CAMBRIDGE ASSESSMENT

Maintaining standards by expert judgment of question difficulty

Tom Bramley
Frances Wilson

Paper presented at the AEA-Europe annual conference
Glasgow, Scotland, 4-7 November 2015.

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

Bramley.T@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

Title
Maintaining standards by expert judgment of question difficulty

Authors
Tom Bramley (Cambridge Assessment)
Frances Wilson (OCR)

Abstract
This paper describes two methods for using expert judgments about test items to arrive at a grade boundary (cut-score) on a new version of a test where none of the items has been pre-tested. The need for this frequently arises in high-stakes assessments in England and elsewhere, where the need for item security, amongst other factors, means that pre-testing is not possible. In this standard-maintaining context, three sources of evidence are often used to inform decisions about where to locate the grade boundaries: i) evidence about the ability of the cohort of examinees; ii) evidence about the difficulty of the examination; and iii) evidence about the quality of work produced in the examination. In England the first and third of these sources of evidence have been more dominant over the years: the work reported here represents an attempt to strengthen the second source. Although there is certainly research evidence casting doubt on the ability of experts to provide accurate and reliable information about difficulty based on their informed judgments about examination questions, not all of the evidence is negative. In some circumstances there can be reasonable agreement between judged and empirical difficulty, particularly when judgments of experts are pooled, and when judgments of difficulty are relative, rather than absolute. Most of the research on judgment of difficulty has been in the context of standard-*setting* methods for tests comprising objective (usually multiple-choice) dichotomous items. It is therefore an open question as to whether, and how best, expert judgment of difficulty can be used in the standard-*maintaining* context such as found in GCSEs and A levels in England for those components of the assessment where the majority of items are polytomous, short answer questions.

Both the methods reported here required item statistics from several previous versions of the test. The first method involved asking two expert judges to make Angoff-type judgments about the expected mean score to be obtained on the new items by examinees at the grade boundary, basing their judgments on the equivalent actual statistics from items on previous versions of the test that they judged to be similar. The second method only required the experts to identify items on previous test versions that they judged to be effectively identical in terms of difficulty. Smoothed empirical item characteristic curves were used to find the score on the new test where the expected score on a new item was equal to the expected score for boundary examinees on a previous item judged to be identical. The new boundary was defined as the average of these scores across all identified items. The two methods were applied to a component from an A level Chemistry examination in England. Both methods gave results that were close to the actual boundaries, but in the case of the first method this may have been fortuitous since there were quite large differences between the judges' individual results. The results from the second method were quite stable when the criteria for defining identical items were varied, suggesting this method may be more suitable in practice.

Selected references

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series, 2014*(2), 1-8.

Benton, T. & Bramley, T. (2015). *The use of evidence in setting and maintaining standards in GCSEs and A levels.* Cambridge: Cambridge Assessment.

Bramley, T. (2010). 'Key discriminators' and the use of item level data in awarding. *Research Matters: A Cambridge Assessment Publication, 9*, 32-38.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59-88.