References

AoC (2009). *AoC Diploma Survey Report – Results, analysis, conclusions and recommendations*. London: Association of Colleges. Available at: http://www.aoc.co.uk/en/newsroom/aoc_news_releases.cfm/id/C5B28C9E-78EC-4E75-BAE3D47ECDF3BD48

DfES (2005a). *14–19 Education and Skills White Paper*. London: DfES. Available at: http://publications.dcsf.gov.uk/eOrderingDownload/CM%206476.pdf

DfES (2005b). *14–19 Education and skills implementation plan*. London: DfES. Available at: http://publications.teachernet.gov.uk/eOrderingDownload/2037-2005PDF-EN-01.pdf

Ertl, H., Stanley, J., Huddleston, P., Stasz, C., Laczik, A., & Hayward, G. (2009). *Reviewing diploma development: evaluation of the design of the diploma qualifications*. London: Department for Children, Schools and Families. Available at: http://www.dcsf.gov.uk/research/data/uploadfiles/DCSF-RW080%20(Rev).pdf

McCrum, E., Macfadyen, T., Fuller, C., & Kempe, A. (2009). *Vocational education and training: some perspectives from Year 11*. Paper presented at the British Educational Research Association Annual Conference, 3–6 September 2009, Manchester.

O'Donnell, L., Lynch, S., Wade, P., Featherstone, G., Shuayh, M., Golden, S., & Haynes, G. (2009). *National evaluation of Diplomas: Preparation for 2008 delivery*. London: Department for Children, Schools and Families. Available at: http://publications.dcsf.gov.uk/eOrderingDownload/DCSF-RW079.pdf

Ofsted (2009). *Implementation of 14–19 reforms, including the introduction of Diplomas*. London: Office for Standards in Education, Children's Services and Skills.

Richardson, W. & Haynes, G. (2009). *National evaluation of diplomas – findings from the 2008 survey of Higher Education Institutions on their implementation and impact*. London: Department for Children, Schools and Families. Available at: http://www.dcsf.gov.uk/research/data/uploadfiles/DCSF-RR145.pdf

The effect of changing component grade boundaries on the assessment outcome in GCSEs and A levels

Tom Bramley and Vikas Dhawan Research Division

Acknowledgements

This paper is based on one section of a report commissioned by Ofqual (the exams regulator in England) as part of its Reliability Programme. For more details on this programme, see http://www.ofqual.gov.uk/research-and-statistics/92-articles/20-reliability. We would like to thank Ofqual and its technical advisory group for their feedback. The opinions expressed are those of the authors. Figure 1 and its commentary were not part of the original report.

Introduction

Investigations of assessment reliability are concerned with answering the question 'how would the assessment outcomes change if the assessment were replicated?' The answer to this question depends on what factors are held constant and what factors change on replication. For example, the examination questions could be different, or the markers (examiners) could be different – or both these could be held constant and the only change might be in the mood or level of preparation or other factors internal to the examinees. A further factor relevant to GCSE and A level assessments is that these are graded examinations, where grade boundaries are set on the raw mark scale of each of the units/components comprising the assessment. These boundaries are then aggregated in a particular way depending on the type of assessment to produce the overall grades for the assessment. It is therefore possible to consider a replication scenario where questions, markers and examinee internal factors remain the same, but the grade boundaries (and hence the grade outcomes) are different.

A variety of sources of evidence can be used to inform the decisions about where to set the grade boundaries, including:

- 'archive' scripts at the key grade boundary marks from previous sessions;
- information about the size and composition (e.g. type of school attended) of the cohort of examinees;
- teachers' forecast grades;
- the distribution of scores (mean, SD, cumulative % of examinees at each mark);
- at GCE, 'putative' grade distributions (grade distributions generated by matching examinees with their GCSE results and taking account of changes in the 'ability' of the cohort of examinees from a previous[1] session, as indicated by changes in the distribution of mean GCSE scores;
- experts' judgements about the quality of work evident in a small sample of scripts covering a range of consecutive marks (total scores) around where the boundary under consideration is expected to be found;
- experts' judgements about the difficulty of the question paper;
- other external evidence suggesting that the particular unit/component (or assessment as a whole) had previously been severely or leniently graded and needs to be 'brought into line' with other examination boards, or with other similar subjects or specifications within the same board.

