Shapley, K. S. & Bush, M. J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: building on practical experience. *Applied Measurement in Education*, **12**, 11–32.

Shavelson, R. J., Baxter, G. P. & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher*, **21**, 22–27.

Stobart, G. (1988). *Differentiation in practice: A summary report*. London: University of London Schools Examination Board.

Stobart, G., Elwood, J. & Quinlan, J. (1992). Gender bias in examinations: how equal are the opportunities? *British Educational Research Journal*, **18**, 3, 261–276.

Taylor, M. (1992). *The reliability of judgements made by coursework assessors*. AEB Research Report RAC 577.

Thomas, I. (2006). *Internal Grading Report*. Cambridge: Cambridge International Examinations.

Trew, K. & Turner, I. (1994). Gender and objective test performance. In: K. Trew, G. Mulhern & P. Montgomery (1994). *Girls and women in education*. Leicester: The British Psychological Society, Northern Ireland Branch.

Wikström, C. & Wikström, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper school grades. *Economics of Education Review*, **24**, 3, 309–322.

Wilmut, J., Wood, R. & Murphy, R. (1996). *A review of research into the reliability of examinations*. Nottingham: University of Nottingham, School of Education.

Wood, R. (1991). *Assessment and testing: A survey of research*. Cambridge: University of Cambridge Local Examinations Syndicate.

Using simulated data to model the effect of inter-marker correlation on classification consistency

Tim Gill and Tom Bramley Research Division

Introduction

Measurement error in classical test theory is defined as the difference between a candidate's observed score on a test and his or her 'true' score, where the true score can be thought of as the average of all observed scores over an infinite number of testings (Lord and Novick, 1968). The observed score X on any test is thus:

$$X = T + E$$

where T is the true score and E the (random) error component. Whilst classical test theory recognises several sources of this measurement error, arguably the source of most concern to an awarding body is that due to markers – in other words the question 'what is the chance that a candidate would get a different mark (or grade) if their script were marked by a different marker?' (Bramley, 2007). Therefore, for the purposes of this article, the E in the above equation refers to marker error only. Other factors affecting measurement error such as the candidate's state of mind on the day of the exam or whether the questions they have revised 'come up' may be thought of as more acceptable by the general public; these are considered to be the luck of the draw. Getting a different grade dependent on the marker is much harder to accept.

However, the marking of exam papers is never going to be 100% reliable unless all exams consist entirely of multiple-choice or other completely objective questions. Different opinions on the quality of the work, different interpretations of the mark schemes, misunderstandings of mark schemes, or incorrect addition of marks all create the potential for candidates to receive a different mark depending on which examiner marks their paper. Awarding bodies put great effort into annual attempts to increase reliability of marking with standardisation meetings, scrutiny of sample scripts from each marker and scaling of some markers. However, these measures are far from perfect: examiners may make different errors in the scripts that are sampled than in other scripts. Scaling is a broad-brush approach, and it has been shown that it can cause more than 40% of the marks given by the scaled examiner to be taken further away from the 'correct' mark (Murphy, 1977 quoted in Newton, 1996).

Arguably, however, the real concern for examinees is not that they might get a different mark from a different examiner, but that they might be awarded a different *grade*. Investigations of the extent to which this occurs have been relatively few, judging by the published UK research literature (see next section for a review), probably because of the cost associated with organising a blind double-marking exercise large enough to answer some of the key questions. The purpose of this study was to use *simulated* data to estimate the extent to which examinees might get a different grade for i) different levels of correlation between markers and ii) for different grade bandwidths.

To do this we simulated sets of test scores in a range of scenarios representing different degrees of correlation between two hypothetical markers, and calculated the proportion of cases which received the same grade, which differed by one grade, two grades, etc. The effect of grade bandwidth on these proportions was investigated. Score distributions in different subjects were simulated by using reasonable values for mean and standard deviation and plausible inter-marker correlations based on previous research. The relative effect on unit grade and syllabus grade was also investigated.

Correlation is traditionally used as the index of marker reliability. Here we discuss some other indices and explore different ways of presenting marker agreement data for best possible communication.

Background and context

It is important at this point to emphasise a distinction that comes up in the literature on misclassification in tests and exams. This is the difference between classification *accuracy* and classification *consistency*. 'Accuracy' refers to the extent to which the classification generated by

the observed score is in agreement with that generated by the candidate's true score (if we knew it). 'Consistency' refers to the proportion of examinees that would be classified the same on another, parallel form of the test (or for our purposes, classified the same by a different marker in the same test). The indices we are interested in are those relating to classification *consistency*, since we do not know the 'correct' mark. In this paper we ignore the impact of other sources of error attributable to the examinee, the particular test questions, etc.

The simplest consistency index is the proportion of candidates (P_0) getting the same grade from the two markers. As an illustration, the following cross tabulation shows the proportion of candidates given each grade by the two different markers, x and y, with an inter-marker correlation of 0.995:

Table 1 : An example cross tabulation of proportions of candidates awarded each grade (simulated data)

y grade	*x grade*						
	A	*B*	*C*	*D*	*E*	*U*	*Total*
A	0.160	0.010	<0.001	0	0	0	0.170
B	0.010	0.088	0.014	<0.001	0	0	0.111
C	<0.001	0.014	0.109	0.016	<0.001	0	0.138
D	0	<0.001	0.016	0.117	0.016	<0.001	0.149
E	0	0	<0.001	0.016	0.152	0.013	0.181
U	0	0	0	0	0.013	0.239	0.252
Total	**0.170**	**0.111**	**0.138**	**0.148**	**0.181**	**0.252**	**1.000**

Hence the proportion of candidates consistently classified is the sum of the diagonal values (P_0=0.865) and therefore the proportion inconsistently classified is 1-0.865 = 0.135.

