When developing and validating assessments, what are the key issues in psychological research? Some ideas from the educational assessment community

Irenka Suto and Stuart Shaw

CAMBRIDGE ASSESSMENT

When implementing educational qualifications entailing novel structures and/or topics, evidence is generated to support and justify claims of assessment validity including reliability. A theoretical underpinning coupled with empirical investigation promotes confidence among students, teachers, and educationalists. Arguably, the research approaches used are applicable but under-utilised in other assessment contexts. Here we review some key methods used in psychological research surrounding the development and validation of educational qualifications, which may be relevant to users and developers of many other types of assessment.

'Think aloud' method

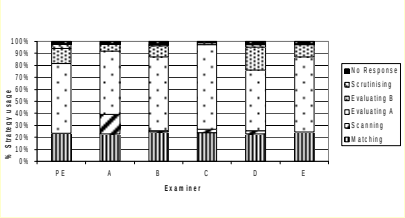
This primarily qualitative research method has a long history and can be used with research participants (e.g. examiners, teachers, students) in person or over the telephone (Greatorex and Suto, 2008; Shaw, 2008). Researchers at Cambridge Assessment have investigated examiners' cognitive marking strategies by inviting them to think aloud whilst marking students' responses to past examination questions from GCSE Business Studies and Maths (Suto and Greatorex, 2008 a, b).
Here are some typical instructions that have been read out to research participants:

"In this study, we are interested in what you are thinking to yourself whilst you are in the process of marking scripts. I will be asking you to 'think aloud' whilst marking, and I will record your thoughts using recording equipment. Beforehand, you can have a short practice at thinking aloud, which will not be recorded.
By 'think aloud', I mean that I want you to say out loud everything that you would normally say to yourself silently whilst you are marking. It may help if you imagine that you are alone.
Please speak clearly so that the recording equipment picks up everything you say. If you are silent for any period of time, then I shall remind you to keep talking by saying "Please keep talking." Please try to mark as 'normally' as you can, but do try to mention when you are looking at the mark scheme and what you are looking at. Do you have any questions?"

In a qualitative analysis of the verbal protocol data generated, Suto and Greatorex identified the following five cognitive marking strategies:

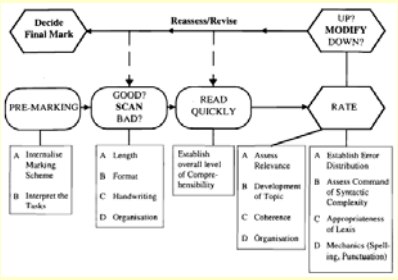
Matching	Examiner looks at a pre-determined location and compares the response with the correct answer (either held in the working memory or recollected using the mark scheme), making a simple judgement about whether the two match up.
Scanning	Examiner scans the whole of the space in the script allocated to a question, to identify whether a particular detail in the mark scheme is present or absent in the student's response
Evaluating	Examiner attends to all/part of the space dedicated to a question. S/he processes the information semantically, considering the student's response for structure, clarity, factual accuracy and logic or other characteristics given in the mark scheme.
Scrutinising	Examiner uses this only for unexpected or incorrect responses. S/he tries to identify where the problem lies and whether the response is a valid alternative to what is in the mark scheme. To do this, s/he scans and evaluates multiple aspects of the response with the overarching aim of reconstructing the student's line of reasoning or working out what s/he was trying to do.
No response	Examiner uses this when nothing is written in the answer space allocated to the question. S/he looks at the space once or more to confirm this; s/he can then award zero marks.

A quantitative analysis of marking strategy usage was conducted. This graph shows the proportions of marking strategy usage for each of six GCSE Business Studies examiners (Suto and Greatorex, 2008 b):



Notes: Examiner 'PE' is the Principal Examiner. 'Evaluating A' entails an end judgment only, whereas 'Evaluating B' entails one or more interim judgments.

