CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Evaluating the simplified pairs method of standard maintaining using comparative judgement

Conference Abstract

Tom Benton

Tim Gill

3rd November 2021

Author contact details:

Tom Benton
Assessment Research and Development,
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

tom.benton@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & 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.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division
If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Benton, T. & Gill, T. (2021, November 3rd). *Evaluating the simplified pairs method of standard maintaining using comparative judgement*. Presentation at AEA Europe conference 2021, online.

Abstract

Standard maintaining refers to any activity designed to ensure that it is no easier (or harder) to achieve a given grade or above in one year than in another. Various ways of using comparative judgement (CJ) to inform standard maintaining have been suggested in the past (see, for example, Curcin et al, 2019). This paper describes and evaluates a new method (simplified pairs) of using CJ in this context. The method is simple to implement as it does not require the estimation of script quality measures using a Bradley-Terry model and is described in more detail by Benton et al (2020).

A simplified pairs study works by asking expert judges (e.g. examiners) to compare pairs of examination scripts from two different test versions and decide which script represents the better performance in the subject being assessed. Data of this type is collected across a large number of pairs with varying mark differences between the scripts being compared. Rather than relying upon measures of script quality from a Bradley-Terry model to provide an anchor between marks on different tests versions, simplified pairs uses logistic regression to directly model the relationship between the marks awarded to each script and the decisions judges make about which is superior. The results of this model can be used to estimate the relative difficulty of the two tests. More specifically, the results allow us to identify pairs of marks where the probability of a script from either test version being selected as superior is exactly 50 per cent – that is, pairs of marks representing equivalent performance on each test version.

Since simplified pairs does not require us to estimate measure of script quality, each script need only be included in a single pairwise comparison meaning that we can improve efficiency by including greater numbers of exam scripts in any given study.

The method of simplified pairs was evaluated in three experiments designed to test its accuracy in three examination subjects (English Literature, Mathematics and Science). In each experiment, the CJ approach to estimating the relative difficulty of two tests was compared to statistical equating based upon common students. Furthermore, in order to make the simplified pairs task challenging, the different test versions within each subject were deliberately chosen to represent substantially different levels of difficulty.

For the experiment based upon English Literature, the mapping of scores on one test version to equivalent values on the other using simplified pairs was extremely similar to that identified using statistical equating. For the Mathematics experiment, the two sets of results were also close although there were some small but statistically significant differences at certain points. Nonetheless, the results were close enough to indicate that simplified pairs may provide a useful source of evidence. The results from the experiment based on the science exams were somewhat different. Not only did the results from simplified pairs diverge from statistical equating, but expert judges did not even correctly identify which of the two examinations was more difficult. A possible explanation for this is that the difference in content between the two science tests was rather greater than was the case for either of the other experiments. Regardless of the explanation, these results demonstrate the need for further experimental work before we can be confident about the use of CJ in standard maintaining in all subjects.

References

Benton, T., Cunningham, E., Hughes, S. & Leech, T. (2020). *Comparing the simplified pairs method of standard maintaining to statistical equating.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/599225-comparing-the-simplified-pairs-method-of-standard-maintaining-to-statistical-equating.pdf.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual report Ofqual/19/6575. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf.

Further reading

Benton, T. (2021). Comparative Judgement for Linking Two Existing Scales. *Frontiers in Education*. 6:775203. https://doi.org/10.3389/feduc.2021.775203.

Benton, T., Leech, T. & Hughes, S. (2020). *Does comparative judgement of scripts provide an effective means of maintaining standards in mathematics?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. https://www.cambridgeassessment.org.uk/Images/603922-does-comparative-judgement-of-scripts-provide-an-effective-means-of-maintaining-standards-in-mathematics-.pdf.