Please (1971) used this method of measuring misclassification in terms of the difference between the observed grade and the true grade. Thus, he was referring to a measure of classification accuracy and not classification consistency.

He estimated levels of misclassification using this method with reliability coefficients of between 0.75 and 1 for A-levels (on the assumption of a known fixed percent getting each grade – 10% getting A, 15% getting a B etc[1]). For example, with a correlation of 0.93 between true and observed score (and thus reliability, the square of the correlation, equal to 0.865) only 74% of A grades were classified correctly with 24% getting a B and 2% a C. For an exam with reliability of 0.83 or less, more than half the candidates would be wrongly graded. He determined that a reliability of 0.97 was required before less than 25% would be wrongly graded.

Two other UK authors (Cresswell, 1986; Wiliam, 2000) also looked at the reliability of tests by simulating data and reporting the proportion of candidates with the same observed and true grades (although Wiliam actually reported the percentage *incorrectly* classified). By comparing observed score with true score classifications, they were again looking at classification accuracy, not consistency. Both papers showed that increasing the reliability of the test increases the proportion correctly classified, and that increasing the number of grades or levels reduces the proportion. This second conclusion makes intuitive sense, merely because there are a larger number of categories into which to be misclassified.

As Cresswell points out however, increasing the number of grades has the compensatory factor of reducing the severity of any misclassification. For instance, misclassification by one grade on an exam with ten different grades is less serious than a misclassification on an exam with only two grades (pass/fail).

Livingston and Lewis (1995) used the mark distribution on one form of a test to estimate classification consistency on an alternate form. However, they did not look at the overall level of classification, but at the level at each of the grade boundaries in turn. Thus at grade B, the inconsistently classified candidates would be those that would be awarded *at least* a B on one form of the test (marker x in our case), but would get a lower grade from another form (marker y). This gives a series of 2x2 contingency tables for each grade. Using the data from Table 1 we have:

Table 2 : 2x2 contingency tables of proportion of candidates classified at A and B boundaries

x grade	*y grade*	
	A	*B–E*
A	0.160	0.010
B-E	0.010	0.820

inconsistent classification =0.01+0.01= 0.02

x grade	*y grade*	
	A,B	*C,D,E*
A,B	0.268	0.014
C,D,E	0.014	0.705

inconsistent classification =0.014+0.014= 0.028

This index is relevant for UK exams when considering what the results of GCSE or A-level exams are used for: for instance, GCSE results are often summarised in terms of the number getting 5 grade C or above, in which case a candidate misclassified from a grade C to a B or A is less serious than one misclassified from a C to a D. Similarly, A-level results are often used to select candidates for university. The index could then be used to measure candidates who would have been awarded grades good enough to achieve the university's offer by one marker, but not by another.

Lee, Hanson and Brennan (2000) used three different models to simulate data. For each they estimated classification consistency indices, which were calculated for all of the grade boundaries at once or each boundary separately. They also calculated the above indices dependent on the true score. These had the unsurprising outcome that on true scores around where the cut-off points lay the levels of inconsistent classification were higher than on scores in the middle of the categories.

Methodology

We generated a set of exam scores from two hypothetical markers, such that the correlation between the two sets of marks was at a certain level. This was done by simulating a large[2] set of normally distributed data (with mean zero and unit standard deviation): this was the data for marker x. Another set of normally distributed data was generated which correlated with the first set of data to a certain level (say 0.90): this was the data for the second marker (y). Both sets of data were then un-standardised using real means and standard deviations based on past live exam data. This converted the data into two possible sets of normally distributed marks based on the real mean and standard deviation for that

1 In 1971 the number of grades available at A-level was different than today, being A, B, C, D, E, O and F.

2 N>1,000,000. In the scatter plots (Figures 1 and 2) the number of data points has been reduced for clarity.

subject/unit and with the required level of correlation between the two hypothetical markers. This is represented graphically in Figure 1. It should be noted at this point that the simulated data gave both markers the same severity or leniency. The correlation between two examiners, one who marks consistently higher than the other may be very high, but would tend to lead to more inconsistent classification than with two markers with the same level of severity. However, the impact of this is beyond the scope of this article.

The next step was to add in the grade boundaries on both axes. By using the actual boundaries that were used for awarding we determined the number and proportion of candidates that might have been awarded a different grade if their script had been marked by a different marker, for a given level of correlation:

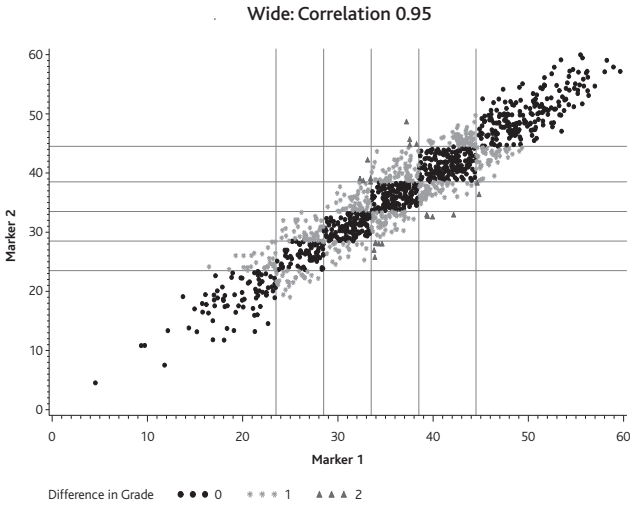


Figure 1 : Scatter plot of marks from two hypothetical markers with grade boundaries (inter-marker correlation = 0.95).

