where examiners can be monitored and supported more effectively than when marking on paper. In the present study all marking was carried out on paper, and the standardisation tasks adapted to match as closely as possible with those used operationally with online marking. Operational standardisation meetings are conducted by Principal Examiners and focus on either the short-answer examination or the essay examination, but not both. Examiners typically mark only one examination. However, the number of questions used in the study was far fewer than would be used in an operational setting.

- All participants knew that the marks did not 'count', and were only for use in the research. Whilst it is our impression that all participants were highly diligent and professional, we have no way of quantifying what effects, if any, were introduced by the low stakes nature of the exercise.

Finally, it should be noted that in operational marking settings examiners are given additional standardisation if necessary and are removed from the marking panel if their accuracy remains unsatisfactory. Additionally, examiners' operational marking is sampled on several occasions after initial standardisation, to check that accuracy levels are maintained. For these reasons operational marking is likely to be more accurate than was found in this study.

References

Baird, J., Greatorex, J. & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education*, **11**, 3, 331–348.

Greatorex, J. & Bell, J.F. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education*, **23**, 3, 333–355.

Greatorex, J., Nádas, R., Suto, I. & Bell, J.F. (2007). *Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training*. Paper presented at the ECER conference, Ghent, Belgium in September 2007.

Qualifications and Curriculum Authority (March 2009). *GCSE, GCE and AEA Code of Practice*. London: QCA.

ASSURING QUALITY IN ASSESSMENT

A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice

Milja Curcin Research Division

Introduction

Marking reliability contributes in important ways to the overall reliability and validity of assessment. It refers to the extent to which different examiners' marks agree with each other or with a definitive mark when they mark the same material (inter-marker agreement), and is also affected, for instance, by individual examiners' consistency throughout marking (intra-marker consistency). Validity of assessment is compromised without high marking reliability since the same mark from different examiners cannot be assumed to mean the same thing (e.g. Massey and Raikes, 2006; Cambridge Approach, 2009). However, as Wilmut *et al*. (1996) observe, "[f]or a variety of reasons, perfect reliability is not going to happen. The aim must be to get as close as possible, given irreducible constraints."

This review article focuses mainly on the literature relevant for the inter-marker agreement aspect of marking reliability in the context of on-screen marking. The increasing use of on-screen in place of paper-based marking presents new possibilities for monitoring of marking and ensuring higher agreement levels, but also raises questions with respect to the most efficient and beneficial use of marker agreement information that is routinely collected in this process, both in monitoring practice and in research.

Current Ofqual[1] regulations (Code of practice, April 2009) for on-screen marking require that the marking of individual examiners be compared to that of a senior examiner at regular intervals throughout the marking process. Although the specifics of this procedure differ across awarding bodies, this is generally implemented by means of "seeding" pre-marked "seeding scripts" (or items)[2] into live marking at regular intervals. The markers' marks are checked against the scripts'/items' "definitive marks",[3] these having been determined in advance by a single senior examiner or by a panel of senior examiners, depending on awarding body practices.

In this monitoring process, marker agreement data are collected at item level, potentially providing a rich source of information, particularly with respect to which features of items are associated with high or low marker agreement. Furthermore, since some awarding bodies use expert panels to decide on definitive marks, presumably under the assumption that groups make better decisions than individuals (cf. Levine and Moreland, 2006), it is conceivable that the group dynamics of these panels could affect the choice of the definitive marks and subsequent individual marker agreement with them. It is useful, therefore, to consider research to date on marker agreement, particularly at item level, as well as social psychology research on group dynamics, as this might inform both current marking monitoring processes and future research in this area, particularly in respect of what marker agreement levels can be

1 Ofqual (Office of the Qualifications and Examinations Regulator) is responsible for regulating public examinations.

2 Script: whole candidate work on one question paper. Item: candidate response on one question or question part.

3 The definitive marks are not visible on the scripts.

expected in different assessment contexts and with different assessment types.

The article first briefly reviews several studies into marker agreement at script level, focusing subsequently on research investigating finer-grained factors affecting agreement at item level, particularly with respect to marking task demands. This is followed by a brief overview of research into group dynamics and small-group decision making relevant to the group dynamics in expert panels deciding on definitive marks.

