Quantifying marker agreement: terminology, statistics and issues

Tom Bramley Research Division

Introduction

One of the most difficult areas for an exam board to deal with when communicating with the public is in explaining the extent of 'error' in candidates' results. Newton (2005) has discussed this in detail, describing the dilemma facing the exam boards: increased transparency about accuracy of results may lead to decreased public trust in those results and the agencies producing them. Measurement error is often conceptualised as the variability of an individual's score across a set of hypothetical replications (for a critique of the underlying philosophy of this approach, see Borsboom, 2005). In everyday language, this could be presented from the point of view of the candidate as a series of questions:

- Would I have got a different result if I had done the test on a different day?
- Would I have got a different result if the test had contained a different sample of questions?
- Would I have got a different result if the test had been marked by a different person?

I would suggest that whilst all these sources of variability (error) are inherent, it is the third one (marker variability) which is of most concern to the public and the exam board, because it seems to be the one most related to the fairness of the outcome. A great deal of effort goes into standardising all procedural aspects of the marking process and investing in marker training.

The advent of new technologies in mainstream live examinations processing, such as the on-screen marking of scanned images of candidates' scripts, creates the potential for far more statistical information about marker agreement to be collected routinely. One challenge facing assessment agencies is in choosing the appropriate statistical indicators of marker agreement for communicating to different audiences. This task is not made easier by the wide variety of terminology in use, and differences in how the same terms are sometimes used.

The purpose of this article is to provide a brief overview of:

- the different terminology used to describe indicators of marker agreement;
- some of the different statistics which are used;
- the issues involved in choosing an appropriate indicator and its associated statistic.

It is hoped that this will clarify some ambiguities which are often encountered and contribute to a more consistent approach in reporting research in this area.

There is a wide range of words which are often seen in the context of marker agreement, for example: reliability, accuracy, agreement, association, consistency, consensus, concordance, correlation. Sometimes these words are used with a specific meaning, but at other times they seem to be used interchangeably, often creating confusion. In this article I will try to be specific and consistent about usage of terminology. It will already be clear that I have chosen to use 'agreement' as the general term for this discussion, rather than the more commonly used 'reliability'. This is because reliability has a specific technical definition which does not always lead to the same interpretation as its everyday connotation (see section 3).

As might be expected, there are several aspects to marker agreement, and sometimes confusion is caused by expecting a single term (and its associated statistic) to capture all the information we might be interested in. We should be aware that different indicators might be appropriate in different situations. Some considerations which could affect our choice of indicator are listed below:

- Level of measurement – are we dealing with nominal, ordinal or interval-level data?
- Are the data discrete or continuous? (The numerical data is nearly always discrete, but sometimes it is thought to represent an underlying continuum).
- Is there a known 'correct' mark with which we are comparing a given mark or set of marks?
- Are we comparing two markers, or more than two markers?
- How long is the mark scale on the items being compared?
- Where does this marking situation fall on the continuum from completely objective (e.g. multiple-choice item) to subjective (e.g. holistic high-tariff essay)?
- Is the comparison at the level of sub-question, whole question, section, or test?
- What is the intended audience for communicating the information about marker agreement?
- What is the range of situations across which we would like to make generalisations and comparisons?

Rather than attempt an exhaustive survey of all possible combinations of the above factors, I will concentrate on a selection of scenarios which might seem to be most relevant in the context of on-screen marking.

1. Objective mark scheme, comparison at sub-question level, low[1] mark tariff (1-3 marks), known correct mark, comparing a single marker

This is probably the most commonly occurring situation. If the mark scheme is completely objective then the correct mark could be determined (in principle) by a computer algorithm. However, I would like

to include in this scenario cases where the mark of the Principal Examiner (PE) could legitimately be taken as the 'correct' mark (for example, in applying expert judgment to interpret a fairly objective[2] mark scheme). This scenario should therefore cover the situation which arises in on-screen marking applications where 'gold standard' scripts (where the correct marks on each item are known) are 'seeded' into a marker's allocation of scripts to be marked. I have arbitrarily set the mark limit for this scenario at questions or sub-questions worth up to three marks – a survey of question papers might lead to a better-informed choice.

