Adaptive Comparative Judgment for reliable on-line assessment

Proposing Organisations:
Goldsmiths, University of London and TAG Assessment (www.tagassessment.com)
Main Presenter:
Professor Richard Kimbell, Director Technology Education Research Unit, Goldsmiths, University of London.
Presenter Contact Details:

- E-Mail: matt.wingfield@tagassessment.com
- Telephone: +44 203 176 0394

Co-Presenter:
Matt Wingfield, Managing Director TAG Assessment and Chairman of The eAssessment Association.

Abstract:

Project e-scape[1] was commissioned by the UK government to solve a continuing problem with the assessment of project-based coursework. Coursework is widely appreciated as the most valid expressions of learners' capability – be that in music, sciences, languages or technology, but despite this, they have consistently proved to be difficult to assess with an acceptable degree of reliability.

Our approach to solving this problem was highly innovative. Rather than one teacher marking their own students' portfolios, all teachers collaborate in assessing all the portfolios. Making use of web-connectivity we developed an on-line adaptive methodology in which teachers make simple comparative judgements (comparing the performance in portfolio A and portfolio B). A string of such simple binary judgements, by a connected group of teachers, results [via a Rasch modelling engine] in a rank order of astonishing reliability. Typically our trials have produced reliability statistics of 0.95 or better. The software adapts to the emerging consensus of the judging team, producing the reliability for minimal investment of judging time.

Assessment projects using this technology have been conducted successfully in Western Australia, Atlanta (USA), Sweden, Israel, Ireland, the UK and most recently in Singapore with the MOE (Humanities and Art Divisions).

[1] See: http://www.gold.ac.uk/teru/projectinfo/projecttitle,5882,en.php
---

Adaptive Comparative Judgment for reliable on-line assessment

In 2004, the Technology Education Research Unit at Goldsmiths University of London was commissioned by the UK government to develop a quite new approach to assessment. The brief was from the Qualifications and Curriculum Authority ('QCA') who had responsibility for schools curriculum and assessment policy.

> "QCA intends now to initiate the development of an innovative portfolio-based (or extended task) approach to assessing Design and Technology at GCSE. This will use digital technology extensively, both to capture the student's work and for grading purposes. The purpose of Phase I is to evaluate the feasibility of the approach...'
> (QCA specification June 2004)

The resulting project 'e-scape' moved through three phases; proof of concept (2004-2005), prototype development (2005-2007), national trial (2007-2009). The concern was always with *performance assessment* to be judged by reference to portfolios of learners' work undertaken in normal classrooms, studios and workshops. There were therefore two principal areas of research & development:

i) how best to capture performance in real time in real workplaces
ii) how to make reliable assessments of the resulting portfolios

We did not believe that learners' activity would be captured principally using desktop computers, keyboards and screens. Such technologies were typically (in 2004/5) retained in special rooms – computer suites – that were remote from the REAL learning activity in any subject of study (eg science labs or art studios). We thought that peripheral, hand-held technologies would be more appropriate. At least at the 'input' level, these technologies enable activities in classrooms, workshops and studios to go ahead almost as normal. A further element of the capture problem was that we sought to develop a system that automatically uploaded any captured elements to learners' web-portfolios.

Concerning the second area of R&D, whilst coursework projects are widely appreciated as the most *valid* expressions of learners' capability – be that in music, sciences, languages or technology, they have always proved very difficult to assess with acceptable *reliability*. We therefore developed a quite new methodology for assessment that took full advantage of the web-based nature of the portfolios. Unlike paper-based portfolios, web-based ones can be easily distributed and marking teams can scrutinize them at the same time in any location. We developed an approach of adaptive comparative judgement (hereafter ACJ) linked to a Rasch modeling engine to create a system that would enable learners' web-portfolios to be assessed by teams of web-connected teachers and examiners.

Project e-scape ran from 2004-10; see http://www.gold.ac.uk/teru/projectinfo/projecttitle,5882,en.php In the national trials in 2009, 500+ learners from 20 secondary schools in England created real-time performance web-portfolios in science, geography and technology and the reliability statistics on the ACJ assessments was astonishingly strong (inter-rater statistics of 0.95 or better). The work is now being rolled out in association with national

assessment agencies in several countries – including Sweden, Australia, Singapore, USA, Israel, and Ireland.

