CAMBRIDGE ASSESSMENT

Investigating and reporting information about marker reliability in high-stakes external school examinations

Tom Bramley & Vikas Dhawan

Abstract of presentation at the annual European Conference on Educational Research (ECER), Berlin, Germany, September 2011.

Research Division
Cambridge Assessment
1, Regent Street
Cambridge
CB2 1GG

Bramley.T@cambridgeassessment.org.uk
Dhawan.v@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Investigating and reporting information about marker reliability in high-stakes external school examinations

Tom Bramley & Vikas Dhawan

The reliability of marking (scoring) in high-stakes assessment is arguably of more concern to the examinees than other aspects of reliability. That is, they are more willing to accept that they might have scored better or worse with different questions, or on a different day, than they are to accept that the performance (script) they produced might have received a different score if it had been marked by a different person. The research reported here aimed to find the most appropriate way to conceptualise and quantify marker reliability in externally-marked school examinations of the kind taken in England at age 16 (GCSEs) and age 18 (A levels). These examinations contain a wide range of question (item) types from highly constrained objective questions to open-ended essays. Although the data came from a single Awarding Body in England, the issues involved in monitoring marker reliability affect all Awarding Bodies and should be relevant to assessment systems in other European countries.
The appropriateness of different frameworks for conceptualising reliability (Classical Test Theory and Item Response Theory) is considered in the light of the operational procedures that are used in practice for monitoring the quality of marking. Some examinations are marked on-screen and monitoring is achieved by 'seeding' scripts for which the 'correct' or 'definitive' marks are known into each examiner's allocation of scripts to be marked. The ultimate aim of monitoring processes is to ensure that the final grades awarded to examinees reflect as accurately as possible their performance in the examination. The purpose of this research was to find ways of presenting information about marking reliability based on the data collected in 'live' examinations processing (as opposed to a research exercise) that are clear, informative, and allow fair and relevant comparisons to be made between examinations in different subjects.

The data was made available by one of the three Awarding Bodies in England. It came from the live examination session in June 2009. The analysis focussed on the distribution of differences at whole script level between the definitive mark and the examiner's mark in 21 selected examination components. This included variance components analysis that attempted to quantify in some way whether the differences between awarded mark and definitive mark arose mainly because markers differed systematically in their levels of severity, or because seed scripts differed systematically in how severely or leniently they were marked.

Our main findings were: i) on average, markers tended to be neither severe nor lenient compared to the definitive mark; ii) systematic differences in severity among markers made the smallest contribution to score variability – less than systematic differences among seed scripts and much less than random error; and iii) marker-related variability in scores was relatively less than test-related variability as quantified by the standard error of measurement (SEM) calculated via Cronbach's Alpha. A limitation of our findings was that the examinations marked on screen were not a representative selection of all examinations – those containing open-ended essays were under-represented.
We found that much of the published research on marking reliability of GCSEs and A levels had presented little or no information about the simple distribution of differences between mark and definitive mark, and had instead used correlation coefficients, or more complicated statistical analyses that were less readily interpretable. We concluded that various graphical displays of the distribution of differences are the most informative and useful way to present information about marker reliability.

Main references

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

Black, B., Suto, I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice, 18(3)*, 295-318.

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

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

Bramley, T., & Dhawan, V. (2010). *Estimates of reliability of qualifications*. Coventry: Ofqual. http://www.ofqual.gov.uk/files/reliability/11-03-16-Estimates-of-Reliability-of-qualifications.pdf Accessed 23/5/11.

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Assessment Europe. Ofqual/10/4709.

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

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability.* Manchester: AQA.

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

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

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

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

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