I would suggest that the best term for describing marker agreement in this scenario is **accuracy**. This is because the correct mark is known. In this scenario, markers who fall short of the desired level of accuracy should be described as 'inaccurate'.

The most informative (but not the most succinct) way to present information collected in this scenario is in an $n \times n$ table like Table 1 below where the rows represent the correct mark and the columns represent the observed mark. The cells of the table contain frequency counts for an individual marker on a particular sub-question or question. This kind of table is sometimes referred to as a 'confusion matrix'.

Table 1 : Cross-tabulation of frequencies of observed and 'correct' marks on a 3-mark item

		Observed mark				
		0	*1*	*2*	*3*	*Row sum*
Correct mark	0	n_{00}	n_{01}	n_{02}	n_{03}	$n_{0.}$
	1	n_{10}	n_{11}	n_{12}	n_{13}	$n_{1.}$
	2	n_{20}	n_{21}	n_{22}	n_{23}	$n_{2.}$
	3	n_{30}	n_{31}	n_{32}	n_{33}	$n_{3.}$
	Column sum	$n_{.0}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

The shaded cells are those containing the frequencies of exact agreement between the observed and the correct mark.

The simplest indicator of accuracy would be the overall proportion (or percentage) of raw agreement (P_0), which is the proportion of the total frequency coming from the shaded cells.

$$P_0 = \frac{\sum_{0}^{m} n_{ii}}{N}$$

where m is the maximum mark on the question (in Table 1 $m = 3$).

However, it is likely that we might want to present more information from the data in the cross-table than can be obtained from the single statistic of overall agreement.

For example, we might be interested in whether the observed mark tended to be higher or lower than the correct mark (which might indicate a specific misunderstanding of the mark scheme), and in how far away from the correct mark the observed mark tended to be.

1 The research literature describes many extra statistical possibilities for measuring agreement with dichotomous (1-mark) items, but in the context of examination marking I do not believe there is much to be gained from treating them as anything other than instances of low-tariff items.

2 In practice there can be considerable difficulties in implementing a computer algorithm for marking 'fairly' objective questions – see, for example Sukkarieh *et al.* (2003).

This could be shown by presenting a frequency table of the differences between observed and correct mark. This essentially reduces the $n \times n$ cells in the table of frequencies to a single row of frequencies in the $(2n-1)$ diagonals of the table, as shown in Tables 2 and 3 below.

Table 2 : Hypothetical data from responses to 90 three-mark gold standard items

		Observed mark				
		0	*1*	*2*	*3*	*Row sum*
Correct mark	0	11	2	1	0	14
	1	4	18	1	0	23
	2	1	4	26	2	33
	3	1	1	3	15	20
	Column sum	17	25	31	17	90

Accuracy (overall exact agreement P_0) = (11+18+26+15) / 90 = 70 / 90 = 0.778.

Table 3 : Frequencies of differences between observed and correct mark

N	Difference	-3	-2	-1	0	1	2	3
90	Frequency	1	2	11	70	5	1	0

A table in the form of Table 3 would allow the reader to see at a glance:

- how accurate the marker was (relative proportion of cases with zero difference)
- whether the marker tended to be severe (entries with negative numbers) ...
- ... or lenient (entries with positive numbers)
- the size and frequency of the larger discrepancies.

For completeness, it would be helpful to add a column indicating the total mark for the item, and for comparisons it might be more helpful to show percentages rather than frequencies in the table, as in Table 4 below.

Table 4 : Percentages of differences between observed and correct mark

N	*Item max*	*Difference*	*-3*	*-2*	*-1*	*0*	*1*	*2*	*3*
90	3	%	1.1	2.2	12.2	77.8	5.6	1.1	0

In this form, the table shows the percentage of overall agreement in the highlighted box. For questions worth larger numbers of marks, we might decide to include the boxes either side (or two either side) of the zero difference box in calculating the indicator of accuracy (see section 2).