Marker agreement at script level

Most marking reliability studies conducted before the rise of on-screen marking have been conducted at whole script level, partially replicating common marking monitoring practices in paper-based marking. In several experimental studies in this context, Murphy (1978, 1979, 1982) used blind[4] re-marking to investigate mark/re-mark agreement. Overall, for nearly all of the 20 different GCE O and A-level examinations that were investigated, the correlation coefficients comparing prime and re-mark were above 0.90, except for English, where they were between 0.73 and 0.93 for individual papers (between 0.80 and 0.95 for combined papers). More recently, Massey and Raikes (2006) conducted a blind multiple-marking study on sample items taken from GCE A-Level and IGCSE examinations in a range of subjects and reported on intraclass correlations (ICCs)[5] at paper level for each subject. The ICCs they reported were in the range of 0.77 (Economics) to 0.99 (French).

The usual monitoring procedures in paper-based marking, however, involve a senior examiner re-marking a sample of each of their team's allocation of scripts at several points in the marking process, while re-marking is non-blind. Pinot de Moira *et al*. (2002, on A-level English) and Bramley (2008, on 38 different subjects) investigated mark/re-mark agreement data collected as part of such monitoring process. They found that mark/re-mark correlations generally exceeded 0.95. However, both studies acknowledge that non-blind re-marking may have boosted marker agreement. Indeed, Murphy (1979) and a number of other studies (e.g. Wilmut, 1984; Massey and Foulkes, 1994; Vidal Rodeiro, 2007) have demonstrated that inter-examiner agreement tends to be lower when the re-marking process is blind.

Importantly, most studies reviewed above report somewhat different agreement levels for different subjects. Murphy's (1978) findings also indicated that question type is an important factor, as suggested by different levels of agreement on differently structured papers within, for example, Geography O-level and English A-level, where papers with more structured questions had higher mark/re-mark correlations. This is further demonstrated in his 1982 study, where he noted that the examining technique (i.e. using essay-type vs. objective questions) tended to outweigh between-subject differences. These findings were replicated by Newton (1996) for English and Mathematics.

Clearly, investigating marker agreement at script level rather than at item level makes it difficult to separate the relative effect on marker agreement of various fine-grained factors including question type. The following section reviews studies that investigate marker agreement at item level mainly in the context of on-screen marking, which attempt to establish relative importance of these different factors and determine the operational potential and value of controlling for at least some of them in order to increase marker agreement in problematic areas.

Fine-grained features affecting marker agreement

Factors affecting marker agreement can be grouped into two general categories, depending on whether they reside in the demands of the marking task or in the marker's personal expertise (see Black, Suto and Bramley, *in submission*). The first group of factors includes item features, mark scheme features, and candidate response features. Some of the prominent factors residing in the marker include expertise, level of education and amount of training. This review will focus on the first group of factors as they are particularly relevant in the context of on-screen marking monitoring by means of seeding items in that they might inform the choice of seeding items and predictions regarding where marker agreement might be low or high.

Since in some awarding bodies (e.g. OCR[6]), the definitive marks of the seeding items are agreed by an expert panel of senior examiners, the group dynamics of these panels could be expected to interact in complex ways with factors related to the marking task and affect the choice of the definitive marks as well as subsequent marker agreement with these marks. A separate section below is therefore dedicated to an overview of research dealing with small group decision making and group dynamics.

Item and mark scheme features

One of the first studies specifically designed to investigate how different features of marking task could affect marker agreement at item level was Massey and Raikes (2006, see previous section), who investigated several surface features of items and their mark schemes (subject; maximum mark available for item; implied time restriction for candidates; type of marking: objective, points-based or levels-based; and number of levels available for levels-based marking).

Their results were mixed. Overall mean ICCs were the highest for objective items (0.97), next highest for points-based items (0.82) and lowest for levels-based items (0.77). On average, agreement decreased with rising maximum mark for points-based items, but this trend was unexpectedly reversed for Chemistry. Another interesting finding was that Sociology essay questions marked against a levels-based mark scheme were marked very reliably (average ICC=0.83, with little variation between items), indicating that it is possible to mark longer pieces of work using less constrained mark-schemes quite reliably. In general, although indicative of interesting patterns in terms of item type and other effects, these findings called for further study on larger quantities of data, and, as suggested by Suto and Nádas (2008), potentially indicate the need for a more sophisticated system of classifying questions according to marking demands.

Hudson *et al*. (2007) investigated on-screen marking reliability on seeding items for nine papers from three AQA[7] subjects. They investigated various factors, including: item type; item maximum mark; number of times the examiner had previously seen the same seed; at what time of day the marking was done. The effects of the first two factors are particularly relevant for inter-examiner agreement, and thus

4 In blind re-marking, the examiners who re-mark cannot see the original markers' marks.

5 Statistic describing how strongly units (in this case, marks) in the same group resemble each other, thus an indicator of examiner agreement.

