CAMBRIDGE ASSESSMENT

The reliability of Adaptive Comparative Judgment

Tom Bramley (Cambridge Assessment)
Chris Wheadon (NoMoreMarking Ltd)

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
The reliability of Adaptive Comparative Judgement

Authors
Tom Bramley (Cambridge Assessment)
Chris Wheadon (No More Marking Ltd.)

Abstract
Comparative Judgement (CJ) is a novel assessment method introduced by Alastair Pollitt (e.g. Pollitt & Murray, 1993; Pollitt, 2004; Bramley, 2007) that has received a lot of interest in both research and applied settings (e.g. Kimbell et al., 2009; Jones & Alcock, 2014). Based on Thurstone's method of paired comparisons, it allows a group of experts to create a reliable scale of perceived quality by making holistic comparative judgements about pairs of objects, which in the assessment context are usually examinees' work (scripts, artwork, performances etc.). Reliability is quantified by a scale separation reliability (SSR) statistic that is derived in exactly the same way as the person separation reliability index in Rasch/IRT analyses (e.g. Andrich, 1982). It is interpreted as the proportion of 'true' variance in the estimated scale values.

One area that has attracted attention in Comparative Judgement studies is the optimisation of the selection of pairs of items for judgement. One proposed optimisation technique is the collection of rankings of more than two objects rather than paired comparisons (Bramley, 2005). A different approach, capitalising on technological developments that allow paired comparison judgements to be made on-line, distributed to larger pools of judges, and analysed on-the-fly, is to make the judgements 'adaptive'. This technique is known as Adaptive Comparative Judgement (ACJ), described by Pollitt (2012).

The adaptivity in ACJ refers to the selection of pairs of objects to present based on the results of previous comparisons, such that objects that are similar in quality are more likely to be compared than objects a long way apart. It has been repeatedly claimed in the literature that ACJ produces very high reliability, often higher than can be obtained by conventional marking with a mark scheme (scoring rubric). Furthermore, this reported high reliability is achieved with significantly fewer comparisons than would be needed in a conventional CJ study.

This paper first of all shows, by simulation, that adaptivity can substantially inflate the SSR statistic, and that high values of SSR (above 0.8) can even be obtained from random data. The paper then explores the conditions under which adaptivity inflates the apparent reliability, and attempts to explain why. Finally, suggestions are made for other means by which Comparative Judgment studies can be optimised, and how reliability should be calculated and reported in a way that does not lead to any bias.

Download the full report here:

http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf

Selected references

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research & Perspectives, 9*(1), 95-104.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.