Is it desirable to summarise the information still further? For routine reporting the above format might still take up too much space. The obvious way to reduce the table further would be simply to summarise the distribution of differences, for example by the mean and standard deviation (SD). It may be that in practice it is difficult to get a feel for the meaning of the SD in this context, and if so the mean absolute difference from the mean (MAD) could be used instead.

Table 5 : Summary of distribution of differences between observed and correct marks

N	*Item max*	P_0	*Mean*	*SD*	*MAD*
90	3	0.78	-0.12	0.63	0.36

Obviously, the more the data are reduced, the more information is lost. Ideally as much information as possible should be preserved in order to facilitate comparisons (for example, between items worth different numbers of marks, between markers who have marked different numbers of items, etc.).

Other possible statistics

Kappa

A more complex statistic than P_0 which is often used in this situation is Cohen's Kappa (Cohen, 1960) or weighted Kappa (Cohen, 1968). This indicates the extent of agreement over and above what would be expected by chance. The problem with using it in our context is that we are not really interested in showing that our markers are better than chance at agreeing with the PE, but in finding out how far they fall short of perfection!

A second problem with Kappa is that it is influenced by both the shape of the marginal distributions (i.e. the distribution of row and column totals in the confusion matrix) and the degree to which the raters agree in their marginal distributions (Zwick, 1988). This could be controlled to some extent in a 'gold-standard seeding' scenario by ensuring that equal numbers of pupil responses worth 0,1, 2 and 3 marks were used as the seeds.

However, the verdict of Uebersax (2002a) is that Kappa is only appropriate for testing whether there is more agreement than might be expected by chance, and not appropriate for quantifying actual levels of agreement. A statistic which has attracted so much controversy in the research literature is probably best avoided if the aim is for clear communication.

Krippendorff's Alpha

A still more complex statistic is Krippendorff's Alpha (Krippendorff, 2002). This has been designed to generalise to most conceivable rating situations – handling multiple raters, different levels of measurement scale, incomplete data and variable sample size. The same problems apply as for Kappa, with the added disadvantage that the necessary computations are much more long-winded, and do not seem yet to be implemented in standard statistical packages (unlike Kappa). In my opinion it is unlikely that this single statistic could live up to the claims made for it.

Correlations

The familiar Pearson product-moment correlation would obviously be inappropriate because it requires continuous data on an interval scale. However, the Spearman rank-order correlation coefficient is also inappropriate (as an indicator of accuracy) because it measures covariation rather than agreement and could thus give misleadingly high values even when exact agreement (P_0) was relatively low. This might happen, for example, if the observed mark was consistently one mark higher than the correct mark.

Summary for scenario 1

The indicator of agreement should be called 'accuracy'.
N and P_0 should be reported as a minimum, followed by (in order of increasing amount of information):

- mean and SD of differences;
- frequency (%) distribution of differences between observed and correct mark;
- full $n \times n$ cross-table of frequencies.

2. Holistic or levels-based mark scheme, high tariff question (10+ marks), single marker compared with team leader or PE

This is another commonly occurring scenario, for example, where a team of markers has been trained to mark a particular essay question. It may be that the PE's mark has a privileged status (i.e. would be given more weight than that of a team member), but it is not necessarily true that the PE's marks are correct. This scenario could also apply where the comparison mark was taken to be the median or mode of several markers, instead of using the PE's mark.

There are several important differences with scenario 1 which need to be taken into account.

