambridge English Placement Test (Jones 2014). At Cambridge English, linear computer-based tests are employed to measure English language ability in four skills – reading, listening, writing and speaking – while CAT tests assess reading and listening ability.

The focus of this article is on the process of arriving at meaningful test scores for CAT tests offered by Cambridge English, such as *BULATS* Online or the *Cambridge English Placement Test*. It will be shown how such tests differ from standard paper-based linear tests with regard to test administration and test production. Computer-based linear tests will not be discussed here as test administration and production largely resemble the processes in place for paper-based tests.

Test production – test specifications and item bank

The process of test production of computer-adaptive tests closely resembles the process in place for the linear test. However, there are a few differences that will be explained in this section.

The development of each computer-adaptive and linear test starts with the decision on the *test specifications* and the construction of an *item bank*. Consideration has to be given to the following: 1) which ability or trait the test is intended to measure; 2) what kind of test questions (further referred to as *items* in this chapter) need to be developed to measure this trait; and 3) how big the item bank has to be to allow for test administration without running the risk of items being exposed too frequently.

The test specifications cover the first two points – *what* is exactly to be measured and *how* it will be measured. The 'what' refers to an aspect of language ability to be tested (e.g. reading comprehension), while the 'how' requires decisions on task types (e.g. multiple choice, multiple matching, true/false), the number of items and tasks, test content, assessment criteria, etc. (see Docherty and Corkill, this issue).

Once the test specifications have been established, test material needs to be commissioned – this test material needs to be appropriate for the test in terms of content and level of difficulty. For linear tests the commissioned material usually reflects a certain level of difficulty that the test is aimed at (e.g. B2 in terms of the Common European Framework of Reference (CEFR)) plus items at adjacent difficulty levels. For some linear tests, such as *IELTS*, the commissioned material reflects all CEFR levels. Similarly, material commissioned for a CAT test may either span all CEFR levels or a narrower range of difficulty. This mainly depends on the purpose for which the test is developed and for the intended population – for example, the *BULATS* Online test aims to measure candidate ability across all CEFR levels, while another CAT test, such as a workplace test, may be designed to differentiate between candidates from A1 up to B2 levels. While developing material for computer-based tests, particular aspects of the delivery mode need to be considered. For example, it is important to make sure that items are suitable for administration on computer – such as whether the graphic or picture accompanying the item will be clear and presentable on the screen. Once the test material is available, it undergoes the process of vetting and editing.

The crucial difference between computer-adaptive and computer-based linear tests pertains to how the test is constructed given a certain item bank. In the linear approach, a set of items is selected from the item bank for a particular test. For this purpose, items are selected with a required content and difficulty and the number of items needs to be large enough to ensure a reliable test. However, in the adaptive approach, a wider pool of items needs to be put together, which cover the appropriate range of content and difficulty and allow the algorithm to work. A CAT test is constructed each time a candidate takes a test – items are selected for a candidate from a wider pool of items based on their responses and their ability, which is estimated after they answer each item. This means that the item pool in the adaptive approach needs to be big enough to allow for tests to be constructed repeatedly over a period of time. For instance, while in a linear test 40 items would be administered to 10,000 candidates, in an adaptive test at least 400 items would need to be included in the item bank to test the same number of candidates. The adaptive testing has several advantages (discussed towards the end of this paper), but the requirements it imposes with regard to the size of the item bank are demanding.

In order for an item bank to be created, items need to be pretested (see the subsequent section). The pretesting process is a trial of future test material on a sample of candidates, followed by quantitative and qualitative analyses of pretest items with a view towards selecting the most suitable material for use in live examinations. When a CAT pretest has been constructed, it is trialled to see whether it performs as expected (e.g. whether the grading scale is appropriate or what the rater effects are) and whether the test delivery on computer is not interrupted by technical problems. Pretest items are then analysed to gather information on item functioning after which the material is either rejected, submitted for revision or included in the final item bank (for details regarding item analysis see Corrigan and Crump, this issue).

After the item bank for a CAT test has been assembled, a number of research activities are undertaken. First and foremost, simulations of the test and the item bank need to be conducted in order to ensure that the test runs as expected given the available item bank. Furthermore, a field trial is scheduled in order to assess whether tests run smoothly in the live environment.

Online pretesting

If new items need to be added to the item bank, pretesting in a computer-adaptive test offers an attractive option compared to traditional paper-based pretesting. There are two approaches to introducing pretest items in a CAT test. The first one is the 'seeded' approach in which pretest items can be administered to the candidate at any time during the test. The second approach is the 'block' approach where the test specifications in the test only contain a separate block of pretest items.

Irrespective of the pretest method, pretesting in a CAT approach has a number of advantages in terms of efficiency and precision compared to traditional ways of pretesting. Firstly, pretesting in CAT tests is more administratively efficient. It removes the need to produce separate tests composed of uncalibrated pretest items, thus eliminating any procedures and paperwork surrounding traditional paper-based pretesting. Secondly, pretesting in CAT is flexible, allowing for the activation of pretest items at any time and their deactivation after they have been exposed to a sufficient number of candidates. As a result, the pretesting process no longer consists of a series of discrete administrations and analyses but is a continuous process, allowing items to be pretested as often as necessary. This also means that pretest items are not exposed to more candidates than necessary for a meaningful analysis. Thirdly, online pretesting offers advantages in terms of efficiency of analysis, as items are calibrated against the entire calibrated item bank. Fourthly, pretest items in CAT are targeted at a much narrower and therefore more appropriate range of candidate abilities, meaning that information about items is gathered more efficiently. Last but not least, online pretesting through CAT tests has the advantage of pretest material being exposed to the live population instead of a smaller subset of candidates (for an in-depth explanation of pretesting in CAT tests see Cope and Somers 2011).

CAT test delivery

In a linear test (in a computer-based or paper-based format), each candidate receives the same items in the same order which should be completed within a specified amount of time. However, in a CAT test, candidates see different items and test duration may vary across candidates. A calibrated item bank with items across a range of difficulties is crucial for a CAT test and the most important item statistic here is the item difficulty.

