Rank ordering and paired comparisons – the way Cambridge Assessment is using them in operational and experimental work

Tom Bramley & Tim Oates, July 2010.

On this page we describe the method of paired comparisons and its close relative, rank-ordering. Despite early origins, these scaling methods have been introduced into the world of assessment relatively recently, and have the potential to lead to exciting innovations in several aspects of the assessment process. Cambridge Assessment has been at the forefront of these developments and here we summarise the current 'state of play'.

In paired comparison or rank-ordering exercises, experts are asked to place two or more objects into rank order according to some attribute. The 'objects' can be examination scripts, portfolios, individual essays, recordings of oral examinations or musical performances, videos etc; or even examination questions. The attribute is usually 'perceived overall quality', but in the case of examination questions it is 'perceived difficulty'. Analysis of all the judgments creates a scale with each object represented by a number – its 'measure'. The greater the distance between two objects on the scale, the greater the probability that the one with the higher measure would be ranked above the one with the lower measure.

Background

The method of paired comparisons has a long history, originating in the field of psychophysics. Within psychology it is most closely associated with the name of Louis Thurstone, an American psychologist working in the 1920s – 1950s who showed how the method could be used to scale non-physical, 'subjective' attributes such as 'perceived seriousness of crime', or 'perceived quality of handwriting'.

The method was introduced into examinations research in England in the 1990s principally by Alastair Pollitt, at that time Director of Research at Cambridge Assessment (then known as UCLES – the University of Cambridge Local Examinations Syndicate). He showed how the method could be used for scaling video-recorded performances on speaking tasks in the field of language testing (Pollitt & Murray 1993), and then went on to apply it to the perennially problematic task of comparing work produced in examinations (in the same subject) from different examination boards, or from different points in time. A detailed description and evaluation of the method's use in 'inter-board comparability studies' can be found in Bramley (2007). Rank ordering is now used extensively in the comparability research work of Cambridge Assessment, and its use in operational aspects of examinations – awarding etc – is being explored and validated. But as with all approaches, it has not and will not be adopted in specific settings without testing its suitability – principally its validity and utility. This requirement for validation is in line with the standards and criteria laid down in The Cambridge Approach.

Although the mathematical details of the method can appear quite complex to non-specialists, at heart the method is very simple, the key idea being that the more times one object 'beats' another in a paired comparison, the further apart they must be on the scale. The resulting scale values are taken to be 'measures' of whatever the comparison was based on, for example 'quality of work produced'. It is assumed that, when comparing work produced in different examinations, the experts making the judgments can allow for any differences in the overall difficulty of the questions or tasks that the examinees were required to respond to.

The main theoretical attraction of the method from the point of view of comparability of examination standards is that the individual judges' personal standards 'cancel out' in the paired comparison method (Andrich, 1978). For example, a judge with a 'severe' personal standard might think that two pieces of work were both worthy of a grade B, while a judge with a more lenient personal standard might think they were both worthy of a grade A – but the two might still agree on which of the pair was better, i.e. on the relative ordering of the two pieces of work.

Using the approach in research and assessment

In practice, the paired comparison method typically is very demanding – it can be extremely resource- and time-intensive. The issue for its deployment depends not least on reaching a judgment regarding its benefit-effort ratio in a specific context. In an effort to increase the efficiency of the process, Bramley (2005) showed how the same principles could be used to create a scale if the experts were asked to put several objects into a rank order rather than comparing just two. Using rankings of several objects allows many more comparisons to take place in the same time,

with the advantage of allowing whole mark scales to be linked, rather than just grade boundary points. This idea of using expert judgment to link the mark scales on two (or more) tests has been the subject of a great deal of research at Cambridge Assessment, leading to several conference papers and publications (see bibliography). Black & Bramley (2008) have argued that it is a better (more valid) use of expert judgment than the method that is currently used as part of the regulator-mandated grade boundary setting process in GCSEs and A levels, and that it could have a role to play in providing one source of evidence for decisions on where to set the grade boundaries. A detailed evaluation of the rank-ordering method as a method for maintaining standards, or for investigating comparability of standards, can be found in Bramley & Gill (in press).

Paired comparison / rank-ordering methods have mainly been applied to the problem of comparing or maintaining standards across different tests or examinations that have been marked in the usual way. However, a far more radical use of paired comparisons / rank-ordering has been proposed by Alastair Pollitt – as an *alternative* to conventional marking (e.g. Pollitt, 2004; further examples in bibliography). An assumption within this is that the resulting scale is, in some situations, *more valid* than the raw score scale that results from conventional marking. In this scheme, both marking and standard maintaining (setting of grade boundaries) can be carried out in a single, coherent, judgment-based process. Paired comparison / rank-order judgments of work from the same examination create the scale that replaces conventional marking. Involving some pieces of work from previous examinations can 'anchor' the scale to previous scales - and hence maintain standards. In principle – although trammelled by practical problems - work from other examinations (e.g. those from other boards) could also be incorporated to ensure comparability across facets other than time.