First of all, there is (often) assumed to be an underlying continuous trait of quality (or level of performance, or ability), and the responses are assumed to have a 'true' location on this trait. Each marker has their own conceptualisation of this trait, and each response is perceived to lie at a certain position on the trait, this position being a function of the true value, marker-specific effects and residual random error. There is no specifiable algorithm for converting a response (e.g. an essay) into an absolutely correct numerical mark. (This is not the same as saying that there is no rationale for awarding higher or lower marks – the whole point of a well-designed mark scheme and marker training is to provide such a rationale, and to ensure that as far as possible the markers share the same conceptualisation of the trait).

Secondly, although the trait is assumed to be continuous, the marker usually has to award a mark on a scale with a finite number of divisions from zero to the item's maximum. In this scenario with a long mark scale it is often assumed that the marks can be treated as interval-level data.

Thirdly, as mentioned above, it is also often assumed that some kind of random error (again continuous and often assumed to be normally distributed) is an inextricable component of any individual mark.

This means that (even more than with scenario 1) a single statistic cannot capture all the relevant information about marker agreement. This is because markers can differ in:

1. their interpretation of the 'true' trait (i.e. what is better and what is worse);
2. severity / leniency (a systematic bias in the perceived location of the responses on the trait);
3. scale use (a different perception of the distribution of the responses on the trait);
4 'erraticism' – the extent to which their marks contain random error.[3]

3 Conceptually, erraticism can be distinguished from differences in interpretation of the 'true' trait by considering the latter as differences *between* markers in where they perceive the response to lie on the trait, whereas erraticism is differences *within* a marker as to where they perceive the same response to lie on hypothetical replications of marking. In practice, these two are difficult to separate.

There is less likely to be a 'correct' mark in this scenario, and gold standard items are less likely to be used because of the time investment in creating and using them. However, there may well be occasions where a single marker's mark on a set of items needs to be compared with those of a senior marker (I will assume a PE for brevity), whose marks can be treated as the correct marks.

In this case it is possible to use the same approach as scenario 1, but just to concentrate on the distribution of differences between the marker and the PE. With a 15-mark item, the differences would need to be grouped into ranges – seven 'bins' seems a reasonable number[4], as shown in Table 6 below (which uses the same percentages as Table 4).

Table 6 : Example distribution of differences between a marker and the PE on a 15-mark item

N	*Item max*	*Difference*	*≤ -8*	*-7 to -5*	*-4 to -2*	*-1 to +1*	*+2 to +4*	*+5 to +7*	*≥ +8*
90	15	%	1.1	2.2	12.2	77.8	5.6	1.1	0

Again, the percentage of cases in the bin containing zero (the highlighted box) could form one indicator of agreement. It might be less appropriate to refer to this as accuracy – perhaps simply **agreement** is a better term for this kind of agreement. My suggestion for terminology for the agreement statistic in this case would be 'P_{agr1}' which would be interpreted as 'the proportion of cases with agreement between marker and PE within a range of ±1 mark'.

As in scenario 1, this distribution could further be reduced to the mean and SD of differences.

Table 7 : Mean and SD of mark differences between marker and PE

N	*Item max*	*Mean*	*SD*
90	15	-0.8	2.39

If we were prepared to assume that the differences were normally distributed (this could be checked graphically) then we could infer from the data in Table 7 that the marker was on average 0.8 (≈1) mark below the PE and that approximately 95% of the time their mark was between 6 marks below and 4 marks above that of the PE (these are the rounded mark points ± 2 SDs either side of the mean of –0.8). If we did not want to make assumptions about the shape of the distribution of differences it might be preferable to report the mode or median and the interquartile range (IQR), instead of the mean and SD.

The mean (or median) difference indicates the severity or lenience in the marker's marks, and the SD (or IQR) of the differences indicates the extent to which the marker's interpretation of the trait disagrees with the PE's and/or their degree of erraticism. Is there a way to extend this approach to assess differences in scale use between markers?

One solution is to plot the difference between marker and PE against the PE's mark, as shown in Figure 1. Any patterns in this plot would reveal differences in scale usage – for example, Figure 1 shows that this marker was on average about 1 mark severe, but less so at the low end of the

4 If the mark scheme is levels-based, there may be a natural 'bin' corresponding to the range of marks awarded in each level.