6 OCR (Oxford, Cambridge and RSA) is one of three main awarding bodies in England.

7 AQA (Assessment and Qualifications Alliance) is one of three main awarding bodies in England.

for this review. Item type was defined in terms of whether an item could be marked by a (i) 'general' marker who was not a subject expert, or (ii) by a subject expert. Clearly, this definition conflates several item properties that could potentially be dissociated (e.g. expected response type, mark scheme properties, etc.).

Regarding item maximum mark, their findings replicate the findings elsewhere in the literature that higher tariff items tend to have higher absolute mark differences between definitive and examiner mark. However, the findings regarding item type (as defined in this study) are less clear-cut. In some subjects, the expert items tended to be associated with lower absolute mark differences, while in others this was the reverse. The authors acknowledge that there is probably a complex relationship between item type, marker expertise, marking variability and seed tolerances. Similarly to the Massey and Raikes (2006) study, it is clearly necessary to identify finer-grained distinctions when classifying item types for the purpose of marking reliability investigation.

Bramley (2008) attempted to identify some of these finer distinctions and coded a number of salient features of items and their mark schemes in order to investigate the relationship of the coded features with the level of marker agreement. The study made use of a large database of marker agreement data collected as part of the usual non-blind re-marking process at item (i.e. sub-question) level in June 2006 (OCR) and November 2006 (CIE[8]) from 38 subjects. The features coded included item maximum mark; item type (here defined in terms of whether the mark scheme was objective, points-based or levels-based); the amount of space available to the candidate to present their answer; the amount of writing required; the ratio of acceptable answers (points) allowed by the mark scheme to the number of marks available (points/marks ratio); whether the mark scheme specified qualifications, restrictions or allowable variants to the creditworthy responses; and whether the mark scheme specifically identified wrong answers.

The study used exact agreement (P_0)[9] as the measure of marker agreement, and logistic regression modelling to estimate the size and significance of the effect of coded features on this statistic. All the features were shown to be associated with marker agreement to a greater or lesser extent. However, three features were found to account for most of the explainable variance in marker agreement on objective and points-based items worth up to 9 marks. These were the number of marks available for the item, item type (objective vs. points-based), and the points/marks ratio. These features affected marker agreement in the expected direction: lower tariff, more constrained items with the number of acceptable answers equal to the number of marks had the highest agreement. In general, as Bramley observes, these findings fit the expectation that the amount of constraint in the mark scheme affects the marking accuracy and agrees with the findings of Massey and Raikes (2006) and other studies reviewed in this section.

A comparison of the relative influences of points-based vs. levels-based items did not yield clear-cut results though, that is, exact agreement was actually higher for levels-based items above 10 marks, perhaps contrary to expectation. Although this finding needs further investigation, Bramley suggests that a more 'subjective' mark scheme will not always necessarily lead to less accurate marking (cf. Massey and Raikes, 2006). Another possible explanation is that the re-marking in this study was non-blind, which may have affected the reliability patterns observed (cf. Black, Curcin and Dhawan, *in submission*, below) and also might have caused higher overall levels of agreement than would be expected in a blind re-marking situation (see previous section).

Influence of some of the above-mentioned features was also detected in the studies by Suto and Nádas (2008; 2009) investigating how examiners' thinking and their marking accuracy are affected by marking task demands defined in terms of cognitive marking strategy complexity (Greatorex and Suto, 2006; Suto and Greatorex, 2008a, b). Suto and Nádas (2008) found a strong relationship between the *apparent* cognitive marking strategy complexity (coded by researchers)[10] and marker agreement. While such findings obviously have practical implications in terms of allocating "simple-strategy" questions to general markers, and "complex-strategy" questions to expert markers, Suto and Nádas (2009) point out that it may not always be straightforward to categorise questions in terms of a relatively abstract characteristic such as marking strategy complexity.

In Suto and Nádas (2009), expert examiners used Kelly's Repertory Grid technique to identify the most influential features of questions that in their view contribute to marking strategy complexity. They identified about ten relevant features, five of which were particularly likely to demand the use of complex marking strategies and affect marker agreement: complexity of the candidate's presentation of ideas; amount of careful reading; independent vs. follow-through marks; use of words/formulae by candidate; whether the question involves application or recall of ideas; and scope/range of acceptable answers (i.e. points/marks ratio, cf. Bramley, 2008). All these features were identified as relevant for at least one subject (Biology, Mathematics or Physics) by Suto and Nádas (2008). In addition, Suto, Nádas and Bell (2009) found that the most important predictors of marker agreement for more complex strategy items were: target grade (reflecting predicted difficulty of question for candidate) and total mark (i.e. maximum mark, see for example, Bramley, 2008; Massey and Raikes, 2006).