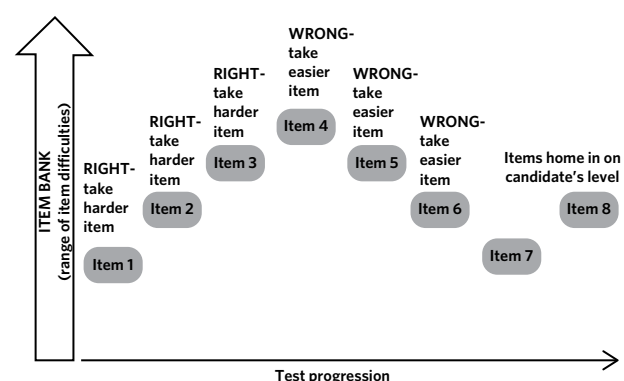
The algorithm in a CAT test successively selects questions based on whether the candidate answered the previous

question correctly or incorrectly. If a candidate answered the preceding question correctly, the next question administered will be more difficult. In turn, if a candidate answered the preceding question incorrectly, the next question administered will be less difficult.

How a computer-adaptive test generally works is described below and the basic principle of a CAT test is depicted in Figure 1.

1. The first question (item 1) that a candidate sees in a CAT test is of a certain, mostly mid-level, difficulty.
2. The computer assesses the candidate's response to the first question as right or wrong and estimates the candidate's ability level at this stage of the test.
3. Next, the algorithm selects a question at a difficulty level appropriate to the currently estimated candidate ability. In Figure 1 item 1 was answered correctly, hence the next item administered (item 2) is more difficult than item 1.
4. The process of administering items based on candidate responses continues until the test homes in on a candidate's ability level.

Figure 1: Schematic diagram illustrating the basic principle of an adaptive test



Owing to the adaptive aspect of CAT tests, test duration may vary across candidates. For some candidates the algorithm takes longer to arrive at the candidate's ability level than for others, which mostly depends on the candidate's response pattern.

Usually a CAT algorithm contains criteria which prescribe when the test should stop. In the case of Cambridge English CAT tests, the test stops when: 1) the standard error of ability estimation has fallen below a predetermined level, and 2) the candidate has reached a maximum number of items specified in the test. The first criterion ensures that the test stops when the measurement of candidate ability has reached a certain level of precision, while the second criterion makes sure that the test does not continue endlessly before the candidate ability is estimated.

Just as for linear tests, the question arises how precise the measurement of candidate ability in CAT tests is. So, besides being used to determine when a test should stop, Standard Error of Measurement is also made public to inform about test precision. For adaptive tests, the conditional Standard Error of Measurement is reported, which provides information on how precisely the test measures candidate ability at each CEFR level.

The adaptive method of testing and estimating candidate ability can be seen as more efficient than conventional tests. In the CAT approach, test items are well targeted to measure the ability level of an individual candidate; an adaptive test is 'tailored' to the individual's performance in order to obtain a precise estimate of candidate ability. In contrast, items in a linear test might not necessarily be appropriately targeted to the ability level of the same individual. In conventional linear testing, test items are targeted at a certain candidature and a lot of work is undertaken to ensure that the test is appropriate for the target population (for details see Docherty and Corkill and Corrigan and Crump, this issue). However, this may also mean that in linear tests items are more appropriate for an average examinee and at the same time some items can be too difficult for some examinees and too easy for others. This means that a linear test may not be the best choice to measure the ability level of such candidates (Weiss 2004). In sum, both linear and CAT tests are reliable tests, but the adaptive aspect of the latter allows the testing experience to be tailored to the individual pace and performance level of candidates.

From estimated candidate ability to reporting scores

The purpose of assessment, regardless of its mode of delivery (computer-based or paper-based) or nature (linear or adaptive) is to measure candidate ability, knowledge, or a certain skill. Once the ability of candidates has been estimated, it needs to be reported to the candidates in a way that is understandable and that allows comparison to other candidates in the same test and in a broader context.

Measuring candidate ability requires a choice of an appropriate measurement model. At Cambridge English, the Rasch models are used for measuring candidate ability with regard to receptive skills such as listening and reading and for establishing difficulty of test items (Rasch 1960, Wainer 2000). The most commonly used dichotomous Rasch model defines a probability of a candidate achieving a correct response as a function of the difference between the candidate's ability and the difficulty of test items (for further details see Elliott and Stevenson, this issue). As Cambridge English uses the Rasch model for item bank and test construction, the items included in a test are of a known difficulty level. Both in adaptive and linear tests, candidate ability is estimated using a Rasch formula based on the difficulty of the items seen by a candidate and whether a candidate responded correctly or incorrectly to those items. The final estimate of ability is used for scoring and grading. However, the major difference between adaptive and linear tests lies in test delivery, i.e. in the way the questions are administered to arrive at the final ability estimate (discussed in detail in the section on CAT test delivery).

The process of arriving at meaningful scores from the conclusion of a CAT test until score reporting is no different from what is done in paper-based tests. Once a candidate has taken a test, the candidate's ability level is estimated. The estimate of ability level is initially expressed in logits (for explanation of Rasch ability estimates see Elliott and Stevenson, this issue) and is subsequently translated into a

score on a standardised scale. In CAT tests, the estimated candidate ability is translated to standardised scores expressed on a scale from 0 to 100. The standardised score is more meaningful as candidates can compare their performance to performance of other candidates from the same test. The standardised score is further translated to a grade, such as 'excellent', 'good', 'satisfactory' or 'A', 'B', 'C' or 'D', or to a level such as the CEFR (for explanation of different formats of score reporting see Somers, this issue). The reported CEFR level situates a candidate's achievement in the test within a wider context and provides information both to the candidate and to the interested parties regarding what the candidate can do at this level of language proficiency (Jones 2014).

Computer-adaptive testing expedites the process of arriving at meaningful scores. Immediately after a candidate has completed a CAT test, test results are displayed on the screen – candidates see on the screen the score and CEFR level that they have achieved on the test. In computer-adaptive tests, there is no need for a candidate to wait several weeks to receive the score, as is usually the case in paper-based tests. Despite this massive advantage to the candidate, computer-adaptive testing comes at a cost. As candidate performance is immediately assessed, the immediate score reporting does not leave time for any malpractice checks before the release of test results to the candidates.

Computer-adaptive tests, similarly to linear computer-based or paper-based tests, may be constructed to measure candidate ability on a single skill (e.g. reading comprehension) or a group of skills (e.g. receptive skills – reading and listening comprehension together). Whether reading and listening comprehension abilities are estimated separately in one test largely depends on the length of the test sections measuring these abilities: a test section should be long enough to reliably estimate a candidate's ability.