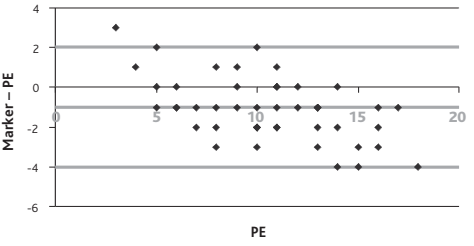

Figure 1 : Difference between marker and PE's mark (on a hypothetical 20-mark essay) plotted against PE's mark

mark range and more so at the high end. These differences could be highlighted by fitting a smoothed line to the points.

The broad lines superimposed on the plot show the mean difference, and two SDs above and below the mean. Altman and Bland (1983, 1986) recommend this kind of graph as providing the most informative visual display of differences between two measurement instruments or methods purporting to measure the same quantity[5].

It might be argued that if a plot is to be produced it would probably be easier for non-experts to interpret a simple plot of marker's mark against PE's mark (Figure 2). If the plot contained an identity line (i.e. the line representing where the marker and PE's marks would be identical) then inspection of this plot could reveal all the types of differences discussed above:

- if the points do not lie in a straight line this indicates that the marker and PE perceive the trait differently – the lower the correlation the greater this difference;
- if the points tend to lie above (or below) the identity line this indicates lenience (or severity);
- if the points tend to be above the identity line at low marks and below at high marks, or vice versa, (or if any other non-linear patterns are observed) this indicates different scale use.

Note that the dotted line in Figure 2 is not a best-fit regression line, but the identity line.

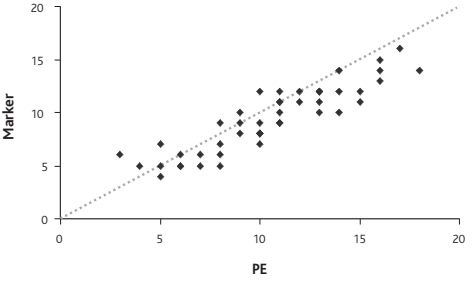

Figure 2 : Marker's mark plotted against PE's mark (on the hypothetical 20-mark essay)

5 Their paper was in a medical measurement context, where the question of interest was 'can instrument 2 be used interchangeably with instrument 1?' I would argue that this is a reasonable analogy for this scenario 2 context where we want to know the extent to which the marker is interchangeable with the PE. (They used the average of the two instrument readings for the x-axis of their graph, but I have used the PE's mark for simplicity).

Although it might be easier for non-experts to comprehend data presented in the form of Figure 2, Altman and Bland (1983) argue that plots like Figure 1 are preferable, for the following reasons:

- much of the space in a plot like Figure 2 will be empty (as this example illustrates well);
- the greater the range of marks, the greater the agreement will appear to be;
- the less highly correlated the data, the more difficult it becomes to perceive severity/lenience and different scale use by visual inspection.

A similar approach could also be used in a situation where there is multiple marking of a set of responses. Each marker could be compared against the average mark (or average of the other markers excluding their own mark) instead of against the PE. However, such situations are unlikely to arise outside a research exercise because of the costs involved in multiple marking.

Comparisons of the marker agreement statistics from this scenario with those from other situations are possible, but should be made with caution. In particular, it is important to allow for any differences in the length of the mark scale. It may also be necessary to specifically select samples of responses which cover the full mark range in order to detect any differences in scale use[6]. Comparisons will only be valid if the situations compared have used similar schemes for sampling responses.

