Cambridge Assessment

Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests

Conference Abstract

Tom Bramley & Tom Benton

Presented at the annual conference of the Association for Educational Assessment - Europe
Prague
November 2017

Author contact details:

Tom Bramley
Assessment Research and Development,
Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
UK

Bramley.T@cambridgeassessment.org.uk
Benton.T@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of Cambridge University, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

How to cite this publication:

Bramley, T. & Benton, T. (2017). *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests.* Paper presented at the annual conference of the Association for Educational Assessment – Europe (AEA-Europe), Prague, 9-11 November 2017.

Abstract

Setting cut-scores representing comparable performance standards on different versions of the same test or exam is a problem faced by assessment agencies around the globe. The solutions to this problem are dictated or constrained by many factors including the stakes of the test, the need for security of items, the size and variability of the cohort of test-takers, stakeholder involvement in procedures etc. Hence while the ideal method might be statistical equating based on large randomly equivalent groups of test-takers, in practice standard setting methods involving the judgment of experts are often used.

If standard setting is conceived as a process whereby an abstraction (the performance standard) is made concrete as a cut-score on the raw score scale of a real test (e.g. Cizek & Earnest, 2015), then carrying out a standard-setting exercise on two tests is conceptually closely related to IRT true-score equating where score points on two tests corresponding to the same latent trait location are deemed equivalent. Popular standard setting techniques like the Angoff method rely on the ability of a group of experts to estimate the difficulty of the items in an examination. Can their judgements provide enough information about the relative difficulty of two tests to be used for equating?

In this paper we first show, using PISA data, that within a single test the average correlation between facility value based on just 3 students and facility value based on the full sample is usually higher than the average value of 0.6 reported in Brandon (2004) as being typical for Angoff exercises. This suggests that even a small sample of data from actual examinees may provide a better basis for ensuring equivalent standards are set on two tests than any procedure based on expert judgment.

We then compare, using simulations, the accuracy as measured by root mean square equating error (RMSE) of linking cut-scores by small-sample equating (using the chained linear method with a non-equivalent groups + anchor test (NEAT) equating design) with the accuracy of linking using simulated expert judgment. For the equating, the simulations used 90 examinees but varied whether i) they were a simple random sample or were clustered within schools in three groups of 30; and ii) the strength of the anchor – a 'strong' anchor being a test of the same kind of items as the tests being equated, and a 'weak' anchor being a covariate measure of general ability. For the simulated judgments we varied the correlation between judged and true difficulty, using a 'realistic' value of 0.6 and an 'optimistic' value of 0.9. We used two cut-scores, one low and one high, to evaluate the linking.

As expected, for the standard-setting method more accurate equating arose from a higher level of correlation between simulated expert judgments of item difficulty and empirical difficulty. For small sample equating with 90 examinees per test, more accurate equating arose from: i) using simple random sampling compared to cluster sampling at a given sample size; and ii) using a stronger rather than a weaker anchor. The simulations based on the more realistic value for the correlation between judged and empirical difficulty (0.6) were worse (higher RMSE) than the small-sample equating with random sampling and a strong anchor. The simulations based on the optimistic correlation of 0.9 had the lowest RMSEs of all.

References

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

Cizek, G. J., & Earnest, D. S. (2015). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 212-237). New York: Routledge.

Full paper

Bramley, T. & Benton, T. (2017). *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests*. Manuscript submitted for publication.