It is important to note that for Cambridge English CAT tests, where both the overall score and section scores are reported, the overall score is not the average of the section scores. The overall score in such adaptive tests is calculated using the candidate's performance on the questions from both reading and listening sections. Calculating the overall score using all the questions answered provides a more accurate score than using the average of the two sections.

As we have seen above, estimating candidate ability and translating it to more meaningful scores in computer-adaptive tests is not much different from what happens for linear tests at Cambridge English. Estimates of candidate ability are translated to standardised scores and to a wider framework such as the CEFR. Also, the way candidate ability is calculated is the same for adaptive and linear tests, although the process of arriving at the candidate ability in CAT tests is different from linear (paper-based or computer-based) tests, which will be explained in the following section.

Advantages and disadvantages of computer-based testing

Computers allow a new way of test delivery that offers advantages to both exam providers and candidates. The major

advantages of using computers in assessment revolve around test delivery and score reporting.

Computer-based tests expedite the process from test development to test administration – test administration can follow shortly after the test has been developed as there is no need to produce and distribute paper copies. Moreover, computers allow certain types of test items to be scored immediately after the candidate has provided a response. Candidates can be provided with immediate results, especially in computer-based tests that assess skills which can be immediately marked by computer (such as reading and listening). Even in assessing productive skills where examiners are involved (such as speaking and writing) computers constitute an advantage as candidate responses can be assessed remotely. The development of automated assessment for speaking and writing could bring further advantages in terms of immediate availability of results. In addition, computer-based marking eliminates the problem of printing answer sheets and ambiguity regarding erasures or response marks being marked incorrectly (Wainer 2000).

Furthermore, the use of computers means improvements in terms of data storage and processing as candidate responses and scores stored electronically can be accessed at any time. This is particularly useful for malpractice detection as the data on candidate performance and candidate responses to the test is instantaneously available and can be retrieved quickly (for an overview of malpractice detection methods at Cambridge English see Bell, this issue).

In a similar vein, computer-adaptive tests have a number of advantages over standard approaches to testing. Firstly, a computer-adaptive test contributes to a positive candidate experience as thanks to the adaptive aspect such tests set a level for the candidate that is appropriate to them. Secondly, in a CAT test fewer items are necessary to achieve a sufficiently high reliability level than in a linear test as in the former the items administered to a candidate are selected based on candidate ability and candidate response pattern. This implies that CAT tests are shorter than linear tests, which has cost-savings implications for test providers. Thirdly, the CAT approach eliminates some traditional threats to test security such as the temptation to look up responses from a neighbour during the test. With a sufficiently large item bank each test instance differs from another.

At the same time, computer-adaptive tests have a few disadvantages. The first disadvantage is related to the size of the item bank – in order to produce a CAT test an item bank needs to be developed that contains more items than would be necessary for a paper-based test. In a scenario when a test is created from scratch, it would be quicker and less costly to produce a paper-based test. Similarly, a CAT test makes more stringent demands on its items than a paper-based test. As a CAT test is shorter, each item can be seen as more critical for measuring candidate ability. In addition, a flawed item can affect some examinees and not others because not every candidate gets the same set of items. This may throw the fairness of the test into question (Wainer 2000). Another disadvantage of a CAT test is related to malpractice detection – if a computer-adaptive test is administered frequently, the probability of items reoccurring in subsequent tests is higher. As a result, candidates can break the item bank by learning the items and sharing this knowledge

more widely. Although CAT tests have in-built control mechanisms to prevent overuse of items, items with certain characteristics tend to be overexposed. Such overexposure needs to be monitored, which imposes operational costs on the test provider.

Conclusion

This article highlighted the particular characteristics of computer-adaptive testing from test production up to the reporting of meaningful scores, and shed light on similarities and differences between CAT and conventional, linear tests.

Computers open vast possibilities to introduce new kinds of tests and streamline data gathering and analysis. With the help of purpose-built test delivery platforms linear and computer-adaptive tests can be created but also combined. Such CAT and mixed approaches to testing have been well researched in the recent years and offer improvements to the traditional testing approaches. In addition, computer-based testing offers many advantages for candidates and exam providers, from administration of tests on demand, immediate reporting of results, and quicker data processing to availability of shorter and yet reliable tests.

Given its advantages, computer-based testing has a promising future in assessment in general and language testing in particular. Once the initial investment of building a platform to deliver computer-based tests has been made, the advantages of computer use in language assessment will bear fruit.

References

- Cope, L and Somers, A (2011) Effective pretesting: An online solution, *Research Notes* 43, 32–35.
- Jones, N (2014) *Multilingual Frameworks: The Construction and Use of Multilingual Proficiency Frameworks*, Studies in Language Testing volume 40, Cambridge: UCLES/Cambridge University Press.
- Rasch, G (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Nielsen and Lydiche.
- Suvarov, R and Hegelheimer, V (2014) Computer-assisted language testing, in Kunnan, A J (Ed), *The Companion to Language Assessment. Approaches and Development Volume 2*, London: Wiley Blackwell, 594–611.
- Thompson, N A and Weiss, D (2011) A framework for the development of computerized adaptive tests, *Practical Assessment, Research & Evaluation* 16 (1), 1–9.
- Thurlow, A, Lazarus, S S, Albus, D and Hodgeson, J (2010) *Computer-based Testing: Practices and Considerations (Synthesis Report 78)*, Minneapolis: University of Minnesota, National Centre on Educational Outcomes.
- Wainer, H (2000) *Computerized Adaptive Testing: A Primer*, Mahwah: Lawrence Erlbaum Associates, 2nd edition.
- Weiss, D J (2004) Computerized adaptive testing for effective and efficient measurement in counselling and education, *Measurement and Evaluation in Counselling and Development* 37 (2), 70–84.

The role of Quality Management in ensuring accurate and meaningful test scores

NICK BERESFORD-KNOX RESEARCH AND VALIDATION GROUP, CAMBRIDGE ENGLISH LANGUAGE ASSESSMENT

Introduction

This paper explores how Quality Management supports Cambridge English in ensuring that tests are fit for purpose and test scores are accurate and meaningful. The building blocks of the Cambridge English quality approach will be discussed as well as the processes required to deliver our products at the right quality, meeting our customer requirements. Three key components of our quality approach will be presented: the Validity, Reliability, Impact, Practicality and Quality (VRIPQ) framework, the process approach, and the Plan, Do, Check, Act cycle.