Other possible statistics

Correlation

The correlation coefficient is a very widely used statistic, often seen in this context. It indicates the extent of linear association between two variables and thus could legitimately be used to show the extent to which the marker and PE share the same concept of what is better and what is worse. (This has been referred to as 'consistency' by some authors, e.g. Stemler, 2004). However, it cannot tell us anything about relative severity/lenience or scale use. Also, it requires there to be some variability in both sets of marks. Although in an ideal situation we would seek to ensure an appropriate range of marks, the 'mean difference' method described above does not require this. We could still produce the distribution of differences between marker and PE if all the responses had received the same mark from the PE – but the correlation between marker and PE in such a case would be zero. It should also be noted that a high value for the correlation coefficient can mask some fairly large differences – for example, the correlation in the data displayed in Figure 2 is 0.90, but Figure 1 shows that there are several cases where the absolute difference between marker and PE was three marks or more.

Regression

It is possible to summarise the data in graphs like Figure 2 by a regression equation of marker's mark (y) on PE's mark (x). This is essentially fitting the model:

$y = a + bx + e$

where a is the intercept, b is the slope, and e is random error.

The extent to which the regression line differs from the identity line could be assessed by testing whether a is significantly different from 0 and b is significantly different from 1.

This regression approach has yet to convince me of its worth. The slope parameter b confounds the correlation and the SD ratio of the two sets of marks, and both parameters might be more sensitive to sample size and outliers in the data than the simple mean of the differences would be. Also, for the results to apply more generally the responses should be sampled at random. Altman and Bland (1983) only recommend the use of regression in the context of prediction, not comparison. However, other researchers may feel that this approach has more to recommend it than I have suggested.

Summary for scenario 2

The indicator of agreement should be called 'agreement'.

The PE's mark has been used as the comparison mark in scenario 2 for brevity, but this could be replaced by the average of a group of markers in a multiple-marking scenario.

If a single indicator is to be used, P_{agrN} has been suggested, which is the proportion of scripts with a difference between marker and PE in a $\pm N$-mark range around zero. N could be increased as the total mark for the question (or sub-test or test) increases.

For fuller diagnosis of the different kinds of differences between marker and PE, the distribution of differences between their marks should be examined:

- The higher the SD, the more they perceived the trait differently, or the more their marks contained random error.
- The more positive (or negative) the mean, the more lenient (or severe) the marker compared to the PE.
- Scatter plots of the difference between marker's mark and PE's mark versus PE's mark can reveal differences in perceived distribution of responses on the trait, in addition to the above two points.

3. Reliability of marking

The previous scenarios have concentrated on methods for assessing a single marker's performance in terms of agreement with the correct mark on an objective item (scenario 1), and agreement with the PE's mark on a more subjective item (scenario 2). The term 'reliability' has been deliberately avoided. I would suggest we do not talk about the reliability of an individual marker, but reserve the term 'reliability' for talking about a set of marks. Thus reliability is a term which is perhaps best applied to an aggregate level of marks such as a set of component total scores.

The definition of reliability comes from what has been called 'true score theory', or 'classical test theory' (see, for example, Lord and Novick, 1968). The key point to note is that reliability is defined as the ratio of true-score variance to observed score variance. This very specific technical definition means that it is easy for non-experts to be misled when they read reports about reliability. Reliability refers to a set of scores, not to an individual score. The size of the variance ratio (which can range from 0 to 1) depends on the true variability in the sample. If there is no true score variance, all the observed differences will be due to error and the reliability coefficient will be zero – so the size of the reliability coefficient depends both on the test and on the sample of pupils taking the test.

Cronbach's Alpha

There are several ways of estimating test reliability, which vary depending on what the source of the errors is deemed to be. One commonly used

6 This will also help to mitigate any floor and ceiling effects when interpreting differences between marker and PE.

index of reliability is Cronbach's Alpha (Cronbach, 1951). One way of viewing this statistic is that it treats the individual item responses (marks) as repeated 'ratings' of the same pupil. The proportion of the total variance due to inter-item covariance estimates the reliability[7]. Alpha is referred to as 'internal consistency reliability' because it indicates the extent to which the items are measuring the same construct – or in other words the extent to which pupils who are above (or below) the mean on one item are above (or below) the mean on other items.