1. Usually this is the previous session with a cohort believed to be most similar to the current session's cohort, e.g. for a June 2009 unit, the June 2008 session might be used rather than the January 2009 session.

These pieces of evidence do not necessarily always 'point in the same direction', and therefore they need to be weighed appropriately – a matter which ultimately requires human judgement, although it is fair to say that most weight is given to statistical methods that take account of changes in the 'ability' of the cohort. Given that it is therefore not possible to determine exactly what the grade boundaries 'should' be, it is of interest to investigate what the impact of slightly different decisions at unit/component level would be on the grade distributions at whole assessment level. In particular, it seems likely that the evidence for any particular grade boundary decision could support two possible boundary marks, and perhaps more.

Whilst it would in principle be possible to carry out an actual replication of the grade boundary setting process, varying some of its characteristics (e.g. decision-making personnel, scripts viewed etc.), considerable if not prohibitive logistical (and financial) problems would arise.

Therefore, we carried out a simple 'sensitivity analysis' in order to determine the effect on assessment grade boundaries of varying the (judgementally set) key grade boundaries on the units/components by ±1 mark. In this paper we report the results of this analysis for two assessments with different structures – a tiered 'linear' GCSE, and a 6-unit 'modular' A level. The data came from the June 2009 examination session administered by OCR.

The effect of varying component grade boundaries on a tiered, linear GCSE examination

In this assessment, Foundation Tier examinees took two written papers and a coursework component. Higher Tier examinees took two different written papers and the same coursework component (which therefore had the same grade boundaries for each tier).

In linear assessments, there are two ways of deriving the aggregate grade boundary from the component grade boundaries. The first, known as 'Indicator 1', is the simple aggregate of the component grade boundaries, taking account of the weight of each component in the aggregate total. In this GCSE the two written papers each carried 40% weight and the coursework 20%, and their paper totals were in these proportions, which meant that indicator 1 could simply be obtained by adding up the grade boundary marks on the three components. Tables 1 and 2 below show the range of possible boundaries at grade C (Foundation) and grade A (Higher) obtainable if the boundaries on some or all of the three components were changed by ±1 mark. The column '# combinations' shows how many ways there were of arriving at that particular aggregate boundary mark. Clearly there is only one way of arriving at a mark 3 lower or higher – that is, by raising or lowering each component boundary by 1 mark. However, the various other permutations lead to more ways of arriving at boundaries within this range.

Tables 1 and 2 show that the possible values for the actual aggregate grade boundary could have led to fluctuations covering a range of up to 9.5 percentage points in the pass rate at grade C on the Foundation Tier, and up to 6 percentage points in the pass rate at grade A on the Higher Tier. Even a ±1 mark difference from the actual boundary would have given a range of ≈3 percentage points at grade C on the Foundation Tier and ≈2 percentage points at grade A on the Higher Tier.

Table 1: Foundation Tier – possible aggregate grade C boundaries (Indicator 1 only)

Aggregate boundary	*# combinations*	*Cumulative % of examinees at Grade C*
93	1	34.50
92	3	35.96
91	6	37.74
90	7	39.47
89	6	41.04
88	3	42.50
87	1	43.96

Table 2: Higher Tier – possible aggregate grade A boundaries (Indicator 1 only)

Aggregate boundary	*# combinations*	*Cumulative % of examinees at Grade A*
167	1	13.71
166	3	14.47
165	6	15.41
164	7	16.55
163	6	17.30
162	3	18.47
161	1	19.70

The second method of calculating the aggregate boundary, known as 'indicator 2', involves finding the mark on the aggregate distribution of marks (the distribution obtained by adding together each examinee's mark on each component, appropriately weighted) where the cumulative percentage of examinees obtaining that mark corresponds most closely to the percentage obtained by taking a weighted average of the cumulative percentage of examinees at that particular boundary on each of the components. Indicator 2 is usually closer to the mean aggregate mark than indicator 1, which means it is usually lower than indicator 1 at the higher boundaries, and vice versa. The Code of Practice (Ofqual, 2009) allows the awarding panel to choose any mark between (and including) the two indicators as the final aggregate boundary mark. The default position is to take the lower of the two indicators[2].