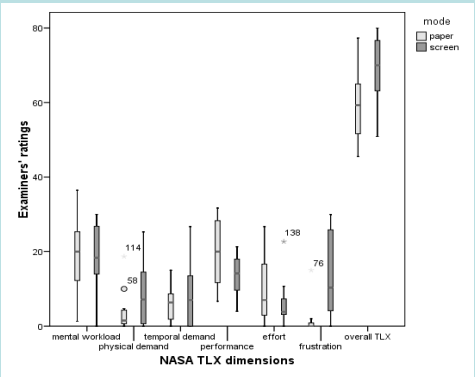
Verbal protocol analysis has also been used to identify reading strategies in the assessment of second language writing (Milanovic, Saville, and Shuhong, 1996; Falvey and Shaw, 2006; Shaw and Weir, 2007). Group interviews, introspective verbal reports, and retrospective written reports with experienced examiners of Cambridge ESOL First Certificate in English (FCE) and Certificate of Proficiency in English (CPE) compositions were conducted (Milanovic et al, 1996). They revealed four discernible approaches to composition marking:
- Principled and Pragmatic Two-Scan Approaches
- Read Through Approach and
- Provisional Mark Approach



Questioning methods

Interviews (in person and by telephone), focus groups, and questionnaires are all used to gather data on the perspectives of research participants, who include examiners, teachers, and students.

Johnson and Nádas (2009) used the NASA-TLX questionnaire with telephone interviewing to explore the relative mental workloads of examiners during paper-based and on-screen marking of GCSE English Literature essays. This graph shows the examiners' ratings of different aspects of their marking experiences:



Kelly's Repertory Grid (KRG) method

KRG technique has been used by Suto and Nádas (2009 a, b) to explore the features of questions and mark schemes that most influence marking accuracy. Studies have focused on examinations for GCSE Maths and Physics and IGCSE Biology. Following some warm-up exercises, a one-to-one meeting is held between the researcher and an expert examiner. The examiner reviews a selection of three examination questions and is asked:
'Which 2 of these 3 questions are the same in some way or some aspect, and different from the third?'
The examiner is usually able to identify a distinguishing feature (a KRG construct), such as *amount of writing by the student* or *number of acceptable answers given in the mark scheme*. The researcher and examiner use the feature to create a 5-point rating scale. The examiner then uses the scale to rate each of the questions in the examination under investigation. A 'repertory grid' is thereby completed.
Subsequently, the researcher uses the feature rating data together with marking accuracy data to identify those features that are related to accuracy. If the dataset is too small to be analysed statistically, a frequency table may be inspected for trends. In the example below, there is a trend for questions that make fewer different demands on the student to be marked more accurately than those making multiple demands:

Relative marking accuracy of examination questions	Ratings given by an expert examiner (on a 5-point scale)					Total N examination questions
	1 Question makes a single demand on the student	2	3	4	5 Question makes multiple demands on the student	
Relatively low accuracy	0	1	2	2	3	8
Medium accuracy	1	5	1	0	0	7
Relatively high accuracy	4	2	2	0	0	8
Total N examination questions	5	8	5	2	3	23

EXAMINERS' PERSONAL EXPERTISE

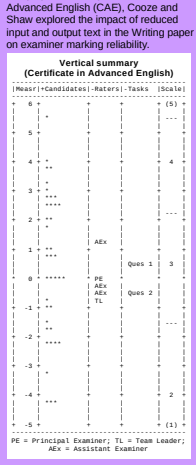
VALIDITY AND RELIABILITY OF ASSESSMENT PROCESS

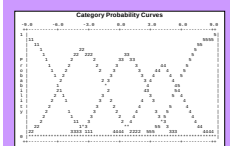
QUALITY OF ASSESSMENT INSTRUMENTS

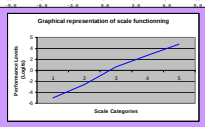
Multi-faceted Rasch methods

Examiner scores and questionnaire responses have been analysed using multi-faceted Rasch (Cooze and Shaw, 2007; Shaw and Falvey, 2008). As part of a project to modify the Certificate in Advanced English (CAE), Cooze and Shaw explored the impact of reduced input and output text in the Writing paper on examiner marking reliability.