Prototype developments in qualifications

The E-scape project led by Richard Kimbell and colleagues at Goldsmith's University (e.g. Kimbell, 2007) is a very well-funded enterprise (≈ £1.8 million over its 3 stages so far) where rank-order approaches to marking are being incorporated at a larger scale than would be possible in most research exercises. The E-scape project is innovative in a number of ways, in particular for its use of technology and its attempts to achieve more valid assessment of creativity and the design process (within Design & Technology assessment). The assessment requires the creation of electronic portfolios of evidence, which are then assessed by experts using paired comparisons and rank-ordering via a customised on-line interface. So far it has been used to assess parts of GCSE Design & Technology, GCSE Geography fieldwork and GCSE science practicals (all in non-'live' pilot projects). It is also being used in several other contexts such as formative and peer assessment (see bibliography).

State of play

Innovation and openness to new ideas are fundamental to the core values of Cambridge Assessment, and the use of paired comparisons and rank-ordering in the assessment process appears to hold considerable potential. However, we are also committed to providing good evidence to support any innovations we introduce. As can be seen in the bibliography below, we have investigated and are continuing to investigate both the technical/statistical aspects of the methods, and the underlying psychology of expert judgment that they depend upon.

Research is needed in order to evaluate the quality of assessment outcomes based entirely on paired comparison or rank-order judgments, and to identify the circumstances in which these outcomes are 'better' than those produced by conventional marking. The assumptions, underlying processes, and operational issues associated with using paired comparison / rank-order judgments in public examinations require further scrutiny. Crucially, the judgment process moves more towards a 'black box' model of assessment – something which is contrary to the direction in which assessment has been developing. In addition, the increasing demand from schools, pupils and parents for detailed feedback on performance becomes problematic under such arrangements. In terms of validity, 'better' means making the case that the paired comparison / rank-order outcome supports more accurate and complete inferences about what the examinees know and can do in terms of the aims of the assessment. In terms of reliability, 'better' means showing that the paired comparison / rank-order outcomes are more replicable with different judges (markers) or different tasks (questions). In terms of practicality, we need to show that replacing marking with paired comparison / rank-order judgments is technologically, logistically and financially feasible. In terms of acceptability, 'better' means showing that examinees and other stakeholders are more satisfied with the fairness and accuracy of paired comparison / rank-order assessment outcomes, and the information from the assessment meets school, candidate and user requirements. In terms of defensibility, 'better' means showing that it is easier for examination boards, when challenged, to justify any particular examinee's result (which clearly could be a significant challenge for a system based entirely on judgment with no equivalent of a detailed 'mark scheme').

In conclusion, Cambridge Assessment is a sophisticated user and developer of rank ordering methods and has been, and continues to be, actively involved in research into the validity of using paired comparison / rank-ordering methods in the assessment process. Our current position is that they are best deployed in standard-maintaining contexts, when the assessments being compared are as similar as possible (e.g. examinations from the same board in the same subject in consecutive examination sessions). We are actively exploring their applicability to more general investigations of comparability and to mainstream qualifications and assessments.

Bibliography

Paired comparison and rank-ordering methods

Thurstone, L.L. (1959). *The measurement of values*. Chicago: University of Chicago Press.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 449-460.

Pollitt, A. & Murray, N.J. (1993). *What raters really pay attention to.* Language Testing Research Colloquium, Cambridge. Reprinted in Milanovic, M. & Saville, N. (Eds.) *Studies in Language Testing 3: Performance Testing, Cognition and Assessment*. Cambridge University Press: Cambridge.

Bell, J.F., Bramley, T. & Raikes, N. (1998). Investigating A-level mathematics standards over time. *British Journal of Curriculum and Assessment, 8(2),* 7-11.

Bramley, T., Bell, J.F. & Pollitt, A. (1998) Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives*, *25(2),* 1-23.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement, 6(2)*, 202-223.

Bramley, T. (2007). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards.* (pp. 246-294). London: Qualifications and Curriculum Authority.
[Contains many references to studies that have used the paired comparison method in inter-board comparability research]

Gill,T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method.* Paper presented at the British Educational Research Association annual conference, Institute of Education, London.
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186333_TG_Eng_rankorder_BERA_paper_final.pdf [Accessed 24/6/10].

Black, B. & Bramley, T. (2008). Investigating a judgmental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23(3)*, 357-373.

Bramley, T. & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work.* Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/171143_TB_BB_rank_order_Perth08.pdf
[Accessed 24/6/10].

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI / Northumbria Assessment Conference, Berlin, Germany, August 2008.
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186335_BB_Jan_June_EARLI.pdf [Accessed 24/6/10].

Greatorex, J., Novaković, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods.* Paper presented at the Annual Conference of the International Association for Educational Assessment, Cambridge, September 2008
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/174284_What_attracts_judges__attention.pdf [Accessed 24/6/10].

Raikes, N., Scorey S. & Shiell, H. (2008). *Grading examinations using expert judgments from a diverse pool of judges.* Paper presented at the Annual Conference of the International Association for Educational Assessment, Cambridge, September 2008.
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186337_Raikes__Scorey_and_Shiell__IAEA_2008_-_Grading_examinations_using_expert_judgments_from_a_diverse_pool_of_judges.pdf [Accessed 24/6/10].

