CAMBRIDGE ASSESSMENT

The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement

A paper presented at the European Conference for Educational Research, Helsinki, Finland, August 2010.

Irenka Suto, Tom Bramley, and Beth Black
E-mail: suto.i@cambridgeassessment.org.uk

Abstract
General description: In England, most secondary school students participate in a system of externally designed and assessed qualifications which affect both their entry into higher education and their employment prospects. Examination papers are marked by professional markers rather than by the students' own teachers. Given the high-stakes nature of these examinations and the popularity of international equivalents in over a hundred countries worldwide, it is important that they can be marked accurately. A key research question is therefore that of what determines marking accuracy. In this paper we present an evidence-based framework for considering many of the factors affecting marker accuracy in written examinations. The framework is developed from the perspective of researchers at an assessment agency which has responsibility for providing robust assessments in England and internationally.

The unit of our analysis is the level of agreement between the marks awarded to an examination question by individual markers and the mark deemed to be 'correct' or 'definitive' according to whatever systems are in place to monitor or control marking quality. Marker agreement has been conceptualised as the product of personal expertise and task demands (Suto and Nàdas, 2008). That is, for any given examination question, agreement can be maximised either through improving the marker's expertise or though reducing the demands of the marking task. Other factors identified as influential in the research literature can be grouped according to which of these two broad routes they are most likely to contribute to. For example, a recent empirical study of personal expertise by Suto, Nàdas and Bell (2009) indicated that the amount of training required by a marker to reach a minimum level of agreement with the 'definitive' marks was found to be the best predictor of subsequent marker agreement. The next strongest 'personal expertise' predictor was found to be a marker's highest level of general education, followed by his or her highest level of subject-specific education, followed by teaching experience, then marking experience.

In the present paper, we analyse on the other part of the equation: the demands of the marking task. A logical analysis of the demands of the marking task suggests a grouping of core features comprising: (i) question features; (ii) mark scheme features and (iii) examinee response features. During a marking event, these are the features that provide the key information which informs the marker's judgement. In describing the

construction of our framework, we synthesise several strands of empirical research in which we have investigated how core features co-vary with and influence marking agreement. We also present new data from a recent study to illustrate some effects of these interactions. We intend this paper to make a major contribution to an under-researched area.

Methods: The study entailed observing the social dynamics of the process of creating 'definitive' marks for the examinee responses used to monitor professional markers during on-screen marking sessions. These examinee responses are called 'gold standard items'. Researchers focused upon marker agreement of the gold standard items of five secondary level examinations in diverse subjects. Two observers attended five meetings in which definitive marks for 2,025 gold standard items were determined.

For each gold standard item, the observers noted: (i) discussion time; (ii) contention level, coded on a five-point scale – the degree of difficulty in agreeing a definitive mark due to differences in opinion among panel members; and (iii) democracy level, coded on a five-point scale – the degree to which views of panel members were encouraged, allowed and discussed. Each item was also coded for selected question features, mark scheme features and response features, and marker agreement data was collected.

Outcome: Statistical analyses indicated discussion time and contention level to be strong predictors of subsequent marker agreement on gold standard items for all five subjects, whereas the democracy level was a significant predictor of marker agreement in some subjects but not others. Significant main effects on marker agreement and interactions were identified among many of the question, mark scheme and response features coded, including maximum mark and apparent cognitive marking strategy complexity.

We use our framework of factors influencing marking task demands to argue that the two most important goals of research into marking accuracy are *prediction* and *control*. That is, we need to be able not only to predict the level of agreement between a marker's mark and the 'definitive' mark, but also to anticipate the effects of different possible courses of action on the level of agreement. We conclude with a discussion of the place of marker agreement in the wider context of assessment validity and public trust in examination results.

Keywords: assessment, examinations, marking accuracy, reliability, marker agreement

References:

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