Inspecting the graph gives an idea of the proportion of candidates getting a different grade depending on which marker marked their paper. The candidates who received the same grade are the dots, those who received one grade different are the triangles, and two grades different are stars. The precise proportions of consistent and inconsistent classifications are shown later in Tables 4 and 5.

The next step was to vary the level of inter-marker correlation. It is well documented that this varies between subjects (e.g. Murphy, 1978, 1982; Newton, 1996; Vidal Rodeiro, 2007). Papers with a large number of highly structured questions (Maths, Physics, etc) generate higher correlations than those with a few long answer essay type questions (English, History, etc). This suggests the amount of inconsistent classification will also be different, with a higher level in subjects with lower correlation. Thus we simulated data at different levels of correlation (0.995, 0.95, 0.90, 0.80 and 0.70) and recorded the effect on the amount of inconsistent classification. This is further complicated by the number of grade boundaries and where they lie within the mark range. The closer together the grade boundaries are, and the more grades there are, the more candidates are likely to be inconsistently classified. For example, in an A-level unit with five boundaries all with a width of five marks, the A-E mark range is 25 marks. If a candidate's two scores from the hypothetical examiners differed by three marks then there is a good chance they will get a different grade from each marker, but there is still a fair chance that their classification would be the same under both markers. Now take a unit with grades that are only three marks wide, an A-E mark range of 15 marks. Our candidate with the three mark difference is now sure to get a different grade from each marker. For the same reason, a subject with a narrower grade bandwidth (but the same score distribution) will generate more inconsistent classifications. Whilst it would have been possible to examine the 'pure' effect of changing the grade bandwidth on the same set of simulated data, we felt this would be somewhat unrealistic, since in practice the grade bandwidths depend on the score distributions. Therefore we carried out simulations based on real data for two different subjects, with different A–E bandwidths, and compared the levels of inconsistent classification in each. It is important to emphasise that the 'narrow' and 'wide' units differed in more than the width of their grade bands. Table 3 shows that they also differed in terms of mean, standard deviation and the percentage of candidates in each grade. Therefore comparisons between them in terms of classification consistency do not show the 'pure' effect of spacing of boundaries. However, they do illustrate two realistic scenarios with some interesting contrasts.

Some factors may have a double effect on the inconsistent classification. Increasing the length of an exam for instance is likely to reduce the problem in two ways. First, longer tests tend to increase the inter-marker reliability (Murphy, 1978, 1982) and secondly a longer test is likely to have boundaries that are more separated.

Two A-level units were chosen for this research (from the June 2006 session); both with the same maximum mark but one with relatively closely spaced grade boundaries (A–E width of 13 marks) and one with relatively widely spaced grade boundaries (A–E width of 21 marks). Descriptive data for the two units are shown in Table 3 below.

Table 3 : Descriptive data for the units used

	Narrow		*Wide*	
Candidates	5296		12543	
Max marks	60		60	
Mean	31.86		36.99	
SD	8.01		9.60	
Boundary	*Cut Score*	*% in grade*	*Cut Score*	*% in grade*
A	40	17.75	45	22.72
B	37	11.78	39	23.58
C	34	13.78	34	19.91
D	31	14.41	29	14.68
E	27	16.79	24	10.09
U	0	25.49	0	9.03

We looked at the potential number of candidates inconsistently classified in both units, for different levels of correlation.

Results

We first confirmed that the data we generated could reasonably have come from a real application of the exam by comparing the score distributions generated by each of the simulated markers with the real distribution. Because the simulated data were normally distributed, some observations were above the maximum or below the minimum mark. These were excluded from the analysis. Also, the observations generated were not whole numbers and thus needed to be rounded. These two adjustments had the effect of very slightly altering the mean and standard deviations of the simulated distributions and the correlation

between the two simulated markers. However, these differences were such a small magnitude that they can safely be ignored.

In Table 4 below, P_0 is the overall level of classification consistency, (the sum of the diagonal elements in the cross-tabulations) for the two units at different levels of correlation.

Table 4 : Proportion of candidates consistently classified at different levels of correlation

	Correlation	P_0
Narrow	0.995	0.865
	0.99	0.809
	0.95	0.616
	0.9	0.523
	0.8	0.429
	0.7	0.372
Wide	0.995	0.881
	0.99	0.832
	0.95	0.637
	0.9	0.528
	0.8	0.418
	0.7	0.356