Applying the same reasoning to the situation where we have pupils with papers marked by the same set of markers, we can see that Cronbach's Alpha could be applicable here. The total scores from the different markers are the repeated ratings. The reliability of marking would be the proportion of total variance due to differences between pupils. Alpha would indicate the extent to which pupils who were above the mean according to one marker were above the mean according to the other markers – what we might term 'inter-marker consistency reliability'.

However, it is important to note that the size of this statistic would not be affected by systematic differences in severity or leniency between the markers. Adding or subtracting a constant number of marks from every value for a single marker would not change the size of Cronbach's Alpha. This type of marker consistency reliability could only be obtained from a situation where multiple markers had marked the same set of responses, and thus is likely to be more useful in research exercises than in 'live' monitoring of markers.

Intraclass correlations and general linear models

Cronbach's Alpha can be viewed as a special case of what are known as 'intraclass correlations' or ICCs (Shrout and Fleiss, 1979). These statistics are all based on analysis of variance, and are thus (in my opinion) difficult to communicate to non-specialists. Choosing the appropriate version of the ICC for the given data is of critical importance and should be done by a statistician. It is possible to choose a version of the ICC which is sensitive to differences in both consistency (correlation) and absolute agreement (von Eye and Mun, 2005). Some see this as an advantage, others as a disadvantage (Uebersax, 2003). Most versions of the ICC require double or multiple marking of the same set of responses.

Intraclass correlations themselves arise in more general linear modelling techniques such as generalizability theory (e.g. Cronbach *et al.*, 1972) and multilevel modelling (e.g. Snijders and Bosker, 1999). Approximate global indices of reliability can be derived from these more complex analyses. In fact, one of the main motivations for the development of generalizability theory was to enable the magnitude of different sources of variability in the observed score (e.g. that due to different markers) to be estimated.

Standard error of measurement

Once a reliability coefficient has been estimated it is possible to derive a standard error of measurement, SEM (see, for example, Harvill, 1991). An approximate 95% confidence interval for the observed score around a given true score is given by ±2 SEMs. These standard errors are arguably easier to interpret than reliability coefficients (which are ratios of variances) because they can be treated as distances in units of marks and thus can be compared to other meaningful mark ranges such as a grade band, or the effective range of observed scores. They are less sample dependent than the reliability coefficient, and can also be generated from generalizability theory and from Rasch (and IRT) modelling.

Multi-facet Rasch models

An alternative to the general linear model would be to fit a multi-facet Rasch model (Linacre, 1994). This approach is described by Stemler (2004) as providing a 'measurement estimate' of marker agreement, because the severities/leniencies of the markers are estimated jointly with the abilities of the pupils and difficulties of the items within a single frame of reference – reported as an equal-interval logit scale. Analysis of marker fit statistics can identify 'misfitting' markers who perceived the trait differently from the other markers. Myford and Wolfe (2003, 2004) show that it is possible to use the output from a many-facet Rasch analysis to diagnose other rater effects such as central tendency (overusing the middle categories of the mark scale), a 'halo' effect (a tendency to award similar marks to the same candidate on different questions) and differential severity/leniency (a tendency to be severe or lenient towards particular subsets of candidates).

Both these approaches (general linear models and multi-facet Rasch models) are statistically complex, generating many statistical indicators which can test different hypotheses about individual markers or groups of markers. The indicators from different analyses (i.e. on different sets of data) are unlikely to be comparable. However, both approaches can be used (with certain assumptions) in situations where the script is split into item response groups which are allocated to different markers, without the need for multiple marking of the same responses, which means that both methods are feasible options in some on-screen marking environments.

Summary for scenario 3

- The term 'reliability' should be reserved for use in its technical sense as a ratio of variances.
- Intraclass correlations are appropriate for reporting reliability, but different ICCs are applicable in different data collection scenarios, and expert statistical advice is essential.
- Where possible, it is preferable to report standard errors of measurement rather than reliability coefficients.
- General linear models and multi-facet Rasch models can diagnose many different aspects of rater agreement. Statistics generated from one set of data are unlikely to be directly comparable with those generated from another.