The effect of including indicator 2 was to increase the range of possible boundaries by one mark down to a mark of 86 at grade C on the Foundation Tier (cf. 87 with indicator 1, see Table 1). The effect was greater at grade A on the Higher Tier, where it increased the range of possibilities by a further six marks down to a possible mark of 155 (cf. 161 in Table 2).

In statistical tables of examination results, the outcomes for the two tiers of the examination are combined rather than published separately. Grade A is only available on the Higher Tier, but grade C is available on both tiers, which dramatically increases the number of overall possible outcomes at grade C if the boundaries on all components on *both* tiers fluctuate by ±1 mark. As we have seen, the more extreme outcomes are less likely to arise because they would require a change in the same

2. The Code of Practice states "Whenever the two indicators do not coincide, the grade boundary should normally be set at the lower of the two indicator marks, unless, in the awarders' judgement, there is good reason, as a result of a review of the statistical and technical evidence, to choose a higher mark within the range spanned by the indicators." Ofqual (2009) p.53.

direction on all components. Table 3 below shows the extreme (widest possible) range and a more plausible range based on the most likely aggregate outcomes for the overall grade A and C on the whole assessment (both tiers combined).

Table 3: Overall possible pass rate outcomes (cumulative % of examinees)

	Cumulative % of examinees	
	Extreme range	*More plausible range*
Grade A	8.6 to 16.5	10.3 to 14.4
Grade C	55.9 to 63.2	58.2 to 61.0

In summary, a range of 3–4 percentage points seems like a reasonable range in which the cumulative percentage outcomes at grades A and C on this linear GCSE might fluctuate. This value is contextualised in the discussion section (see later).

The effect of varying unit grade boundaries on a modular 6-unit A level

GCE AS and A levels are 'modular' or 'unitised' – that is, examinees are assessed in discrete units. Most AS levels consist of 2 units, but some contain 3. Most A levels consist of 4 units, but some contain 6. The A levels include the corresponding AS units, plus further 'A2' units. The A2 units do not form a qualification on their own, unlike the AS units. The number and choice of units depends on the specification (syllabus). Most units are 'available' in examination sessions in January and June. Any exceptions or restrictions are stated in the specification. Examinees would generally take AS units in the first year of a 2-year A level course, and the A2 units in the second year. Units can be re-taken individually: in other words if an examinee wishes to improve their aggregate grade they do not need to re-take every unit in the assessment.

Because of this choice and flexibility in modular assessment schemes, a different method for deriving the aggregate grade boundaries is required. AS and A level units have a 'Uniform Mark Scale' (UMS). The key grade boundaries 'A' and 'E' are set on the raw mark scale for each unit, and these raw marks are converted to fixed boundaries on the UMS. The conversion between raw and uniform marks is linear within the A–E range and extended slightly beyond it – see AQA (2009) and Gray and Shaw (2009) for further details. For 6-unit A levels the aggregate grade A boundary is at 480 UMS marks (out of 600), and for grade E it is at 240 UMS.

In terms of the effect on aggregate outcome of changes to the unit boundaries, it is only reasonable to consider the effect of changes made in a particular examination session. This is because once the unit boundaries have been set, they cannot later be changed. So, when considering the effect of changing the boundaries on all units of a 6-unit A level in June 2009 by ±1 mark, it should be emphasised that the vast majority of examinees would already have taken units in previous sessions – probably the three AS units in January and June 2008, and perhaps one A2 unit in January 2009. Table 4 shows part of the breakdown of numbers of examinees taking units in June 2009.