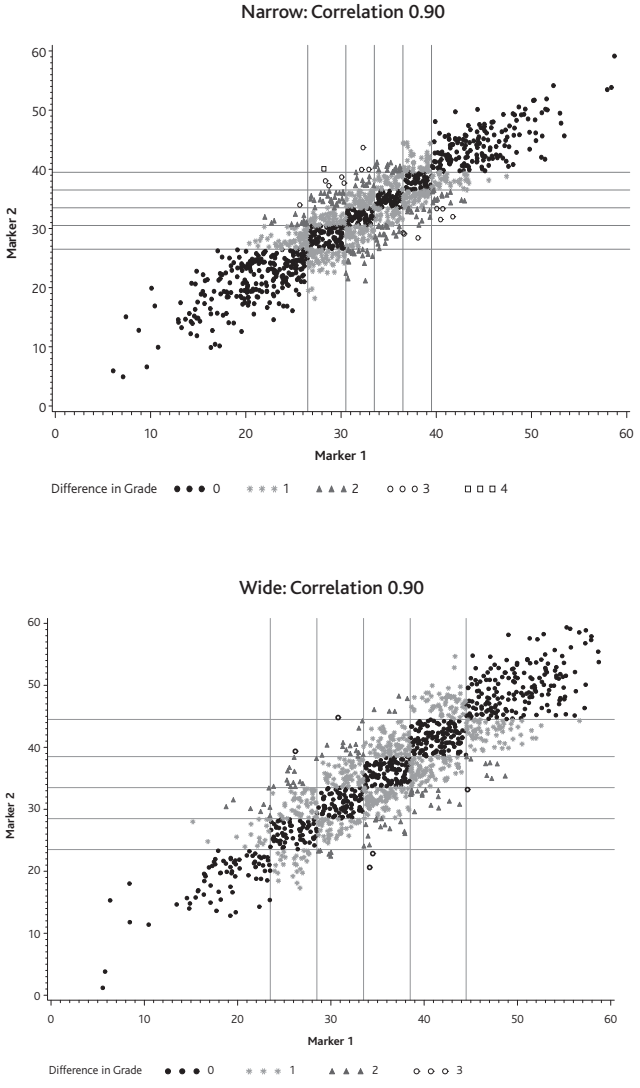


Figure 2 : Scatter plot of marks from two hypothetical markers (inter-marker correlation = 0.90)

It is clear that the impact on the proportion consistently classified of changes in the correlation coefficient between the two simulated markers was substantial. As expected, this fell with the level of correlation. For the narrow unit the percentage consistently classified fell from 86.5% at a correlation of 0.995 to 37.7% at a correlation of 0.7. For the wide unit the fall was slightly larger, from 88.1% to 35.6% consistently classified.

To demonstrate the levels of consistent classification visually, Figure 2 plots the marks from the two markers for both units, with a correlation of 0.90. Note that on the graphs the lines representing the boundaries have been set 0.5 marks below the actual boundaries, to show more clearly which mark points are in a particular grade and which are out.

We also looked at the classification consistency conditional on the *mark* given by one of the markers. This is the proportion of candidates on each mark (from marker x) given the same *grade* by marker y. This is best represented graphically, as shown in Figure 3.

These graphs demonstrate that for both units the levels of consistent classification fell considerably with marks on and around the grade boundaries (the vertical lines represent the boundaries). The peaks in the graphs are at marks in the middle of the boundaries. This is what we would expect, since for a mark on the grade boundary a difference of just one mark between the two markers (in one direction) is enough for inconsistent classification, whereas in the middle of the boundary a difference of two or three marks is necessary. It is worth noting that the differences between the peaks and troughs were much lower for low levels of correlation.

Severity of inconsistent classification

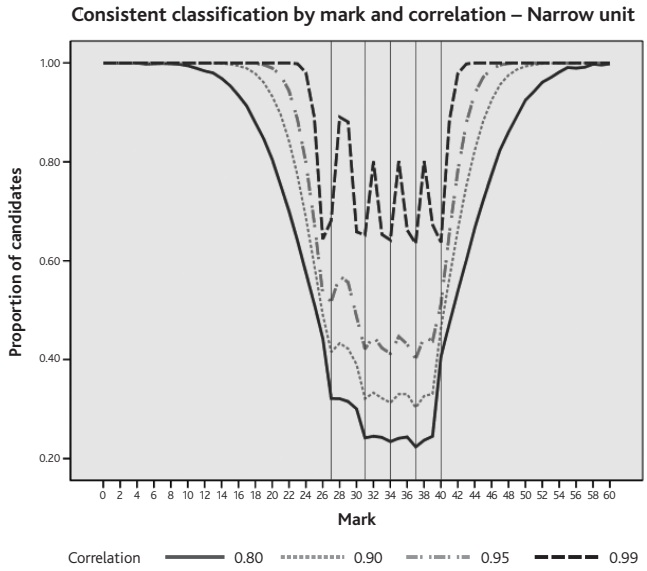
What the above indices do not take account of is the severity of the inconsistent classification – the proportions that were inconsistently classified by one grade, by two grades and so on. This is shown in Table 5 below:

Table 5 : Severity of inconsistent classification

	Correlation	*Proportion inconsistently classified by*					
		0 grades	*1 grade*	*2 grades*	*3 grades*	*4 grades*	*5 grades*
Narrow	0.995	0.865	0.135	<0.001	0	0	0
	0.99	0.809	0.191	<0.001	0	0	0
	0.95	0.617	0.341	0.042	0.002	<0.001	0
	0.9	0.523	0.363	0.099	0.014	<0.001	<0.001
	0.8	0.428	0.353	0.157	0.051	0.010	<0.001
	0.7	0.372	0.336	0.182	0.081	0.026	0.004
Wide	0.995	0.881	0.119	0	0	0	0
	0.99	0.832	0.168	<0.001	0	0	0
	0.95	0.637	0.349	0.014	<0.001	<0.001	0
	0.9	0.528	0.412	0.058	0.003	<0.001	<0.001
	0.8	0.418	0.427	0.132	0.022	0.002	<0.001
	0.7	0.356	0.414	0.175	0.047	0.007	<0.001

At correlations of 0.995 and 0.99 very nearly all of the candidates were classified within one grade for both units. At a correlation of 0.95 this was still the case, but the percentage inconsistently classified by one grade increased to over 30%. At a correlation of 0.90, around 11% of the candidates on the narrow unit and 6% of the candidates on the wide unit were inconsistently classified by two grades or more.