In another study specifically designed to investigate marker agreement on seeding items[11] (Black, Curcin and Dhawan, *in submission*; see also Black, Suto and Bramley, *in submission*), data were collected on the seeding items used in the January 2009 session for five OCR units marked online in scoris®.[12] This study combined the insights from several studies cited above in terms of a comprehensive list of item/mark scheme features investigated. Most importantly, item type was defined more precisely in terms of level of constraint (objective, constrained, short answer question, extended response) while the mark scheme approach was defined separately as either objective, points-based or levels-based. Other features coded included maximum mark, definition of outcome space (whether the mark scheme specifies an exhaustive list of creditworthy responses or not), apparent marking strategy complexity (AMSC), physical answer space, whether wrong answer was specified, etc.

The features which were most strongly associated with differing levels of exact marker agreement were item maximum mark (the higher the tariff, the lower the agreement), item type (the more constrained the item, the higher the agreement), mark scheme approach (again, more constraint leads to higher agreement), definition of outcome space (the more exhaustive the outcome space, the higher the agreement), and AMSC (simple strategy – higher agreement). Thus, this study replicated

8 CIE (University of Cambridge International Examinations) – another awarding body, providing international qualifications.

9 The proportion of cases with no difference between a marker's mark and the definitive mark.

10 The categorisation of marking strategy complexity in this study was based on researcher rather than examiner judgement, hence the *apparent* marking strategy complexity.

11 Using the P_0 statistic (cf. Bramley, 2008).

12 Bespoke software for online marking, developed by RM on behalf of OCR.

some of the important findings from previous research in this domain while providing further evidence for the influence of some previously un(der)explored factors.

Black, Suto and Bramley (*in submission*) suggest that question type and mark scheme approach may be key determining factors of cognitive marking strategy complexity, which they characterise as a fundamental concept that embodies various factors affecting the demands of the marking task and consequently marker agreement. Though question type and mark scheme approach seem indeed to be relevant, there are also other factors that can potentially make an apparently simple strategy question complex to mark for any particular marker. In particular, as noted in Bramley (2008), the difficulty with applying cognitive marking strategy complexity categorisation in advance in order to predict marker agreement (e.g. by researchers, or awarding bodies) is that the actual strategy applied in each case will depend to some extent on what the candidate has actually written. Irrespective of how much constraint is placed on the outcome space, candidates can always respond in an unanticipated fashion thus potentially affecting marking task demands and subsequent marker agreement.

Candidate response features

A number of studies have investigated the features of candidate responses that potentially influence examiners' choice of marks, both in marking and grading contexts. The majority of these features appear to be 'relevant' for the construct that is assessed in any particular subject, but there are also those that may not be, but still might affect examiners' judgement and marks (e.g. Crisp, 2007).

For instance, several experimental studies detected an influence of handwriting neatness and legibility on the marks awarded, with neater responses getting higher marks. The majority of these studies were conducted in experimental settings, where teachers were marking scripts of the same content but written in different handwriting styles (e.g. Briggs, 1970, 1980; Bull and Stevens, 1979; Markham, 1976). Massey (1983), however, failed to detect a significant influence of several potentially construct-irrelevant response features on marks given in A-level English literature exams in a study using a sample of actual marked scripts. He investigated the effect of features such as untidiness, prose complexity and prose accuracy on marks awarded. He suggests that the reason why this study failed to replicate previous findings might be that the markers in earlier studies were teachers, while the markers in this study were experienced examiners. The latter, through their procedures and/or experience, might be less likely than teachers to be influenced by candidates' writing. Another possibility is that there are differences in how markers of different subjects deal with different penmanship styles, or that handwriting and style differences are less pronounced the older the candidates are (i.e. A-level vs. GCSE).

Black, Curcin and Dhawan (*in submission*) also investigated the effect of some candidate response features on marker agreement, namely spelling, communication, legibility of handwriting, crossings-out, whether the response was standard or not, and whether it was in designated response area. Spelling, legibility and quality of communication were found to have only small effect on marking agreement, corroborating to some extent the findings of Massey (op. cit.).

Response features found to be most strongly associated with P_0 in this study were whether the response was standard (associated with higher agreement); the presence of crossings out (associated with lower agreement); and whether the response was entirely in the designated response area (associated with higher agreement). The latter two effects were characterised as unexpected since they are relatively superficial aspects of responses that should not increase the demands of the marking task. If indeed replicable, the latter effect in particular should probably be taken seriously considering the preponderance of out-of-area responses in candidate scripts (cf. Whetton and Newton, 2002). Furthermore, Black, Suto and Bramley *(in submission*) report that these last three features interact with other features of the marking task, in particular question type, mark scheme approach and AMSC, increasing the demand of the marking task even for some apparently simple marking strategy questions.