Table 4 makes it clear that nearly half of the examinees aggregating in June 2009 had just taken two or three A2 units in June 2009. Only 3% had taken all six units in June 2009.

Table 4: A 6-unit A level – number of aggregating examinees taking each unit in June 2009 (total N=11,603). Only combinations with more than 100 examinees are shown

AS units			*A2 units*				
Unit 1	*Unit 2*	*Unit 3*	*Unit 4*	*Unit 5*	*Unit 6*	*N*	*%*
✗	✗	✗	✗	✓	✓	2929	25.24
✗	✗	✗	✓	✗	✓	288	2.48
✗	✗	✗	✓	✓	✓	2348	20.24
✗	✗	✓	✗	✓	✓	657	5.66
✗	✗	✓	✓	✓	✓	500	4.31
✗	✓	✗	✗	✓	✓	239	2.06
✗	✓	✗	✓	✓	✓	736	6.34
✗	✓	✓	✓	✓	✓	327	2.82
✓	✗	✗	✗	✓	✓	738	6.36
✓	✗	✗	✓	✓	✓	476	4.10
✓	✗	✓	✗	✓	✓	405	3.49
✓	✗	✓	✓	✓	✓	283	2.44
✓	✓	✗	✓	✓	✓	317	2.73
✓	✓	✓	✓	✓	✓	352	3.03

Changing the boundaries on all six units by ±1 mark would give 3^6 = 729 possible scenarios. Given the complexity of the computations required to derive the final grade distributions (which involve obtaining unit-level UMS distributions going back several years) it was only feasible to investigate a relevant selection of these possible scenarios. The outcomes are shown in Table 5.

Table 5 shows that the variability of aggregation outcomes at grade A was ≈3 percentage points when the grade A boundaries on the 6 units were moved by ±1 mark. Changing all the AS units simultaneously only affected the outcome by about 0.5 percentage points. Not surprisingly, given the entry patterns shown in Table 4, changes to the A2 units had more impact – changing the boundary on either Unit 4, Unit 5 or Unit 6 had as much impact as changing the boundary on all three AS units. Unit 5 and Unit 6 on the A2 appeared to be more influential than Unit 4, but given that more of the aggregating examinees had taken Unit 5 and Unit 6 in June 2009 this is not surprising.

Table 5: A 6-unit A level – effect of varying June 2009 unit grade A boundaries on overall % of examinees at grade A (actual outcome in bold)

Unit 1 June 2009	*Unit 2 June 2009*	*Unit 3 June 2009*	*Unit 4 June 2009*	*Unit 5 June 2009*	*Unit 6 June 2009*	*Cumulative % aggregate grade A (N=11,603)*
-1	-1	-1	-1	-1	-1	33.41
0	0	0	-1	-1	-1	32.94
0	0	0	0	0	-1	32.35
-1	-1	-1	0	0	0	32.33
0	0	0	0	-1	0	32.31
0	0	0	-1	0	0	32.07
0	**0**	**0**	**0**	**0**	**0**	**31.84**
0	0	0	+1	0	0	31.60
0	0	0	0	0	+1	31.39
+1	+1	+1	0	0	0	31.36
0	0	0	0	+1	0	31.27
0	0	0	+1	+1	+1	30.79
+1	+1	+1	+1	+1	+1	30.35

Discussion

Given all the sources of information that can be used in setting grade boundaries, some of which relate to different definitions of what it might mean to 'maintain a standard' (see, for example, Baird, 2007; Coe, 2010; Newton, 2010) and which therefore can suggest different locations for the grade boundaries, it should be clear that the setting of grade boundaries is not a problem with a clear-cut answer. Therefore, it is perhaps of interest to consider how the outcomes might have been different if different decisions had been taken[3]. The analyses presented here give some indication of what such reporting might look like. Two potentially useful ways of quantifying the potential variability in aggregate outcome are:

- to determine the range of possible aggregate outcomes that could have arisen if all relevant key grade boundary decisions at unit/component level had been 1 mark lower or 1 mark higher;
- to discover the largest change to the aggregate outcome that could have arisen from a 1-mark change in the boundary on a single unit/component.