As with the proportion consistently classified we also produced graphs for the severity of inconsistent classification (by at least one, two or



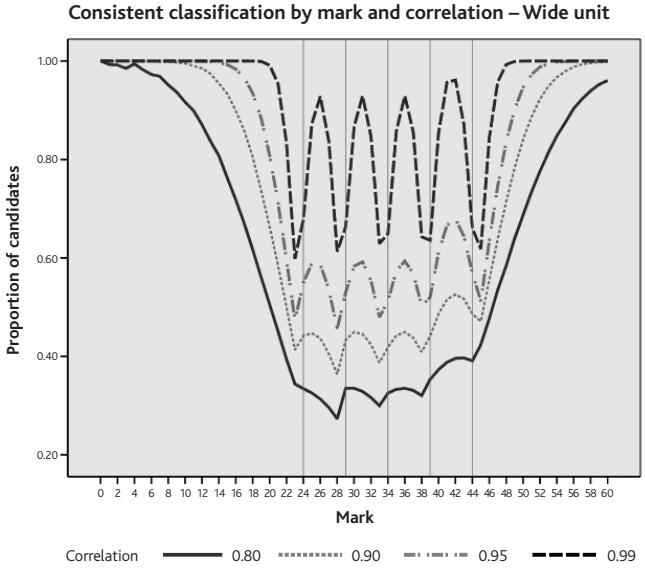
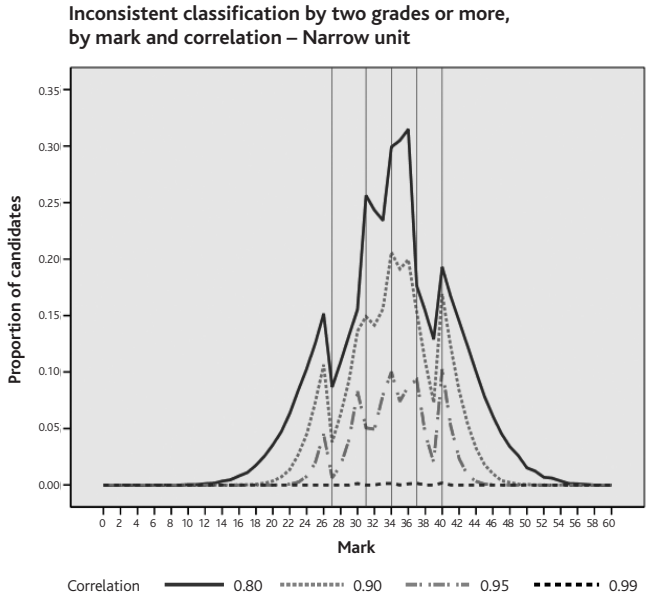


Figure 3 : Consistent classification by mark and correlation – Narrow Unit, Wide Unit[3]



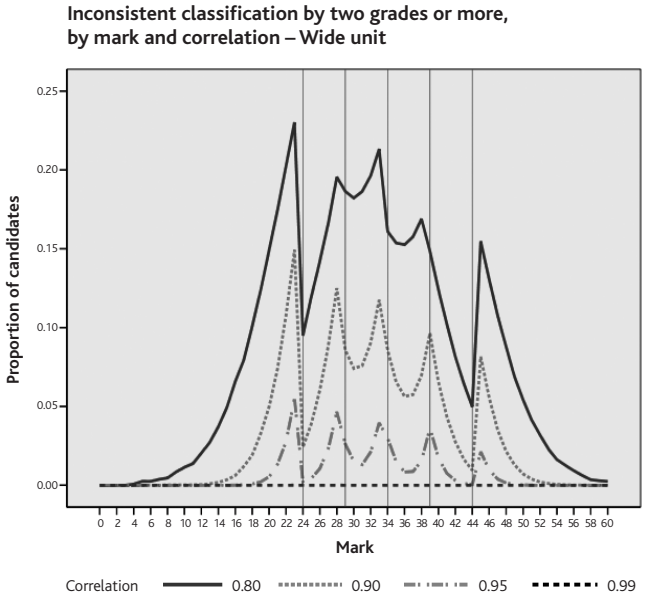


Figure 4: Inconsistent classification by two grades or more, by mark and correlation – Narrow Unit, Wide Unit

three grades), conditional on mark. Figure 4 above shows the proportion of candidates inconsistently classified by at least two grades, for both units.

As expected the graphs are generally the reverse of Figure 3, with the peaks on or around the boundaries; inconsistent classification is more likely on mark points close to the boundaries.

Differences between the units

The effect of altering the correlation between the two markers has been shown to be significant. A reduction in the correlation substantially reduced the proportion of candidates consistently classified and increased the severity of the inconsistent classification. We now consider the differences between the narrow and wide units.