Group dynamics in expert panel decisions about definitive marks

According to Suto and Greatorex (2008a, b), from a cognitive psychological perspective, the individual judgements made in examination marking may not be fundamentally different from those made in other decision-making situations. However, since the decisions about definitive marks for seeding items are sometimes made by expert panels rather than individual examiners, usually by small groups of examiners led by one most senior examiner, these decisions can be seen as additionally subject to the influence of various social factors, for example, group polarisation (Fitzpatrick, 1989), minority influence (Brennan and Lockwood, 1980), the influence of 'authority' figures or personalities, and social conformity (Murphy *et al*., 1995).

Conformity, cohesion and dissenting minorities

A number of studies have investigated the impact of majority influence or conformity in group decision-making, observing that in many cases individuals change their opinions when they find out what is the majority opinion in their group (e.g. Asch, 1951, 1956; Deutsch and Gerard, 1955), and that this can be problematic if the majority opinion is misguided. Conformity in turn can lead to group polarisation. This refers to an initially dominant position becoming more extreme or enhanced as a result of group discussion (Moscovici and Zavalloni, 1969; Myers, 1982, cited in van Avermaet, 1988) which can sometimes lead to group-think, an extreme example of group polarisation (Janis, 1972, cited in van Avermaet, 1988).

According to Kerr and Tindale's review (2004), several recent meta-analyses indicate that more cohesive groups tend generally to be more productive if their group norms favour high productivity and their group members are committed to performance goals. However, high cohesion (and/or conformity) can also cause the loss of the beneficial effects of dissenting minorities (Zimbardo and Leippe, 1991, cited in Murphy *et al*., 1995). Several studies have shown that the presence of a dissenting minority can improve the quality of group decisions through greater consideration of alternatives, divergent thinking, and integration of multiple perspectives (e.g. Moscovici, 1976). This however, depends on a number of factors, particularly in situations when there is no demonstrable correct solution to a problem under discussion, for instance, to what extent the minority members are actually aware of the superiority of their opinion or knowledge (Phillips and Lewin Loyd, 2006).

Leadership styles and group performance

In some decision-making situations, groups may be organised in such a way that multiple people provide advice to a decision maker, but the final decision is in the hands of a single person. This corresponds to the set-up of expert panels deciding on definitive marks. Kerr and Tindale (2004)

discuss a line of research dealing with these "judge-advisor systems" (e.g. Budescu and Rantilla, 2000; Sniezek, 1992, cited in Kerr and Tindale, 2004) and review a number of studies investigating how much influence the "advisors" have on the final decision of the "judges" (e.g. Harvey *et al*., 2000; Budescu *et al*., 2003). A general finding is that advisors influence judges, but judges give their own positions more weight and they also give more weight to advisors whose preferences are similar to their own, or who have been right in the past. However, the best predictor of an advisor's influence appears to be his/her (apparent) level of confidence.

Another line of research deals with leadership styles, distinguishing democratic from autocratic leadership (e.g. Lewin and Lippitt, 1938; Lewin *et al*., 1939, cited in Gastil, 1997). As summarised by Gastil (1997), in the former case the leaders encourage group decision-making and discussion, active member involvement, honest praise and criticism, and a degree of comradeship. By contrast, autocratic leaders are either domineering or uninvolved and do not consult the opinions of others. Research suggests that the interaction of leadership style with the type of task and group is particularly relevant (e.g. Fiedler, 1993, cited in Goethals, 2005; see also Gastil, 1997), with democratic leadership being apparently more productive when experimental groups are given moderately or highly complex tasks, though the link between democratic leadership and satisfaction was found not to be particularly strong or uniform (Gastil, 1994). Gastil (1993a, cited in Gastil, 1993b) also identified a number of obstacles to small group democracy, including excessive meeting length, unequal levels of commitment and involvement of different group members, clique formation and mini-consensus (formed in and/or outside meetings), differences in communication skills and styles, and intense interpersonal conflicts.