One of the facet summary reports generated by the Rasch FACETS software is shown. The report is a visual representation of reliability data. The scale along the left represents the measure (in logits). Column 2 is a candidate ability measure (each script is shown as an asterix). Each script is ordered with the highest level of performance at the top and the lowest level at the bottom of the 'Measr' scale. The third column relates to examiner severity. Harsh examiners are at the top and lenient examiners are at the bottom. The next column relates to task (question) scoring. Again, harshly scored questions appear at the top and leniently scored questions at the bottom. The most likely scale score for each ability level is shown in the rightmost column.

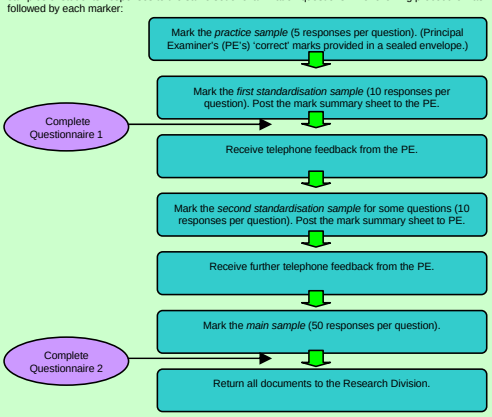




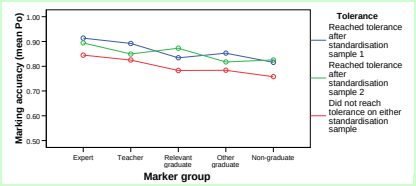


Experimental marking methods

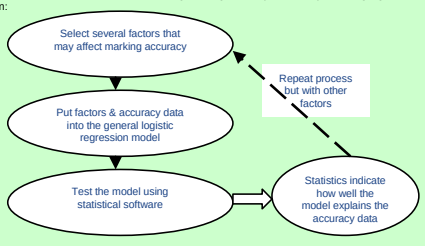
Inviting examiners to mark past students' work in experimental settings enables researchers to explore issues such as inter-examiner reliability and the relative merits of different training and standardisation procedures. Key advantages of experimental marking are that it cannot affect the results of current students, and that the same students' responses can easily be marked by multiple examiners.
Suto, Nádas and Bell (2009) conducted an experimental marking study of a past IGCSE Biology examination, in which 42 markers with diverse background experiences marked the same three samples of students' responses to the same set of examination questions. The following procedure was followed by each marker:



This graph reveals some of the influences of background experience on marking accuracy on the main sample of students' responses. The mean proportion of exact agreement between each marker and the Principal Examiner (mean Po) has been used as a measure of accuracy. Accuracy is very high in general. However, the red line indicates the accuracies of markers of all backgrounds who failed to 'reach tolerance', i.e. who did not demonstrate an acceptable level of accuracy on either standardisation sample. These markers proceeded to mark the main sample of responses the least accurately of all the markers in the study.



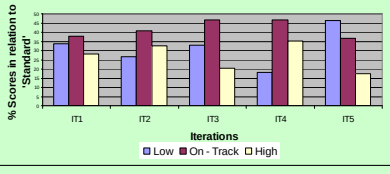
The relative influences of several factors affecting marking accuracy were analysed using logistic regression:



Comparing statistics generated by different models indicated which factors affect accuracy most.

In an experimental marking study of the Certificate of Proficiency in English (CPE), Shaw (2002) observed that an experimental iterative standardisation process of training and successively delivered feedback to examiners did not improve inter-rater reliability (but this was perhaps affected by the fact that inter-rater reliability was already encouragingly high). Shaw's study showed evidence of examiners modifying their behaviour with successive standardisation exercises producing a 'see-saw' effect.

Examiner performance over 5 iterations:
Low, high and 'on track' marking of an optional question



When developing and validating assessments, what are the key issues in psychological research? Some ideas from the educational assessment community

A poster presented at the annual conference of the British Psychological Society, 14th -16th April 2010, Stratford Upon Avon

Abstract

Purpose

We review key issues in psychological research surrounding the development and validation of educational qualifications, which are relevant to users and developers of many other types of assessment.