Adaptive comparative judgement and the 'pairs engine'

The approach to assessment was designed to overcome the reliability problem by making use of web-connectivity. One of the problems with normal (paper) portfolio assessment is that teachers have to be the front-line assessors for *their own* students' work. Moreover, typically (in the UK) only a sample of the assessed portfolios is then sent for moderation – so there is a high probability that any errors in the original teachers' judgements will remain in the final award. With web-portfolios and connected teachers we can completely change this paradigm. Rather than one teacher marking their own students' portfolios, all teachers collaborate in assessing all the portfolios. We developed an on-line adaptive methodology in which teachers make simple comparative judgements (comparing the performance in portfolio A against portfolio B). A string of such binary judgements, by a connected group of teachers, results [via a Rasch modelling engine] in a rank order of astonishing reliability. In the first national trial in 2009, the modeling engine required 17 rounds of judging (each portfolio was compared with 17 others) before the rank stabilized. But thereafter, any more rounds of judging did not alter the rank. The reliability statistic at this point was 0.93 and every run of the engine on subsequent projects has yielded better reliability and 0.96 is now the norm. Moreover, that first run of the engine was four years ago and subsequently we have refined the algorithm in the ACJ engine so that currently the stabilized ranks emerge after only 9 rounds of judging.

In developing the algorithm we worked with Alastair Pollitt who had worked with the University of Cambridge Local Examinations Syndicate, and subsequently headed the research section for Cambridge Assessment (see Pollitt & Crisp 2004 & Pollit and Ahmed 2009). His report on the first run of the engine contains three important observations.

First the *reliability* of the resulting scale.

> "The key figure here is the reliability coefficient of 0.93. This figure allows for unreliability between markers *as well as* for lack of internal consistency within the examination – most traditional reliability coefficients only allow for one of these. Only a few current GCSEs are likely to be as reliable as this if we consider both sources of unreliability." (Pollitt in Kimbell et al 2007 pp51-53)

But this reliability is hardly surprising. Each piece of work has been compared with many others, and judgements have been made by many judges. Any idiosyncratic judgements are soon outweighed by the weight of opinion of the team. The process is almost inevitably more reliable than current GCSE practices.

Second it is important to note the consistency of the judges. In this ACJ approach, the analysis automatically produces a measure of the consensuality of the judging team. The system notes how often, and by how much, one judges' decisions are at variance with other judges and in the end produces a mean score for the whole sample. If I am more than two Standard Deviations from that score, then I am a cause for concern. As Pollitt reported; 'None of the judges failed this test'.

Third, the system also automatically produces data on the consensuality of judgements applied to individual portfolios. There are portfolios over which there was some of disagreement within the judging team – but these are automatically highlighted by the 'standard error' indicator attached to each portfolio. So in the process of generating the rank, the system automatically highlights the pieces of work that need closer attention.

These three features: the reliability of the scale, the consensuality measure of judges, and the identification of any portfolios that generate disagreement, are all automatic virtues of the ACJ process.

Assessment projects using this technology have been conducted successfully in Western Australia, Atlanta (USA), Sweden, Israel, Ireland, the UK and most recently in Singapore with the MOE (Humanities and Art Divisions).

References:

Qualifications and Assessment Authority (QCA) 2004 *E-assessment expert seminar* BECTa 9th Dec.

Kimbell, RA Wheeler A, Miller S, and Pollitt A. 2007 e-scape portfolio assessment (e-solutions for creative assessment in portfolio environments) phase 2 report pp 100 TERU Goldsmiths University of London

Kimbell RA (2008) "Project e-scape: a web-based approach to design and technology learning and assessment" ch 12 (pp219-241) in The episteme Reviews: Research Trends in Science, Technology and Mathematics Education. Eds Choksi B and Natarajan C,. Macmillan India Ltd. New Delhi.

Kimbell R (2009) "Performance Portfolios: problems, potentials and policy" chapter 42 in Internataional Handbook on Research and Development in Technology Education (Jones A. & de Vries M. Eds) Sense Publishers . Rotterdam NL

Pollitt A and Crisp V. (2004) "Could Comparative Judgements Of Script Quality Replace Traditional Marking And Improve The Validity Of Exam Questions?" A paper presented at the British Educational Research Association Annual Conference, UMIST, Manchester, September 2004.

Pollitt A & Ahmed A (2009) The importance of being valid A paper presented at the 10th Annual Conference of the Association for Educational Assessment – Europe. Cambridge Exam Research www.camexam.co.uk Malta, November 2009