In general terms, quality is a characteristic of a product or service and is achieved through customer satisfaction and fitness for purpose. There are three key facets to the management of quality: Governance, Assurance and Improvement (Chartered Quality Institute 2014). The principles of Governance are the management commitment to quality and structures in place to ensure quality is maintained and improved. This includes fostering a culture whereby all staff have a responsibility for quality. The process approach is one tool which assists with maintaining and improving quality at process level; this will be discussed further later in the article. Assurance involves the procedures, processes and policies that help Cambridge English achieve fitness for purpose, supported by audit functions. The final facet of quality is Improvement, which consists of the mechanisms that drive change and innovation in the organisation, forming part of the continuous improvement cycle. Continuous improvement seeks to achieve excellence, in terms of quality, customer satisfaction and efficiency, through incremental and radical improvement. The Plan, Do, Check Act cycle is one method that helps us to review and improve processes and will be discussed in this article.

Recognition of quality at Cambridge English is very important to show stakeholders we work to stringent externally verified requirements. To this end, we gained certification to the ISO 9001 Quality Management system standard in 2007. The standard is internationally recognised with a history typically in the manufacturing sector. We translated these requirements for the assessment context and to support our existing processes and systems.

VRIPQ framework

The VRIPQ framework is used to develop, maintain and revise Cambridge English language examinations (see *Principles of Good Practice* (Cambridge English 2013) for more information). It provides the necessary control mechanisms which help achieve fitness for purpose. The VRIPQ framework has roots in assessment and is used specifically in the

assessment context, while the quality methodology and the quality procedures discussed throughout this paper are generic and can be applied to any industry. This assessment-specific framework and the generic Quality Management procedures both provide a control mechanism for our assessment products and the processes that develop them to ensure fitness for purpose.

Figure 1: The VRIPQ framework

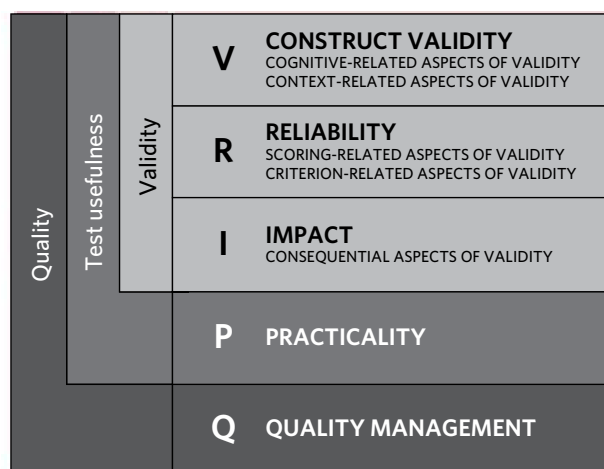


Figure 1 illustrates our approach to achieving fitness for purpose and shows how quality interacts with the key principles of language assessment – validity, reliability, impact and practicality. Validity is the extent to which an assessment can be shown to produce scores which are an accurate reflection of the test taker's true level of ability. Within the VRIPQ framework, we investigate and accumulate the evidence which provides an answer to the question: does the test measure what we want it to measure? Reliability is the extent to which test results are stable, consistent and accurate. Impact is an outward-facing concept, which includes the effects of an assessment on test users and the consequences of an assessment. Practicality focuses more on the internal processes required to produce and administer the test and whether this can be accomplished efficiently. Last but not least, Quality supports the management of the processes which are grounded in the four key principles of language assessment discussed above. Quality Management helps to ensure that tests are fit for purpose and meet customer requirements.

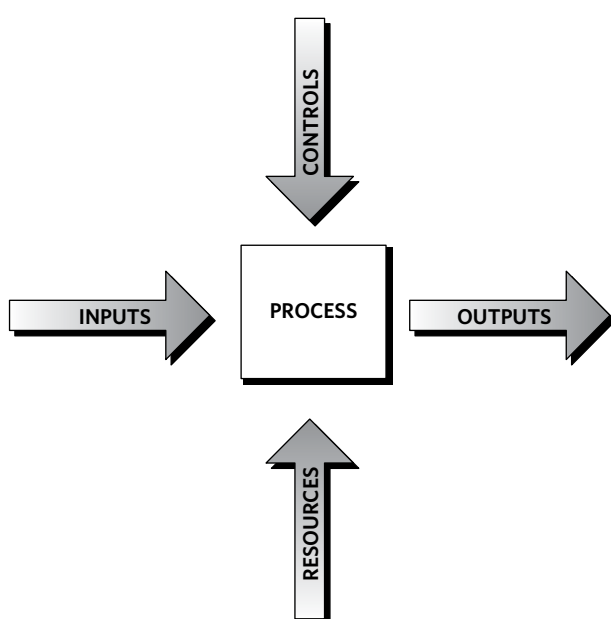
Saville (2012) explores VRIPQ and expands on the relationship between Quality Management in large-scale assessment and validity. He argues the importance of the two areas and 'the necessary convergence between the twin concepts of quality and validity'. For further information see Saville (2012).

The process approach

Within the Quality Management approach, defining processes is essential and helps ensure that quality control and quality assurance procedures can be carried out.

A process is commonly understood as ‘an activity or set of activities using resources, and managed in order to enable the transformation of inputs into outputs’ (British Standards Institution 2009). According to Figure 2, a process consists of Inputs, Controls, Resources and Outputs. The input is what goes into the process (e.g. raw materials and machinery), while the output is the final product or service which comes out of the process. These are supported by controls – which ensure that the process constantly delivers the required output at the right quality – and resources (e.g. people) – which ensure the process can operate.

Figure 2: The Process diagram

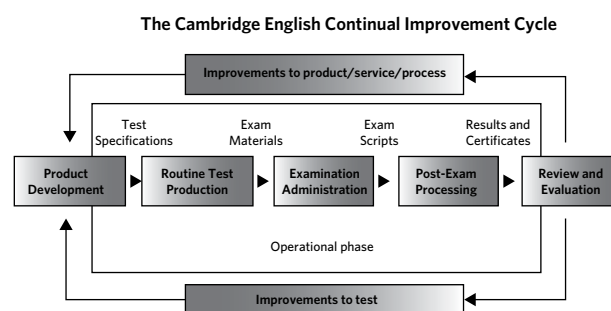


Process metrics can be determined to measure that the outputs meet or exceed expectations in terms of cost, quality and efficiency (time and resources). The model assists us in defining the requirements for the process to succeed, what controls are needed to ensure the required output is always achieved, and what resources (people, hardware, software, etc.) are used/needed to ensure this can be delivered on time. The inputs and outputs to the process may not change over time, but the methods used to turn the input into the output (the process) are continually improved. This may be in terms of time and cost savings or new systems that provide greater control, and therefore improve the quality of the outputs. The statistical analyses and processes discussed throughout this volume are a result of evolution from the Quality Management perspective. They have been modified or introduced to increase the quality of assessment, while keeping in check the practicalities of cost and time.