The most obvious factors affecting the sensitivity of the aggregate outcome to decisions on the individual units/components are: i) the number of units/components to be aggregated – the greater the number the less the effect of changes on any one unit/component; and ii) the percentage of examinees on each mark point at the part of the distribution where the grade boundary lies (on each unit in unitised schemes, but on the aggregate distribution in linear schemes[4]). Units with longer raw mark scales, all things being equal, might be expected to have a lower percentage of examinees on each mark point. The correlation of scores among the units can also be expected to have an effect, with changes to grade boundaries on more highly correlated units/components affecting the aggregate more.

A more subtle point relating to unitised assessments is the effect of potential grade boundary changes to the 'conversion rate' of raw marks to uniform marks. Changes that reduce the distance between the A and the E boundary (i.e. lowering the A boundary and/or raising the E boundary) increase the rate of exchange; and vice versa. So whereas on a linear assessment a change to a component boundary changes the aggregate boundary but does not affect the aggregate totals of any examinees, in a unitised assessment a change to a unit boundary does not affect the aggregate UMS boundary but does affect the unit (and hence the aggregate) UMS total of most of the examinees who took that unit. So on a linear assessment (for example a higher tier GCSE) a change to a component grade A boundary could not affect the cumulative percentage of examinees obtaining aggregate grade C, but on a unitised assessment a change to a unit grade A boundary could conceivably affect the cumulative percentage of examinees obtaining aggregate grade E. Admittedly this effect is likely to be very small for the ±1 mark changes we are talking about. In the case of the 6-unit A level reported here, lowering the grade A boundary by 1 mark on all six units would have resulted in an extra 3 examinees (out of 11,603) obtaining an aggregate grade E. Lowering the A boundary and the E boundary by 1 mark on all six units would have resulted in an extra 30 examinees obtaining an aggregate grade E. Interestingly, lowering the E boundary by 1 mark and raising the A boundary by 1 mark on all six units would have resulted in an extra 38 examinees obtaining aggregate grade E! This illustrates the point that the UMS conversion can have some slightly counter-intuitive effects – but supports the claim that the proportion of examinees affected is likely to be very small.

In unitised assessments it is very difficult to gauge or control the impact of changes at unit level because of the large number of different valid combinations of units, from different examination sessions, that can be aggregated to achieve an overall result at assessment level. Decisions made in a particular examination session cannot have any effect on the UMS scores on units from previous sessions. For the new unitised GCSEs, certificated for the first time in June 2010, 'terminal rules' specify that a certain proportion of the units must be taken in the same session that aggregation will take place, which will presumably mitigate this problem to some extent.

We therefore would argue that an appropriate way to quantify 'grading reliability' would be to consider the range of possible outcomes (grade distributions) that could have been obtained if grade boundary decisions taken in a particular session had been slightly different. We have chosen to define 'slightly different' as 'varying by ±1 mark', that is, the smallest difference possible. There would be some justification for taking a wider range, given that the 'zone of uncertainty'[5] in expert judgement of script quality usually spans a range wider than ±1 mark. The results presented here could then be seen as lower bounds.

To put the kinds of variability we have found into context, Table 6 shows the cumulative percentage of examinees obtaining grade A from June 2006 to June 2009 in the two assessments discussed above. This table uses the 'final' data on the system, rather than the data available at the time of awarding used in the analyses presented above, so the numbers of examinees do not exactly match.