Figure 5 shows the proportion consistently classified, and the proportion classified within one grade and within two grades for the narrow and wide units:

There was virtually no difference in terms of the proportion consistently classified, with the indices for the wide unit very slightly higher at high levels of correlation and the indices for the narrow unit very slightly higher at lower correlations. This was not what might have been anticipated since the wide unit had grade boundaries that were more spaced apart than the narrow unit and thus we expected less inconsistent classification. The reason for the similarity is the difference in the relative mark distributions of the units (see Table 3). The proportions in each grade were different and the standard deviation of the wider unit was larger (9.60 compared to 8.01) and so the distribution was also more spread out.

Where differences did occur between the subjects these were in the severity of the inconsistent classification. Figure 5 shows that the proportion of candidates classified within one grade and within two

3 We have only included four of the levels of correlation in these graphs so that they remain legible.

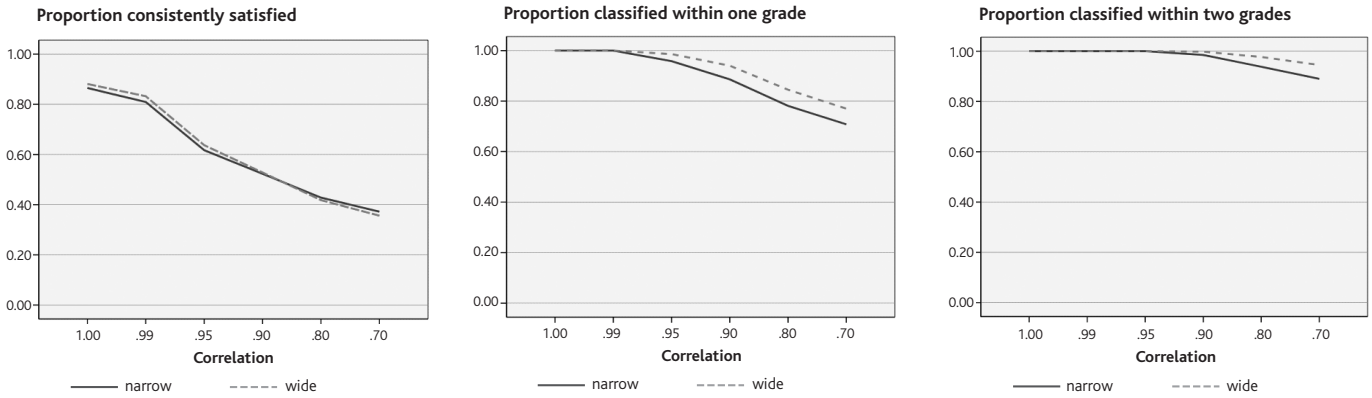


Figure 5: Comparison of levels of consistent classification between units

grades were both lower for the narrow unit. Hence the inconsistent classification in the narrow unit tended to be more severe. For example, in Table 5 we note that at a correlation of 0.90, 11.3% of candidates on the narrow unit were inconsistently classified by more than one grade compared with 6.1% of candidates on the wide unit. At a correlation of 0.80 these values were 21.8% and 15.6% respectively, at 0.7 they rose to 29.3% and 22.9%. Thus, in this example the overall effect of having more widely spaced grade boundaries was to reduce the severity, if not the amount of inconsistent classification.

Aggregation

The above analysis was at unit level. The candidate's overall grade at AS-level is of course based on the sum of the marks from each of three units. Thus the impact of inconsistent classification in any one unit is diluted by performance in the other two. However, inconsistent classification in the other units will also impact on the overall grade, and this could be compensatory to a candidate or it could make things worse. We used simulated data to investigate the impact of inconsistent classification on overall AS grade[4].

The 'wide' unit used above is an AS unit, and so we combined this with two other AS units in the same subject. There was some choice of units, but we chose the most popular units taken by those who took the original unit. We began by generating normally distributed data from the two markers on the first unit as above. We then used the real level of correlation in marks between each pair of pairs (units 1–2, 1–3 and 2–3) to simulate data for these other units, which were also normally distributed. Un-standardising each of these distributions (using the real means and standard deviations) gave a potential mark for each candidate on each unit. For the purposes of aggregation we then converted these to UMS[5]. Thus we had a mark for unit 1 by marker x (M_{1x}), a mark for unit one by marker y (M_{1y}), a mark for unit two (M_2) and a mark for unit three (M_3). For simplicity we started by assuming that there was only one marker on units two and three, so there was only one potential mark on each. The possible overall marks were thus:

$T1 = M_{1x} + M_2 + M3$

$T2 = M_{1y} + M_2 + M3$

From this the relative grades awarded under marker x and marker y, and thus the level of inconsistent classification, were estimated at each level of correlation.

We extended this analysis further by introducing inconsistent classification in unit two as well as unit one. So the totals we were interested in were:

$T_1 = M_{1x} + M_{2x} + M_3$

$T_3 = M_{1y} + M_{2y} + M_3$

T_1 is the total if units 1 and 2 were both marked by marker x and T_3 is the total if units 1 and 2 were both marked by marker y. We could then look at the proportion of candidates who would be consistently classified if not just one, but two of their units were marked by different examiners.

We used the same method as above, but just added another set of marks for the second unit and with a certain level of correlation in marks between marker x and marker y.

Finally, we introduced a second marker in the third unit, giving:

$T_1 = M_{1x} + M_{2x} + M_{3x}$

$T_4 = M_{1y} + M_{2y} + M_{3y}$

This time we were interested in the differences between T_1 and T_4 and the question became: what proportion of candidates would be consistently classified if all three of their units were marked by different examiners?