Background

When implementing educational qualifications and study programmes entailing novel structures and/or topics, evidence is generated to support and justify claims of assessment validity including reliability. A theoretical underpinning coupled with empirical investigation promotes confidence among students, teachers, and educationalists. Arguably, the research approaches used are applicable but under-utilised in other assessment contexts.

Methods

We conceptualise the robustness of the assessment process as the interaction of (i) assessors' personal expertise, and (ii) the quality of the assessment instruments. Key issues reviewed range from examiners' understanding of, and confidence in the assessment process, to aspects of assessment materials that most influence examiners' judgements, to aligning assessment instruments with test-takers' needs and preparedness.

We contrast some well-established methodological and statistical approaches to exploring these issues. For example, data generated in multiple marking and script judgement studies has been analysed using multiple regression and multi-faceted Rasch modelling techniques. Examiners' verbal protocol data has been analysed both quantitatively and qualitatively, whereas Kelly's Repertory Grid technique has generated primarily qualitative data. We illustrate this review with recent studies conducted in the context of a range of English and international qualifications.

Conclusions

This review enhances understanding of some complex conceptual issues at the heart of multiple areas of test construction. We intend to offer new insights and practical ideas for users and developers of diverse assessments.

References

Cooze, M. and Shaw, S. D. (2007) FCE and CAE Review: establishing the impact of reduced input and output length in Part 1 FCE and CAE writing. *Research Notes*, Issue 30, November 2007, 15-19.

Falvey, P. and Shaw, S. D. (2006) IELTS Writing: revising assessment criteria and scales (Phase 5). *Research Notes* 23. Cambridge: Cambridge ESOL.

Greatorex, J. and Suto, W.M.I. (2008) What do GCSE examiners think of 'thinking aloud'? Findings from an exploratory study. *Educational Research* 40(4) 319-331.

Johnson, M. and Nàdas, R. (2009) An investigation into marker reliability and some qualitative aspects of on-screen essay marking. *Research Matters: A Cambridge Assessment Publication*, 8, 2-9.

Milanovic, M, Saville, N. and Shuhong, S. (1996) A Study of the Decision-making Behaviour of Composition markers, in Milanovic, M. and Saville, N. (Eds) *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium and Arnhem*, Cambridge: UCLES and Cambridge University Press, 92–114.

Shaw, S. D. (2002) The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes*, Issue 8, May 2002, 13-17.

Shaw, S. (2008). Essay Marking On-Screen: implications for assessment validity. *E–Learning*. 5(3), 256-274.

Shaw, S. D. and Weir, C. J. (2007) *Examining Second Language Writing: research and practice*. Studies in Language Testing Vol. 26. Cambridge: CUP.

Shaw, S. D. and Falvey, P. (2008) *The IELTS Writing Assessment Revision Project: towards a revised rating scale*. Cambridge ESOL Web-Based Research Report No. 1. Monograph. CUP/Cambridge ESOL.

Suto, W.M.I. and Greatorex, J. (2008a) A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy and Practice* 15 (1) 73-90.

Suto, W.M.I. and Greatorex, J. (2008b) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal* 34 (2) 213-233.

Suto, W.M.I. and Nàdas, R. (2009a) Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features.' *Research Papers in Education* 24 (3) 335-377.

Suto, I. and Nàdas, R. (2009b) *Investigating examiners' thinking: using Kelly's Repertory Grid technique to explore cognitive marking strategies*. Paper presented at the 14th International Conference on Thinking, Kuala Lumpur, Malaysia, June 2009.

Suto, I., Nàdas, R. and Bell, J. (2009) Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*. (Published on line to date.)

For further information please contact Dr Irenka Suto (suto.i@cambridgeassessment.org.uk) or Stuart Shaw (shaw.s@cie.org.uk) at Cambridge Assessment, 1 Hills Road, Cambridge, CB1 2EU)