Table 6: Grade A cumulative percentages and number of examinees, 2006–2009

Qualification		*2006*	*2007*	*2008*	*2009*
Linear GCSE	%	13.6	12.2	12.5	14.6
	N	3323	3977	4764	5244
6-unit A level	%	28.6	29.9	30.9	31.7
	N	10290	11113	11472	11874

It is very striking how similar the cumulative percentages gaining grade A were from year to year in the period 2006–2009, given that the examinees were different and the size of the entry varied somewhat. In no case was the largest difference between any pair of years more than 3.1 percentage points, and most adjacent pairs of years differed by less than 1 percentage point. On the other hand, the analysis above showed that the possible range of variation in percentage at grade A with *exactly the same examinees* could be from around 2 to 4 percentage points, if boundary marks on all units/components were changed by ± 1 mark. This

3. Of course, examination boards do consider the aggregate effect of the decisions made at unit level at the time when those decisions are taken, in 'modelling' exercises. We are suggesting here that the range of possible fluctuation had decisions been slightly different could be reported more systematically.

4. For linear schemes that use indicator 1 only. If indicator 2 is used then the number of examinees at mark points around the boundary on the individual components is also relevant.

5. The term formerly given to the range of marks over which there was no consensus among a panel of experts that the quality of scripts was definitely worth the higher or lower of two adjacent grades. Nowadays this range is referred to simply as the 'zone' – presumably so as not to give the impression that there is any uncertainty in the process!