The results of the simulations are shown in Table 6 with the proportion consistently classified in terms of aggregated grade, compared with the consistent classification at unit level. The pairs of marks in each unit have the same correlation across all units.

Table 6 : Consistent classification in aggregated grade

Correlation	*P_0 (unit)*	*P_0 (aggregated, different markers unit 1)*	*P_0 (aggregated, different markers units 1 & 2)*	*P_0 (aggregated, different markers units 1, 2, 3)*
0.995	0.881	0.944	0.922	0.906
0.99	0.832	0.925	0.896	0.875
0.95	0.637	0.838	0.769	0.738
0.9	0.528	0.770	0.700	0.647
0.8	0.418	0.685	0.606	0.544
0.7	0.356	0.624	0.529	0.482

It is clear that the impact at aggregate level was much less than at unit level. As we suggested above, inconsistent classification in one unit is diluted when aggregated over the three units. In our simulation there was

4 This also applies to overall A-level grade, but it was simpler to use an AS level as an example, as this consists of only three units.

5 UMS=Uniform Mark Scale. See http://www.ocr.org.uk/learners/ums_results.html for a brief explanation. Note that in our example, the first unit had a maximum UMS of 120, whilst units 2 and 3 had a maximum of 90. Thus, the effect of misclassification of the first unit on aggregated grade is slightly greater than if all the units had equal weighting in terms of UMS.

also some 'averaging out' over the three units so that the potential levels of inconsistent classification at aggregate level were less than at unit level even if all three units were marked by different examiners. Thus at a correlation of 0.95 the potential inconsistent classification on one unit was 36.3%, compared to 26.2% at an aggregated level.

We have seen the effect of changes in the level of correlation between markers and the spread of the grade boundaries on the level of inconsistent classification, and also investigated the inconsistent classification at aggregate level. But what might this mean in reality for the number of pupils who would receive a different grade dependent on their marker? We estimated this using the levels of correlation from previous research.

There has been relatively little published research into marking reliability in UK exams. Murphy (1978, 1982) reported correlation coefficients of between 0.73 and 0.99 in 20 O-level and A-level subjects. As expected, subjects with more essay type questions such as English, History and Sociology tended to have lower levels of correlation than Maths and Physics papers, which are generally all short answer questions. Where more than one paper in each subject was investigated the aggregated marks generally correlated better than the marks on the individual papers. The correlations for the short answer questions varied from 0.98 to 1.00, whilst for the longer answer and essay type questions they varied between 0.73 and 0.98 with a mean correlation of 0.86.

More recently, Newton (1996) looked at correlations for Maths and English GCSEs. He reported correlations of above 0.99 for Maths and between 0.85 and 0.90 for English.

The two units in this research were quite different in that the paper for the narrow unit consisted of short answer questions and the paper for the wide unit was essay questions only. Thus if we arbitrarily allocate a correlation of 0.99 to the narrow unit and a correlation of 0.90 to the wide unit, we can estimate the potential levels of inconsistent classification. We should point out that this is not to suggest that these are the true levels of inconsistent classification, which cannot be known without blind double-marking, they are merely the levels that *might* exist, *if* the correlations were as stated. From Table 4, the percentage potentially inconsistently classified on the narrow unit was 19.1%, and the percentage for the wide unit was 47.2%. In other words, almost half of the students on the wide unit could potentially get a different grade dependent on the marker. Even on the narrow unit, where the level of inter-marker correlation is expected to be very high, up to one fifth of the candidates may be inconsistently classified.

The effect of aggregation would be to dilute the potential inconsistent classification. At the same level of correlation in the wide unit (0.90) 23% would be potentially inconsistently classified at aggregate (AS) level if one unit was marked by a different marker. This would increase to 35.3% if all three units were marked by different markers.

Conclusion

Since there is no such thing as a perfectly reliable test, there will always be a certain level of misclassification and/or inconsistent classification in tests and examinations. Exam boards go to great lengths to ensure that their procedures for marker monitoring, result checking and appeals allow all candidates the best chance of receiving the result that they deserve. However, the levels of misclassification/inconsistent classification are not well researched in relation to GCSEs and A-levels. Furthermore, it seems likely that the public underestimate the amount of measurement error that exists in these exams. If they were made aware of the true amount of error the level of trust in exam boards might be affected. Newton (2005) argues that while the level of trust may fall in the short term, there are many reasons why increased transparency about the extent of measurement error is desirable for students, policy makers, the public, and exam boards. His reasoning for this is 'it is crucial for those who use test and examination results to understand what kind of inferences can legitimately be drawn from them and what kind of inferences cannot' (Newton, 2005, p. 431). Because of the lack of understanding of measurement error, inferences might be drawn that cannot be justified. Whether or not this is the case, and whether it is likely that there will be more transparency in the future, we suggest that exam boards should be in a position to report an estimate of the amount of measurement error that exists in the qualifications they produce.

This article has presented the levels of inconsistent classification that *might* exist dependent on the marker used, based on simulating data in two A-level units, one with a particularly wide grade bandwidth and one with a narrow width. This should not be taken as evidence of the true levels of inconsistent classification in all A-level units, since each unit will have a different distribution of marks, a different grade bandwidth, and a different level of inter-marker correlation. However, this research does give an idea of the magnitude of the potential inconsistent classification, something that might come as a surprise to the general public.