Figure 3 illustrates a high-level view of the processes within Cambridge English required to deliver valid examination results to candidates. These processes are of the operational and development nature, delivering the examination and results to customers. The concept of a process approach

can be illustrated by using the analogy of a production environment where the raw materials and resources enter at one end of the process, and the final product comes out of the other end. As the product works its way along the production line, the output of a stage becomes the input of the next one and each stage has its own quality control activities. In the context of language assessment in Cambridge English, the process begins with test development displayed as Product Development in Figure 3 (see also Saville 2003). This first phase in the development of a new or a revised test includes market research, to ensure that there is a need for the test, and trialling, to ensure that the test is valid and reliable. This all informs the final output in this phase which is a test specification.

Figure 3: Process interactions map



Routine Test Production is the following step. This is the first of the ongoing operational processes which create live examination materials using the test specification as a starting point. There are a number of sub-processes which make up the high-level process including Item Production, Pretesting and Test Construction. The output of the process is approved paper-based and computer-based examination materials. See Corkill and Docherty (this issue) for a detailed overview of test construction and Corrigan and Crump for item production and pretesting (also this issue).

The next steps are as follows:

- **Pre-examination Administration:** In this phase our customers begin to interact with the process, with centres making examination entries for candidates. The output of the process is the examination materials being sent to centres based on these entries or, in the case of computer-based assessment, examinations are set up on our computer-based testing platform.
- **Post-Examination Processing:** Candidate responses are returned (hardcopy and electronic) and marked. Our examiners follow a Quality Assurance programme to ensure the reliability of our marking. Following final quality checks, results and certificates are released to test takers, as the final output of the process. See Somers (this issue) for further information on the reporting of test scores and Elliott and Stevenson (this issue) for grading.

Exam Review and Evaluation: This is the final stage of the process model, where our assessment products are reviewed against our requirements. The findings of a review feed into examiner and item writer training, test revision and routine test production.

The process approach enables us to define the processes necessary to deliver our assessment services to our test users

(customers), to meet customer requirements and ensure that a test is at the right level of quality. It provides a standard approach to the definition and management of processes, and the same approach is applied to all test versions and different examinations.

Plan, Do, Check, Act (PDCA) cycle

Complementary to the process approach is the Plan, Do, Review (PDR) cycle shown in Figure 4. 'Plan' refers to defining objectives and processes, risk analysis and contingency planning; 'Do' consists of implementing the plan and collecting measurement data, while 'Review' consists of a process review, lessons learned, internal and external audit and implementing corrective action. The review stage in the Cambridge English approach encompasses 'Check' and includes lessons learnt from the review, which feed in as improvements the next time the process is implemented.

Figure 4: Plan, Do, Review wheel



The PDR cycle is more commonly known as the Plan, Do, Check, Act (PDCA) cycle, which is a process control and improvement model. The PDCA cycle was developed in the 1950s in Japan and is an evolution of earlier work called the Deming wheel. This approach enables Cambridge English to develop robust procedures across the organisation that continually evolve and improve.

The following sub-sections provide a more detailed explanation of the stages forming part of the Plan, Do and Review cycle.

Plan

The planning stage involves ensuring that all the processes and functions needed to deliver the required output of the product, service or process are in place. This may include establishing objectives for the activity or quality requirements to ensure the required output of the process is achieved. Given that objectives are established for the activity, risk assessment should be conducted to consider what could prevent the achievement of these objectives (see below for further information). The concept of PDR is multi-functional and can be applied to numerous situations, for example a project, a major process or even a small activity which makes up a larger process. This stage also encompasses the need to consider legal, regulatory and other requirements on the process see below for further information.

Risk management: Risk analysis and contingency planning

Risk management forms part of the planning stage in Figure 4 and plays an important role in the management of any organisation. It is the process of identifying possible barriers that could prevent the achievement of organisational objectives and implementing preventive measures or mitigation to reduce the likelihood of them affecting Cambridge English. As well as being fundamental to process and business management, risk is core to the management of all projects. The well-known and widely used PRINCE2 methodology incorporates the concept of risk, and risk registers are commonly used in the project lifecycle.

We define risks based on the cause and consider the potential impact or effects on achieving the goals of the organisation – in our case, delivering accurate assessment systems on time and to our stated quality standards.

Commonly, risks are prioritised following an evaluation of their impact and likelihood, with higher scores requiring the most urgent action. Descriptors are used to help standardise the scoring of risks at each of the levels of likelihood and impact (see Table 1).

Table 1: Descriptors for likelihood of risk impact

		Impact				
		Very Low (VL) 1	Low (L) 2	Medium (M) 3	High (H) 4	Very High (VH) 5
Likelihood	Very Low (VL) 1	1	2	3	4	5
	Low (L) 2	2	4	6	8	10
	Medium (M) 3	3	6	9	12	16
	High (VH) 4	4	8	12	16	20
	Very High (VH) 5	5	10	16	20	26

■ Critical ■ Serious ■ Moderate □ Minor □ Insignificant

We can use this information for decision-making purposes; for example, should we terminate an activity if it poses too much risk for the organisation and is above our 'level of tolerance'? More commonly, the risk can be addressed in some way in order to reduce its score. The aim of such 'mitigation' is to effectively manage the risk in order to reduce its likelihood of becoming a problem or even eliminating it all together. As circumstances are always changing, regular review of the risks is required so trends can be reviewed over time and the level of risk re-evaluated if necessary.

As well as using risk management to prevent risks from occurring, we also use it to minimise the impact in the worst case scenario – i.e. if a known risk were to occur. The use of contingency planning related to risks involves the development of backup plans. The aim of the contingency planning is to minimise the impact of the incident on the assessment systems themselves and on the stakeholders who may be affected.