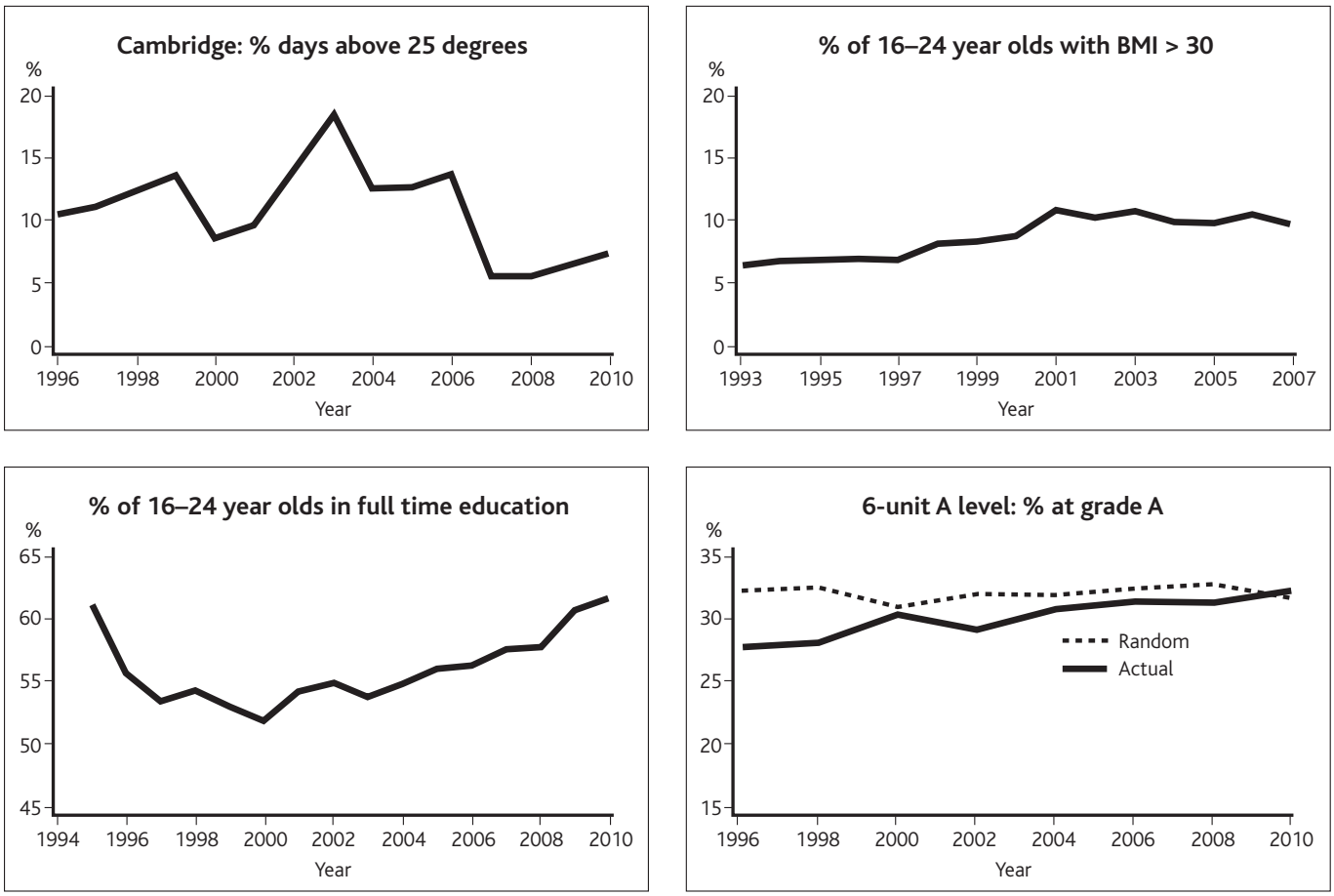

Figure 1: Illustrations of various trends of data expressed as percentages

Data sources for Figure 1

Top left: Cambridge computer laboratory daily weather record http://www.cl.cam.ac.uk/research/dtg/weather/index-daily-text.html Accessed 8/2/11.

Top right: NHS information centre, Body Mass Index (BMI) data. http://www.ic.nhs.uk/webfiles/publications/HSE07/ADULT%20TREND%20TABLES%202007.xls Accessed 8/2/11.

Bottom left: UK National Statistics publication hub: Labour market statistics: educational status, economic activity & inactivity of young people. http://www.statistics.gov.uk/StatBase/xsdataset.asp?vlnk=5740&More=Y Accessed 8/2/11.

Bottom right: Joint Council for Qualifications: inter-awarding body statistics.

suggests that the current statistically driven grade-boundary setting procedures could be 'overfitting' and producing a year-on-year grade distribution that does not fluctuate enough, given all the conceptual conflicts and practical limitations of the standard maintaining process. Of course, given public expectations about 'standards' it might be difficult to explain that a more fluctuating grade distribution is perfectly acceptable. On the other hand, it would help to avoid the pattern that is sometimes seen of steady year-on-year small incremental rises in pass rates that lead to accusations of 'grade drift' (see, for example, Oates, 2009 and its coverage in Paton, 2010).

As an illustration of this point, Figure 1 shows four graphs of time series data where the variable plotted is a percentage. The y-axis covers the same range of percentage points for ease of comparison. It can be seen that while fluctuations in A level % pass rate at grade A in Chemistry are of a similar order of magnitude to the other variables (with the exception of warm days in Cambridge!), there is a tendency for consistent very small increases. By contrast, eight bootstrap samples from the 2009 aggregate distribution in the large-entry 6-unit A level (shown as the dashed 'random' line in the bottom-right graph in Figure 1) showed the kind of fluctuations in pass rate that might be expected if random variation was the only source of year-on-year differences.

The fact that the observed fluctuations are of a similar size to the random fluctuations, but in a more consistent (upward) direction could be explained by saying that in the years when random fluctuations would increase the pass rate, they are the only factor operating, but in the years when they would decrease it, other factors act to cancel them out by more in the opposite direction. However, this does seem rather implausible. A more likely explanation is that awarding panels look for the 'safety in numbers' that the statistical sources of evidence appear to provide, and combine this with a tendency to give examinees the 'benefit of the doubt' when undecided about two adjacent marks for a grade boundary (Stringer, 2008).

In summary, reliability investigations seek to show how outcomes would vary if some factors were changed while others remained constant. One factor affecting outcomes is the decision of the awarding panel on where to locate the grade boundaries on the raw mark scale of each unit/component. Small changes to grade boundaries of the units/components of the linear GCSE and modular A level reported here would have produced fluctuations in the cumulative percentage of examinees reaching the boundary in a 2–4 percentage point range. This is slightly larger than the range of fluctuation that might be expected from random sampling variability (in large entry subjects), and larger than the observed range of changes across a period of several years. We suggest that this finding supports the claim that the observed pass rates do not fluctuate enough in both directions and that the current boundary-setting procedures might be achieving a tighter level of statistical control than is necessary or appropriate.