Decision-making in an educational context

Observational data from educational contexts detected a number of the above-mentioned social factors in, for instance, awarding meetings and Angoff meetings (e.g. Murphy *et al*., 1995; Brennan and Lockwood, 1980). In these studies, dominant group members were found to unduly influence the consensus opinion; there was evidence of individuals being under pressure to conform when presented with a consensus opinion; the meetings were strongly influenced by decisions taken by the Chair or by the ways in which the Chair exercised his or her role, etc. Regarding democratic (non-hierarchical) vs. autocratic (hierarchical) processes in standardisation meetings, Baird *et al*. (2004) note that, according to the questionnaire responses they collected, examiners preferred having a hierarchical discussion to having no discussion in standardisation meetings, and there was some preference for non-hierarchical rather than hierarchical discussion.

Black and Curcin (*in submission*; see also Black, Suto and Bramley, *in submission*) investigated the relationship of various group dynamics factors in expert panels deciding on definitive marks on seeding items with subsequent marker agreement. The researchers coded the discussion surrounding the decisions regarding each mark in five OCR units in terms of level of contention (which encapsulates factors such as minority influence, conformity, cohesion) and democracy levels (subsumes leadership style), as well as discussion time, and investigated these "meeting features" in relation to levels of subsequent marker agreement with the definitive marks.

While democracy was found to be related to P_0 for only two of five units under investigation and further investigation was deemed necessary, the other two features (contention and discussion time) were strongly related to P_0 for all units (higher contention and longer discussion time were associated with lower agreement). Indeed, these two features were two of the strongest single predictors of marker agreement (with similar or higher levels of prediction as maximum mark or item type) and can be seen as an expression of many of the other features that affect marking task demands. Thus, the authors suggest that these meeting features might each be thought of as a composite of the interaction of question features, mark scheme features and response features and thus might be considered as useful heuristics for prediction of subsequent marker agreement.

Conclusion

The overview given here clearly leads to a conclusion that the more objective an item and consequently the more constrained the mark scheme, the higher level of marker agreement will be achieved, though this can become complicated by, for instance, the nature of candidate response. However, marking reliability is only one of the many concerns of assessment. As Newton (1996) points out, changing the format of questions or mark schemes to increase marker agreement may threaten assessment validity as, for instance, more constrained questions may fail to measure the desired construct in some subjects appropriately. On the other hand, low marking reliability also has a negative effect on validity as the same marks given by different markers cannot be assumed to mean the same thing. More detailed and integrated knowledge of various factors that affect marker agreement which can be gleaned from item-level investigations in the context of seeding, as well as from investigations of group dynamics in expert panels deciding on definitive marks, could equip awarding bodies with an understanding of the levels of marker agreement that could be expected in different contexts and that could realistically be aspired to. This in turn could perhaps help boost reliability by improving marker agreement prediction, monitoring, feedback and training practices, without the need for resorting to over-constrained questions in inappropriate contexts.

References

Asch, S.E. (1951). Effects of group pressure on the modification and distortion of judgements. In: H. Guetzkow (Ed.), *Groups, Leadership and Men*. 177–190. Pittsburgh: Carnegie.

Asch, S.E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs*, **70**, 9, (whole no. 416), 1–70.

Baird, J-A., Greatorex, J. & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education, Principles, Policies and Practices*, **11**, 3, 333–347.

Black, B. & Curcin, M. (*in submission*). Group dynamics in determining 'gold standard' marks for seeding items and subsequent marker agreement.

Black, B., Curcin, M. & Dhawan, V. (*in submission*). Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.

Black, B., Suto, W.M.I. & Bramley, T. (*in submission*). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Paper presented at the British Educational Research Association (BERA) annual conference, Heriot-Watt University, Edinburgh, September 2008.

Brennan, R.L. & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, **4**, 219–240.

Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review*, **32**, 2, 185–193.

Briggs, D. (1970). The influence of handwriting on assessment. *Educational Research*, **13**, 1, 50–55.

Budescu, D.V., Rantilla, A.K., Yu, H.T. & Krelitz, T.M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behaviour and Human Decision Processes*, **90**, 1, 178–194. (cited in Kerr & Tindale, 2004).

Budescu, D.V. & Rantilla, A.K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, **104**, 371–98. (cited in Kerr & Tindale, 2004).

Bull, R. & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology*, **52**, 53–59.

Cambridge Assessment (2009). The Cambridge Approach. Principles for designing, administering and evaluating assessment. Available at http://www.cambridgeassessment.org.uk/ca/digitalAssets/181348_cambridge_approach.pdf Accessed 15/02/10.

Crisp, V. (2007). *Do assessors pay attention to appropriate features of student work when making assessment judgements?* Paper presented at the International Association for Educational Assessment Annual Conference, Baku, September 2007.

Deutsch, M. & Gerard, H.B. (1955). A study of normative and informational influence upon individual judgement. *Journal of Abnormal and Social Psychology*, **51**, 629–636.

