CAMBRIDGE ASSESSMENT

Estimates of reliability at qualification level for GCSE and A level examinations

Tom Bramley and Vikas Dhawan*
Research Division, Cambridge Assessment
*Dhawan.V@cambridgeassessment.org.uk

Paper presented at the British Educational Research Association annual conference, University of London Institute of Education, September 2011.
SIG: Assessment

Abstract

This work was part of a larger study commissioned by Ofqual as part of their Reliability Programme. Reporting of reliability (most often internal consistency reliability), is a standard requirement for publishers of psychological tests. Would such a practice also be appropriate for school examinations of the kind taken in England (GCSEs and A levels)? Now that many components of such examinations are marked on-screen, the data necessary for reliability calculations is becoming routinely available. The focus of this study was to investigate ways in which an estimate of the internal consistency reliability of a GCSE, AS or A-level assessment could be derived. These assessments are almost without exception comprised of several components or units. Although the educational measurement literature provides several formulas for calculating composite reliability from the reliability of individual elements, these cannot be applied directly to GCSEs and A levels because of the complexity of the assessment structures, particularly for modular or 'unitised' assessments. The amount of choice available to examinees in which units they take and when they take them makes it difficult to define what 'the' composite might be. No information is generally available on the reliability of units/components that have not been externally assessed, such as coursework or practical examinations, meaning that the reliabilities of these can, at present, only be estimated. The non-linear weightings introduced by use of the Uniform Mark Scale (the single scale on which the results of modular assessments are aggregated) and the possibility of re-sits further complicate the picture. We drew on the literature of Classical Test Theory (CTT) and Item Response Theory (IRT) to estimate the composite reliability of four assessments. We found that while it was just about possible to derive a meaningful estimate of composite reliability under CTT, the assumptions needed to derive an IRT estimate were not plausible and the outcomes were hard to interpret. However, in all cases the reliability of the composite was higher than the reliability of the elements comprising it, as theory would expect. Neither of the above approaches took account of the fact that GCSE and A level outcomes are reported on a grade scale. We created a new index (the ratio of the grade bandwidth to the standard error of measurement) which we argue can allow for more meaningful comparisons between different assessments. We discuss the issues around interpreting indices of reliability for GCSEs and A levels.