References

AQA (2009). Uniform marks in GCE, GCSE and Functional Skills exams and points in the Diploma. http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 16/02/10.

Baird, J.-A. (2007). Alternative conceptions of comparability. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education*, **25**, 3, 271–284.

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

Newton, P.E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, **25**, 3, 285–292.

Oates, T. (2009). 'Standards are up this year' – what does this mean? The question of standards in public examinations. http://cambridgeassessment.files.wordpress.com/2010/01/the-question-of-standards-in-public-examinations-by-tim-oates1.pdf Accessed 17/5/10.

Ofqual (2009). GCSE, GCE and AEA code of practice, April 2009. http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf Accessed 08/01/10.

Paton, G. (2010). GCSE and A level results being 'inflated'. *Daily Telegraph*. http://www.telegraph.co.uk/education/educationnews/7528383/GCSE-and-A level-results-being-inflated.html Accessed 17/5/10.

Stringer, N. (2008). *An appropriate role for professional judgement in maintaining standards in English General Qualifications*. Paper presented at the International Association for Educational Assessment annual conference, Cambridge, September 2008.

An American university case study approach to predictive validity: Exploring the issues

Stuart Shaw and Clare Bailey CIE

Introduction

Predictive validity research is fundamental to test validation (Davies *et al*., 1999). Predictive validity entails the comparison of test scores with some other measure for the same candidates taken some time after the test has been given (see Anastasi, 1988; Alderson *et al*., 1995). In psychometric terms, predictive validity is the extent to which a scale predicts scores on some external (future) criterion measure. It is the prediction of criterion performance that is basic to validation. For tests that are used for university selection purposes it is vital to demonstrate predictive validity.

However, establishing predictive validity through relating secondary school performance to later academic performance is fraught with practical difficulties in mounting tracer studies and the problems associated with confounding intervening variables that obscure the effects of another variable (see Banerjee, 2003, for a critique of such approaches to establishing predictive validity). These difficulties notwithstanding, predictive validity is still regarded a vital aspect of the validation process. Moreover, predictive validity research is becoming increasingly necessary as test providers are being challenged to pay greater attention to issues of test comparability – both in terms of the relationships between their own assessment products and those offered by other competitor, examination boards.

A common need for predictive validity is inherent in the process of selecting students for university. Consequently, this article will focus on the research being conducted by University of Cambridge International Exams (hereafter simply 'Cambridge') to ensure that its international assessments prepare students well for continued studies in colleges and universities. The long-term purpose of the research is to highlight the predictive validity of Cambridge assessments and other students' characteristics to predict preparedness for and continued academic success at U.S. universities in terms of first year Grade Point Average (GPA).

This study takes a case study approach. The research reported here uses data collected from three cohorts of students enrolled at Florida State University. The data include information about each student's performance at high school, ethnicity, gender and first year GPA. Multilevel modelling has been applied to the data using the statistical software package MLwiN[1] to investigate the relationships between the variables, and in particular to determine which are the best indicators of academic success at university, whilst taking into account the effects of individual high schools. Issues relating to choice of predictive and university success measures, intervening variables, controlling for selection bias, data and measurement, and choice of research model will be discussed in the context of an American university.

U.S. secondary school indicators for success

Given the increase in the number of applications for admissions to colleges and universities for the limited number of seats in freshmen classes, students and universities in the U.S. must consider all available indicators for success in higher education. There are many ways a student can gain recognition to contribute towards their university application. The standard high school exam in the U.S. is the SAT (formerly known as the Scholastic Aptitude Test) although in some states an alternative, the

1. www.cmm.bristol.ac.uk/index.shtml

2. Concordance tables are published to find equivalences so that SAT scores can be used for the minority of students who take the ACT.