Curcin, M., Black, B. & Bramley, T. (2009). *Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method.* Paper presented at the British Educational Research Association annual conference, University of Manchester, September 2009.
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/184999_BERA_paper_Curcin_Black_and_Bramley.pdf [Accessed 24/6/10].

Novaković, N. & Suto, I. (2009). *How should grade boundaries be determined in examinations? An exploration of the script features that influence expert judgments.* Paper presented at the European Conference for Educational Research, Vienna, Austria, September 2009.

Bramley, T. (2009). *The effect of manipulating features of examinees' scripts on their perceived quality.* Paper presented at the annual conference of the Association for Educational Assessment - Europe (AEA-Europe), Malta, November 2009.

Novaković, N., and Suto, I. (2010) 'The reliabilities of three potential methods of capturing expert judgment in determining grade boundaries'. *Research Matters: A Cambridge Assessment Publication* 9, 19-24.

Bramley, T. & Gill, T. (in press). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education.*

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments.* Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.
http://www.cambridgeassessment.org.uk/ca/collection1/digitalAssets/186233_TB_locating_objects_Rasch2010.pdf [Accessed 24/6/10].

"Let's stop marking exams"

Pollitt, A. and Elliott, G. (2003) Finding a proper role for human judgment in the examination system, *Qualifications and Curriculum Authority Seminar on 'Standards and Comparability',* April 2003. [online] Available at:
http://www.cambridgeassessment.org.uk/ca/digitalAssets/113954_Finding_a_Proper_Role_for_Human_Judgment_in_the_Examination.pdf [Accessed 25/02/10]

Pollitt, A. (2004) *Let's stop marking exams.* Paper presented at the 30th Annual Conference of the International Association for Educational Assessment, June, Philadelphia, USA. [online] Available at:
http://www.cambridgeassessment.org.uk/ca/digitalAssets/113942_Let_s_Stop_Marking_Exams.pdf [Accessed 25/02/10]

Pollitt, A. and Crisp, V. (2004) *Could Comparative Judgments of Script Quality Replace Traditional Marking and Improve the Validity of Exam Questions?* Paper presented at the British Educational Research Association Annual Conference, September, UMIST, Manchester, UK. [online] Available at:
http://www.cambridgeassessment.org.uk/ca/digitalAssets/113798_Could_comparative_judgments.pdf [Accessed 25/02/10]

Pollitt, A. (2009) *Abolishing marksism and rescuing validity.* Paper presented at the 35th Annual Conference of the International Association for Educational Assessment, September, Brisbane, Australia. [online] Available at: http://www.iaea2009.com/abstract/69.asp [Accessed 24/02/10]

E-scape

Davies, D. (2008) E-scape phase 3 science final report – December 2008. [online] Available at: http://www.bathspa.ac.uk/schools/education/research/docs/08-09/09-e-scape-final-report.pdf [Accessed 24/02/10]

Kimbell, R. (2007) E-assessment in project e-scape. *Design and Technology Education: An international journal* 12(2) 66-76

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., Pollitt, A. and Whitehouse, G. (2009) E-scape e-portfolio assessment. Phase 3 report. [online] Available at: http://www.gold.ac.uk/media/e-scape_phase3_report.pdf [Accessed 24/02/10]

Kimbell, R., Wheeler, T., Miller, S. and Pollitt, A. (2007) E-scape e-portfolio assessment. Phase 2 report. [online] Available at: http://www.gold.ac.uk/media/e-scape2.pdf [Accessed 24/02/10]

Kimbell. R., Brown Martin, G., Wharfe, W., Wheeler, T., Perry, D., Miller, S., Shepard, T., Hall, P. and Potter, J. (2005) E-scape e-portfolio assessment. Phase 1 report. [online] Available at: http://www.gold.ac.uk/static/teru/pdf/e-scape1.pdf [Accessed 24/02/10]

Martin, F. and Lambert, D. (2008) Report for geography 'e-scape' Initial Trial. June 2007-December 2008. [online] Available at: http://www.geography.org.uk/download/GA_PREscapeFinalReport.pdf [Accessed 24/02/10]

Martin, F. (no date) E-scape: Briefing paper for Secondary Phase Committee [online] Available at: www.geography.org.uk/download/GA_PREscapeBriefingDoc.pdf [Accessed 24/02/10]

Newhouse, P., Miller, D., Campbell, A. and Cooper, M. (2009) Digital forms of assessment. Extra report on the analysis of 2008 data. Applied Information Technology. A report for the Curriculum Council of Western Australia. [online] Available at: http://csalt.education.ecu.edu.au/downloads/AIT_Report2008.pdf [Accessed 24/02/10]

Technology Education Research Unit (2009) E-scape project information. [online] Available at: http://www.gold.ac.uk/teru/projectinfo/ [Accessed 24/02/10]

Technology Education Research Unit and CIE (undated) Video of the IGCSE English Experiment in E-assessment http://www.cie.org.uk/qualifications/academic/middlesec/igcse/video/IGCSE_english_experiment_esssessment [Accessed 24/02/10]