Cambridge English follows best practice in the field of risk management with risk identification and treatment across all aspects of our activities – from people to processes and projects. We draw on requirements from the ISO 9001 Standard, the UK's qualifications and examinations regulator (Ofqual) and other sources of best practice. Risks are managed on a hierarchy of risk registers, with the most critical risks being escalated to the senior management team. It is vital for all staff to have an awareness of where the risks lie within processes they work on. This is to ensure that suitable

attention is given to certain tasks and the consequences of actions are clear.

Process requirements

Cambridge English has a number of regulated qualifications in the UK and we are therefore recognised and monitored by Ofqual, the body that regulates qualifications, examinations and assessments in England and vocational qualifications in Northern Ireland. Ofqual provides a document called *General Conditions of Recognition* (2014) which outlines requirements for awarding bodies. A declaration of compliance is made against this document and monitoring visits take place. The document provides requirements for all stages of the assessment process and provides us with another assurance mechanism.

Many Cambridge English examinations have also been externally audited by ALTE (Association of Language Testers in Europe), an association which provides another set of principles that help us to ensure the Cambridge English assessment processes provide accurate results to candidates. The audited Cambridge English examinations have been awarded the Q-mark in meeting ALTE Standards. 'The ALTE Q-mark is a new quality indicator which member organisations can use to show that their exams have passed a rigorous audit and meet all 17 of ALTE's quality standards' (Association of Language Testers in Europe 2014).

Do

The 'Do' stage is where the plans developed are implemented; this may be the operational stage of a process, the manufacturing of a product or the implementation of a project. Where products or activities deviate from the intended plan, corrections should be put in place to eliminate the problem (see incident management below).

Incident management

If a risk event happens, this is classified as an 'incident'. Incidents could include, but are not limited to: complaints, despatch delays, system problems, malpractice, etc. Cambridge English has robust procedures in place to manage all types of incident, above all to ensure minimal impact on external stakeholders. A key part of this is in analysing what caused the problem and how we can stop it happening again (in Quality Management terminology this is known as corrective action). We use 'root cause analysis' to identify the underlying cause and to decide how best we can go about fixing it.

Effective preventive action is vital to ensure the same or a similar problem does not happen again. For example, a problem may have been caused by poor instructions or staff training. In this case, we would revise the induction process and look at increasing the controls and checking of the work instructions for new employees.

Review

This final stage holds the key to continuous improvement. Review is based on an analysis of the performance of a product, process or project in the cycle and looking at what could be improved in the future. Typically, this might take the shape of a process review session or lessons learned

in project methodology. The analysis can be supported by using audit mechanisms (see below) as assurance, review and improvement activities. Furthermore, Cambridge English has an established review mechanism, Exam Review and Evaluation, which is explained below.

Internal audit

As part of our Quality Management system, we are committed to a thorough internal audit programme. We use the risk-based approach in developing a programme of internal audits around our core processes. The audit acts as an assurance and improvement activity, with auditors identifying whether staff are following procedures, identifying potential improvements and ensuring conformance with ISO 9001 requirements. We recruit Cambridge English staff as internal auditors, providing them with thorough training. Staff do not audit their own work and the benefit comes from a 'second pair of eyes' looking at a process from another perspective and with different experiences within our assessment systems. See Beresford-Knox and Rose (2010) for further information on the development of the Cambridge English Internal Audit System.

External audit

As mentioned previously, the Cambridge English Quality Management system is ISO 9001 certified and was first registered in 2007. The external certification provides us with external review of all our processes over a 3-year cycle, with audits covering all our UK-based locations. Our external certification body to ISO 9001 is the British Standards Institution.

In addition to auditing as a business function, which is used to ensure adherence to procedures and ISO 9001, auditing is also an important aspect of language testing. Li (2006) looks at the usage of audit trails in the test development process where 'the onus is placed on test developers and the project team to ensure adequate audit trails relating to the changes of the test and the reasons why the test was changed, and effectively tell the story of the test evolution process'. The audit trail records the decisions and justifications in the test development process, therefore providing evidence for the Validity argument. The audit trail and auditing of this evidence 'makes the test development process more transparent and explicit to the different stakeholder groups'. The QMS approach adopted by ALTE (as discussed in the Process requirements section previously) follows a similar methodology whereby a testing organisation defines their validity argument and an audit is conducted to review the evidence against 17 key principles in language testing. Minimum standards for establishing quality profiles in ALTE examinations are set for the following areas (Association of Language Testers in Europe 2014): test construction, administration and logistics, marking and grading, test analysis and communication with stakeholders.

Exam review and evaluation

The review focuses on the performance of the tests against our requirements. The process takes a systematic approach to the review of data for a given year with the results of the review feeding into improvements to process or test revision. This process helps to ensure that the examinations continually

meet our stated requirements and that we are providing an accurate measurement of candidate performance. The review focuses on the following four areas:

Pretesting: What proportion of pretested tasks and items meet the given statistical criteria? Are the pretests targeting a sufficient variety of candidates, and are they at the right level?

Test Construction: Are the test versions meeting the appropriate test construction criteria?

Live Test Candidature – profile and performance: Is there a change in composition of the candidature (e.g. age, first language)? Is there any change in the overall candidate performance?

Live Test Performance: How reliable are the tests? What proportion of tasks and items meet the given statistical criteria?

Concluding remarks on how a quality approach helps Cambridge English ensure meaningful and accurate results

The key to the quality approach is determining the requirements and objectives necessary to ensure meaningful and accurate results. These then provide the building blocks for creating standard processes and procedures in order to deliver our stated requirements. These procedures ensure that we consistently meet our requirements and provide a repeatable system. Control is essential to ensuring accuracy and the process approach assists us in defining this and determining the controls necessary in each process to maintain quality.

The risk management process helps to identify events that could affect the accuracy of results before they happen. Where risks are identified, and could impact the accuracy of results, extra controls can be put in place to ensure that the risk does not occur.

We provide further assurance to our system through the use of regular internal auditing. Processes which can impact result accuracy follow a thorough auditing regime to ensure we are following the defined procedures. We gain additional external verification of our systems through our external assessment to ISO 9001 as well as other requirements and regulations we adhere to.

This all forms part of the Cambridge English approach to control and continuous improvement through the use of a Plan, Do, Check, Act cycle throughout our process management. Such an approach is necessary to ensure the maintenance of assessment standards over time. Without it, we would risk having frequent and recurrent problems and not consistently meeting or not exceeding customer requirements.