Conclusion

The choice of a statistical indicator of marker agreement depends on the situation and reporting purpose. I have argued that simple statistics, based on the distribution of differences between marker and correct mark, or marker and PE, are the easiest to interpret and communicate.

> *A study that reports only simple agreement rates can be very useful; a study that omits them but reports complex statistics may fail to inform.* (Uebersax, 2002b)

It will be interesting to see whether exam boards pick up the gauntlet thrown down by Newton (2005) and risk the short-term cost in terms of public trust by becoming readier to report indices of marker agreement. If they do, it will be important to choose indices which reveal more than

7 The formula for Cronbach's Alpha also contains an adjustment factor of N/(N-1) to allow it to range between 0 and 1.

they conceal. This last point is well illustrated by Vidal Rodeiro (2007, this issue) – the reader is encouraged to compare in her article tables 4 and 11 with tables 5, 6 and 12.

References

Altman, D.G. & Bland, J.M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician*, **32**, 307–317.

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–310.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cohen, J. (1968). Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213–220.

Cronbach, L.J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability of scores and profiles*. New York: Wiley & Sons.

Eye, A. von & Mun, E.Y. (2005). *Analyzing rater agreement: manifest variable methods*. Mawhah, NJ.: Lawrence Erlbaum Associates.

Harvill, L.M. (1991). An NCME Instructional Module on Standard Error of Measurement. *Educational Measurement: Issues and Practice*, **10**, 2, 33–41.

Krippendorff (2002). Computing Krippendorff's Alpha-Reliability. http://www.asc.upenn.edu/usr/krippendorff/webreliability2.pdf. Accessed January 2006.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement*. 2nd Edition. Chicago: MESA Press

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

Myford, C.M. & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, **4**, 4, 386–422.

Myford, C.M. & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 2. *Journal of Applied Measurement*, **5**, 2, 189–227.

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

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86**, 2, 420–428.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis. an introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Stemler, S.E. (2004). A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, **9**, 4. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4 November, 2005.

Sukkarieh, J.Z., Pulman, S. G. & Raikes, N. (2003). *Auto-marking: using computational linguistics to score short, free text responses*. Paper presented at the 29th conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Uebersax, J. (2002a). Kappa coefficients. http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm Accessed 22/01/07.

Uebersax, J. (2002b). Raw agreement indices. http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm Accessed 22/01/07.

Uebersax, J. (2003). Intraclass Correlation and Related Methods http://ourworld.compuserve.com/homepages/jsuebersax/icc.htm Accessed 22/01/07.

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

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, **103**, 3, 374–378.

ASSURING QUALITY IN ASSESSMENT

Agreement between outcomes from different double marking models

Carmen L. Vidal Rodeiro Research Division

Introduction

The practice of arranging for students' work to be marked by more than one person is a subject of great interest in educational research (see, for example, Cannings *et al*. 2005, Brooks, 2004, White, 2001 or Partington, 1994). However, deciding if double marking is worthwhile incorporates a perennial dilemma. Intuitively, it seems to increase the reliability of the assessment and shows fairness in marking, but this needs to be proven a benefit in order to justify the additional time and effort that it takes. Awarding bodies struggle to recruit enough examiners to mark scripts once, never mind twice, and therefore double marking of all examination papers can be a difficult task.

In the context of GCSE or GCE examinations, double marking can be a means to enhance the reliability of the marking process. One of the principal concerns of any examination board is to ensure that its examinations are marked reliably. It is essential that each examiner is applying the same standard from one script to the next and that each examiner is marking to the same standard as every other examiner. Although Pilliner (1969) had demonstrated that reliability increases as the size of the marking team increases, it was Lucas (1971) who observed that the greatest improvement came from increasing the size of the marking team from one to two and that additional benefits derived from using teams of three or more markers were of smaller magnitude.