Fiedler, F.E. (1993). The leadership situation and the black box in contingency theories. In: M. M. Chemers, & R. Ayman (Eds.), *Leadership Theory and Research*. 1–28. San Diego, CA: Academic. (cited in Goethals, 2005).

Fitzpatrick, A. (1989). Social influences in standard setting: the effects of social interaction on group judgments. *Review of Educational Research*, **59**, 315–328.

Gastil, J. (1997). A Definition and Illustration of Democratic leadership. In: K. Grint (Ed.), *Leadership: Classical, Contemporary and Critical Approaches*. 155–178. Oxford: OUP.

Gastil, J. (1994). A meta-analytic review of productivity and satisfaction of democratic and autocratic leadership. *Small Group Research*, **25**, 3, 384–410.

Gastil. J. (1993a). *Meeting democracy: Participation and decision making in small groups*. Philadelphia: New Society Publishers.

Gastil, J. (1993b). Identifying obstacles to small group democracy. *Small Group Research*, **24**, 1, 5–27.

Goethals, G.R. (2005). Presidential leadership. *Annual Review of Psychology*, **56**, 545–570.

Greatorex, J. & Suto, W.M.I. (2006). *An empirical exploration of human judgement in the marking of school examinations.* Paper presented at the annual conference of the International Association for Educational Assessment, 21–26 May, Singapore.

Harvey, N., Harries, C. & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behaviour and Human Decision Processes*, **81**, 52–73. (cited in Kerr & Tindale, 2004).

Hudson, G., Donahue, B.H., Rutt, S. & Schagen, I. (2007). *Is electronic marking just about efficiency? Further analysis of electronic marking data to investigate factors related to marking reliability*. DRS Data Services Limited.

Janis, I. L. (1972). *Victims of Groupthink*. Boston: Houghton Mifflin. (cited in van Avermaet, 1988).

Kerr, N.L. & Tindale, R.S. (2004). Group performance and decision making. *Annual Review of Psychology*, **55**, 623–55.

Levine, J.M. & Moreland, R.L. (2006). *Small Groups*. New York and Hove: Psychology Press.

Lewin, K. & Lippitt, R. (1938). An Experimental Approach to the Study of Autocracy and Democracy: A Preliminary Note. *Sociometry*, **1**, 3/4, 292–300. (cited in Gastil, 1997).

Lewin, K., Lippitt, R. & White, R.K. (1939).Patterns of aggressive behaviour in experimentally created "Social Climates." *Journal of Social Psychology*, **10**, 271–279. (cited in Gastil, 1997).

Markham, L.R. (1976). Influences of Handwriting Quality on Teacher Evaluation of Written Work. *American Educational Research Journal*, **13**, 4, 277–283.

Massey, A. (1983). The effects of handwriting and other incidental variables on GCE 'A' level marks in English Literature. *Educational Review*, **35**, 1, 45–50.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science national test pilot and the concept of quasi-reconciliation. *Evaluation and Research in Education*, **8**, 119–132.

Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement*. Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6–9 September 2006, University of Warwick, UK.

Moscovici, S. (1976). *Social Influence and Social Change*. London: Academic Press.

Moscovici, S. & Zavalloni, M. (1969). The group as the polarizer of attitudes. *Journal of Personality and Social Psychology*, **12**, 125–135.

Murphy, R.J.L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 1, 58–63.

Murphy, R.J.L. (1979). Removing the Marks from Examination Scripts before Re-Marking Them: Does It Make Any Difference? *British Journal of Educational Psychology*, **49**, 1, 73–78.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, **48**, 2, 196–200.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The dynamics of GCSE awarding (DOGA)*. Report of a project conducted for the School Curriculum and Assessment Authority, School of Education, University of Nottingham.

Myers, D.G. (1982). Polarizing effects of social interaction. In: H. Brandstätter, J. H. Davis & G. Stocker-Kreichgauer (Eds.), *Group Decision Making*. 125–157. New York: Academic Press. (cited in van Avermaet, 1988).

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.

Ofqual (2009) *Code of practice for GCSE, GCE, and AEA*, April 2009.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, **67**, 79–87.

Phillips, K.W. & Lewin Loyd, D. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational Behaviour and Human Decision Processes*, **99**, 143–160.

Sniezek, J.A. (1992). Groups under uncertainty: an examination of confidence in group decision making. *Organizational behavior and human decision processes*, **62**, 159–174. (cited in Kerr & Tindale, 2004).

Suto, W.M.I. & Greatorex, J. (2008a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, **34**, 1, 1–21.