In addition to the generic components of the Quality Management system which can be applied to any industry (i.e. the process approach and Plan, Do, Check, Act cycle), the Cambridge English Quality Management system also rests on the VRIPQ framework which ensures that Cambridge English tests are fit for purpose and that they possess key assessment qualities which, in turn, ensure meaningfulness and accuracy of test results.

References

- Association of Language Testers in Europe (2014) *Minimum Standards for Establishing Quality Profiles in ALTE Examinations*, available online: www.alte.org/attachments/files/minimum_standards.pdf
- Beresford-Knox, N and Rose, D (2010) Implementing an internal audit process: A case study from Cambridge ESOL, *Research Notes* 39, 8-10.
- British Standards Institution (2009) *ISO 9001:2008: Quality Management Systems – Requirements*, London: British Standards Institution.
- Cambridge English (2013) *Principles of Good Practice: Quality Management and Validation in Language Assessment*, Cambridge: Cambridge English Language Assessment.
- Chartered Quality Institute (CQI) (2014) *Managing Organisation Quality*, available online: www.thecqi.org/The-CQI/What-is-quality/
- Li, J (2006) *Introducing Audit Trails to the World of Language Testing*, Illinois: University of Illinois.
- Ofqual (2014) *General Conditions of Recognition*, available online: www.gov.uk/government/uploads/system/uploads/attachment_data/file/371266/2014-11-03-general-conditions-of-recognition-november.pdf
- Saville, N (2003) The process of test development and revision within UCLES EFL, in Weir, C J and Milanovic, M (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913-2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 57-120.
- Saville, N (2012) Quality management in test production and administration, in Fulcher, G and Davidson, F (Eds) *The Routledge Handbook of Language Testing*, London: Routledge, 395-412.

Glossary

Anchor (items, test): A set of items of known difficulty used to calibrate other items.

Assessment scale: A scale used for judging candidate performance with descriptors arranged into a series of levels.

Bespoke test: A test that is customised for a particular purpose or set of candidates.

Biometrics: Metrics related to human characteristics that can be used as a form of identification and access control. Examples include fingerprints, facial recognition and voice recognition.

Calibration: The means by which difficulty of an item is ascertained, relative to other items of known difficulty.

Classical test theory (CTT): An approach to the analysis of item performance based on observed scores.

Construct: An attribute of a person that often cannot be measured directly but can be assessed using test items that are supposed to assess the construct. For example, several similar test items may assess the reading comprehension ability of an individual.

Criterion referenced: Decisions are made by comparison against defined criteria e.g. to pass a driving test candidates must demonstrate specific skills.

Cut score: The mark which is used to define a particular decision point, such as the pass mark. It represents the minimum score (or cut-off point) at which a particular result will be awarded.

Difficulty (of an item/task/test): A measure of how easy or hard an item is. It is determined through calibration which, unlike facility, accounts for the strength of the population taking that item.

Discrete task: A task with only one item.

Discrimination: The extent to which an item differentiates between strong and weak candidates.

Facility (of an item/task/test): Proportion of candidates responding correctly.

Formative assessment: Assessment primarily intended to inform teaching and learning.

Grading: The process of establishing cut scores/pass marks for a test.

Item: A testing point which is given a mark(s). For example, a multiple-choice question with four options is an item.

Item banking: The practice of storing items and information about them, such as difficulty estimates, for future test construction.

Item response theory (IRT): A statistical model used to determine the relative difficulty of tests and ensure that the results on each test are comparable.

Linear testing: Tests of fixed format where all test takers would see the same number of items in a sequential way. Almost all paper-based tests are linear.

Live test: A test administered to candidates in order to measure their abilities, rather than to obtain information about the test.

Malpractice: Action which allows a taker or takers of an examination to gain an unfair advantage.

Measurement scale: A scale such as the numbers on a ruler which divide the property of length into equal units. It could be of four kinds: nominal (using categories such as male/female to assign values), ordinal (ranking individuals on the basis of performance), interval (breaking the intervals into equal distances) and ratio (which starts at zero and has equal intervals). The latter is rarely used in language assessment as language ability cannot be assessed in ratio terms.

Multi-item task: A task with more than one item. Typically the items are linked together by a rubric and some type of input (e.g. a reading passage or listening text).

Norm-referenced: Decisions are made by comparison against other individuals, e.g. in an audition, the best candidate will be chosen.

Objectively marked: Tests/tasks/items which are marked with reference to a defined answer key; it does not require any judgement in choosing what mark to award.

Parameters (item parameters): Under a statistical model for item analysis, parameters are the values that describe that item, such as the difficulty of an item.

Practicality: Practical considerations in constructing and administering tests. The term practicality covers a range of issues such as cost of development, test length, time, font size, candidate's fatigue, modes of test administration, marking, availability of test results, test security and many other practical considerations which will all have an impact on test design.

Pretesting: The administration of items to obtain information about the performance of the items, rather than the candidates. Pretesting is usually employed as part of a quality control procedure.

Productive skills: Writing and speaking.

Psychometrics: The measurement of psychological traits such as language ability. It covers all measurement properties of a test and is based on some measurement theory.

Qualitative: Based on a quality or characteristic, typically not measured or quantified.

Quality assurance: The systems, processes and procedures that help to ensure the delivery of quality products and services.

Quantitative: Based on quantity or quantifiable data.

Rating scale (see Assessment scale)

Receptive skills: Listening and reading.

Reliability: The consistency of scores obtained by the same persons when re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable conditions.

Score scale: A scale that defines the overall results that are possible for an exam, showing which constitute stronger or weaker performance.

Standard Error of Measurement (SEM): Shows the impact of test reliability on the likely score of an individual: it indicates

how close a test taker's score is likely to be to their 'true score'.

Structural Equation Modelling: A statistical technique for building and testing statistical models. It is normally used to investigate the underlying constructs within tests.

Subjectively marked: Tests/tasks/items where marking requires an expert judgement by a trained professional working to set guidelines to decide on the most appropriate mark.

Summative assessment: Assessment primarily intended to measure prior learning.

Task: A collection of one or more items that have been designed to measure a specific language learning trait.