Of course, there will always be a certain level of inconsistent classification since only completely objective tests will ever be free from measurement error attributable to markers. Further debate and investigation is needed into whether awarding bodies should routinely report estimates of these levels to the public. One approach would be to determine an acceptable level, and attempt to develop tests and train examiners so that this level can be attained. However, Newton (2005) argues that to define acceptable levels of accuracy is probably not realistic given the different natures of exams and the trade-offs between 'technical characteristics and pragmatic constraints'.

Alternatively, given that there will always be a level of inconsistent classification, more than one grade could be reported (Please, 1971) or confidence intervals could be reported on the mark given (Newton, 2003). Please suggested reporting grades in the following clusters; A/B, A/B/C, B/C/D, C/D/E, D/E/O, E/O/F and O/F. However, as he himself stated, this could lead to people treating A/B as the top grade, A/B/C as the next and so on, ignoring the implication that the candidate's true grade could be any of those in the group. The idea of confidence intervals is to report a range of marks within which we are almost certain the candidate's observed score will lie for a given true score. This method would give an idea of how much reliance we should put on exam results as an accurate summary of a candidate's skills in a particular area, and would therefore mean it is less likely that the results would be used to make unrealistic inferences.

Another idea would be to report for each grade an estimate of the proportion of candidates with that grade who might have received a higher or lower grade if another marker had marked the paper. As an example, Table 7 shows this for the narrow unit if the inter-marker correlation was 0.90.

Thus 27.2% of the grade B candidates might have got a higher grade from a different marker, and 40.9% might have got a lower grade. This

Table 7: Proportion of candidates getting a higher or lower grade if marked by a different marker

Observed grade	*Proportion Higher*	*Proportion Lower*
A	0.000	0.265
B	0.272	0.409
C	0.302	0.374
D	0.336	0.339
E	0.330	0.255
U	0.227	0.000

would be a relatively easy way of understanding how much reliance should be put on the results given. A table like Table 7 is a more informative version of a reliability coefficient. Like a reliability coefficient it is not a fixed property of the test, but depends on the distribution of scores, the grade bandwidth and (in this case) the inter-marker correlation. The proportions cannot be interpreted as probabilities for individual candidates, however, because this would depend on how close the individual was to the grade boundary. The proportions apply to the grade scale as a whole.

Finally, some limitations of this study should be mentioned. First, we mainly looked at levels of inconsistent classification in one unit only. In reality this may not be as important to candidates, as we have shown the effect is almost certain to be diluted when aggregating over the three units of AS. This would be even more the case when aggregating over six units of A-level. Arguably, it is at the aggregate level that any inconsistent classification is particularly serious: for example, when grades are used to create point scores for university selection. Secondly, it may be that using a normal distribution to simulate the data is not the ideal method. For instance, having to truncate the distribution at zero and the maximum mark meant losing some of the data, and may have slightly distorted the distribution. It may be that other distributions would better match the distribution of the data in reality, such as the beta binomial (see Livingston and Lewis, 1995; Lee *et al*., 2000). Finally, this research only considered inconsistent classification arising from differences in correlation between markers' scores, not differences between markers in severity or bias. Future research could address some of these issues, and widen the scope to other assessments, such as GCSEs or admissions tests.

References

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Cresswell, M. (1986). Examination grades: How many should there be? *British Educational Research Journal*, **12**, 1, 37–54.

Lee, W-C., Hanson, B.A. & Brennan, R.L. (2000). Procedures for computing classification consistency and accuracy indices with multiple categories. *ACT Research Report Series*. Available online at http://www.act.org/research/reports/pdf/ACT_RR2000-10.pdf (accessed 23 October 2006)

Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, **32**, 2, 179–198.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, **48**, 196–200.

Murphy, R.J.L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 58–63.

Newton, P.E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, **22**, 4, 405–420.

Newton, P.E. (2003). The defensibility of national curriculum assessment in England. *Research Papers in Education*, **18**, 2, 101–127.

Newton, P.E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, **31**, 4, 419–442.

Please, N.W. (1971). Estimation of the proportion of examination candidates who are wrongly graded. *British Journal of Mathematical and Statistical Psychology*, **24**, 230–238.

Vidal Rodeiro, C.L. (2007). Agreement between outcomes from different double marking models. *Research Matters: A Cambridge Assessment Publication*, **4**, 28–34.

Wiliam, D. (2000). Reliability, validity and all that jazz. *Education*, **29**, 3, 9–13.

EXAMINATIONS RESEARCH

Statistical Reports: Patterns of GCSE and A-level uptake

Joanne Emery and Carmen L. Vidal Rodeiro Research Division

Two new statistical reports have been added to the 'Statistics Reports' series on the Cambridge Assessment website (http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports):

Statistics Report Series No. 4: *Uptake of GCSE subjects 2000–2006*

Statistics Report Series No. 5: *Uptake of GCE A-Level subjects in England 2006*

Data for these reports were extracted from the 16+/18+ databases. These databases are compiled for the Department for Children, Schools and Families (DCSF) from data supplied by all the awarding bodies in England. They contain background details and national examination data for all candidates who have their 16th, 17th and 18th birthdays in a particular school year. Candidates are allocated a unique number that remains the same throughout their Key Stage tests, allowing matching of examination data for longitudinal investigations. Records are present only if the candidate has sat an examination in a particular subject, not just attended classes.

This brief article outlines some of the results from both reports.

Uptake of GCSE subjects 2000–2006

There were a total of 561,407 students that attempted at least one GCSE examination in 2000. This number increased 12% to reach 629,523 students in 2006. The average number of GCSEs taken by candidates in