Suto, W.M.I. & Greatorex, J. (2008b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practice*, **15**, 1, 73–89.

Suto, W.M.I. & Nádas, R. (2009) Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*, **24**, 3, 335–377.

Suto, W.M.I. & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*, **23**, 4, 477–497.

Suto, W.M.I., Nádas, R. & Bell, J.F. (2009). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*. (Published online to date).

van Avermaet, E. (1988). Social Influence in Small Groups. In: M. Hewstone, W. Stroebe, J-P. Codol & G. M. Stephenson (Eds.), *Introduction to Social Psychology*. 350–380. Oxford: Basil Blackwell.

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

Whetton, C. & Newton, P. (2002). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong SAR, China, September.

Wilmut, J., Wood, R. & Murphy, R. (1996). *Review of Research into the Reliability of Examinations*. A discussion paper prepared for the School Curriculum and Assessment Authority.

Wilmut, J. (1984). A pilot study of the effects of complete or partial removal of marks and comments from scripts before re-marking them. *AEB Research Report RAC315*.

Zimbardo, P. & Leippe, M.R. (1991). *The Psychology of Attitude Change and Social Influence*. New York: McGraw Hill. (cited in Murphy *et al.*, 1995).

Why use computer-based assessment in education? A literature review

Matt Haigh Research Division

Introduction

The aim of this literature review is to examine the evidence around the claims made for the shift towards computer-based assessment (CBA) in educational settings. In this examination of the literature a number of unevidenced areas are uncovered, and the resulting discussion provides the basis for suggested further research alongside practical considerations for the application of CBA.

The review looks at academic literature from UK and international contexts, examining studies that are based in educational settings from primary education to higher education. It should be noted that the literature identified predominantly emerges from higher education contexts in the UK.

Background

CBA first emerged in educational settings in the 1950s and has undergone a steady expansion in use. Burkhardt and Pead (2003) provide a useful summary of the development of CBA in educational settings for each decade between 1950 and 2000:

> *1950s: Early computers offered games, puzzles and 'tests'; compilers were designed to identify errors of syntax, and later of style, in computer programs.*
>
> *1960s: The creators of learning machines, in which assessment always plays a big part, recognised the value of computers for delivering learning programmes.*
>
> *1970s: The huge growth of multiple-choice testing in US education enhanced the attractions of automatic marking, in a self-reinforcing cycle.*
>
> *1980s: A huge variety of educational software was developed to support learning, with less emphasis on assessment.*
>
> *1990s: Along with the continuing growth of multiple-choice testing, integrated learning systems, a more sophisticated development of the learning machines of the 1960s, began to be taken more seriously.*
>
> *Since the 1990s, the explosive growth of the internet has begun to raise the possibility that testing online, on-demand might replace the traditional 'examination day' model, although many technical and educational challenges remain.*
>
> (Burkhardt and Pead 2003, p.134)

This history highlights the varying degree to which assessment has formed part of technology-facilitated pedagogy, along with the dangers of allowing technology to dictate assessment practices such as with the permeation of multiple-choice testing in the US during the 1970s detailed by Clarke, Madaus, Horn, and Ramos (2000).

The accompanying expansion in research activity can be illustrated by interrogating online-databases and filtering by year of publication as illustrated in Figure 1. This indicates that CBA developments in the mid-1990s, highlighted in the quote above, spawned a dramatic increase in the research literature available.

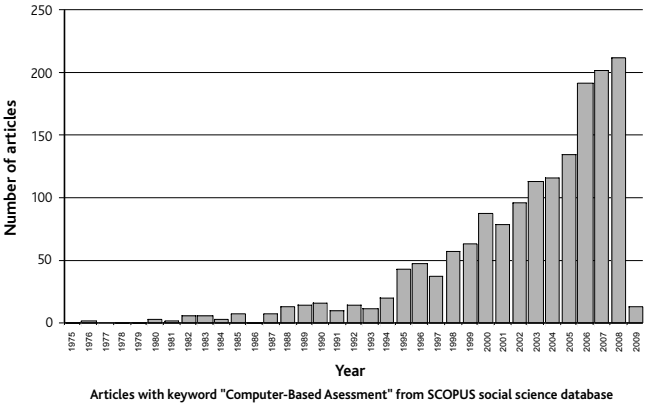

Figure 1: An illustration of CBA research activity

Note that CBA covers a broad range of assessment types, from high-stakes multiple-choice tests through to compilation of assessment evidence in electronic portfolios. This review encompasses this range, however it is quite plausible that the research discussed may only apply to a subset of these assessment types and the reader should consider this caveat throughout.