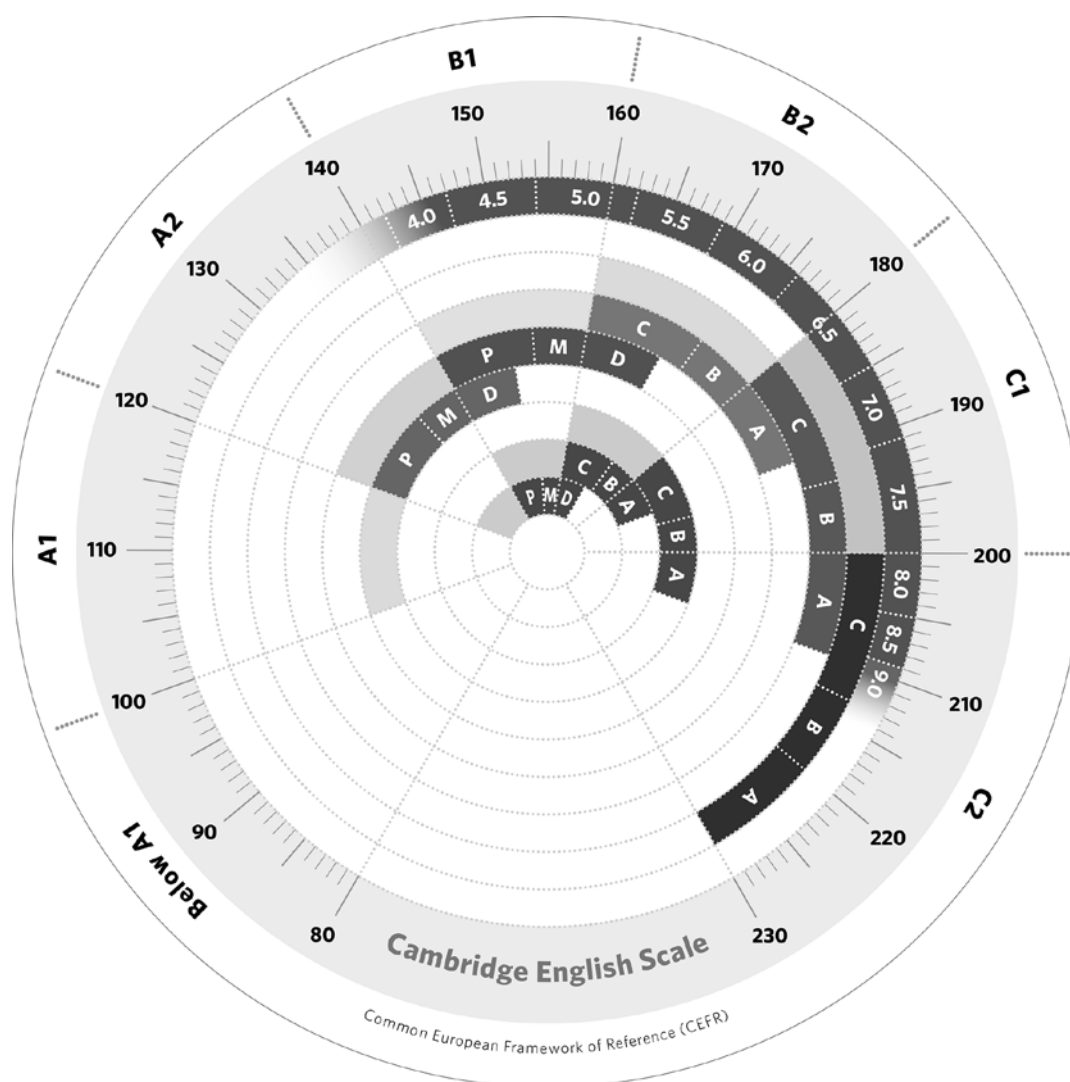
Unidimensionality: There is only a single trait being measured.

An example of unidimensional measurement is measuring the length of an object with a ruler; the longer the object, the greater the value of its length. When we measure the area of an object that is two-dimensional – we need to measure its length and breadth to get an accurate view of its area – increasing one or other or both will increase the area.

Validity: The degree to which evidence and theory support the interpretations of test scores. It refers to whether a test measures what it intends to measure.

The Cambridge English Scale

The Cambridge English Scale is a single range of scores used to report the results of our exams. The diagram below shows how our exams map onto the scale and align to one another. Please note that IELTS is mapped too, but will not be reported on the scale.



Studies in Language Testing

The *Studies in Language Testing* (SiLT) series focuses on important developments in language testing and assessment and profiles work of direct relevance to the field of language assessment. The question of test reliability and the provision of accurate and meaningful results – which is the focus of this *Research Notes* issue – has been a theme running through several SiLT titles. It formed part of the first SiLT title published in 1995 by Professor Lyle F Bachman and colleagues and marked the beginning of reform and development at

Cambridge over the years that followed. Key SiLT titles which have focused on related questions include: SiLT 1: *An Investigation into the Comparability of Two Tests of English as a Foreign Language*, SiLT 3: *Performance Testing, Cognition and Assessment*, SiLT 10: *Issues in Computer-Adaptive Testing of Reading Proficiency*, SiLT 15: *Continuity and Innovation*, SiLT 23: *Assessing Academic English*, SiLT 26: *Examining Writing*, SiLT 29: *Examining Reading*, SiLT 30: *Examining Speaking*, SiLT 35: *Examining Listening*, and SiLT 37: *Measured Constructs*.

Studies in Language Testing

An indispensable resource for anyone interested in new developments and research in language testing



To find out more about our full list of publications:

www.cambridge.org/elt/silt

www.cambridgeenglish.org/silt



To subscribe to *Research Notes* and download previous issues, please visit:
www.cambridgeenglish.org/research-notes

Contents:

Editorial	2
Introduction Ardeshir Geranpayeh	3
Item analysis Michael Corrigan and Paul Crump	4
Test construction: The Cambridge English approach Coreen Docherty and David Corkill	10
Grading and test equating Mark Elliott and Lynne Stevenson	14
Determining standards in assessing language ability Gad S Lim	20
Reporting test scores and the Cambridge English Scale Andrew Q Somers	23
A modern perspective on statistical malpractice detection Chris Bell	31
Computer-adaptive testing Agnieszka Walczak	35
The role of Quality Management in ensuring accurate and meaningful test scores Nick Beresford-Knox	40
Glossary	45
The Cambridge English Scale	47
Studies in Language Testing	47

For further information visit the website:
www.cambridgeenglish.org

Cambridge English
 Language Assessment
 1 Hills Road
 Cambridge
 CB1 2EU
 United Kingdom

www.cambridgeenglish.org/help



All details are correct at the time of going to print in March 2015