

# Creating research programmes to support the development and validation of qualifications: what are the key assessment issues and what are the key research methods?

Irenka Suto and Stuart Shaw



CAMBRIDGE ASSESSMENT

When implementing any new qualification, it is important to create a body of research evidence to support and justify any claims of assessment validity including reliability. A theoretical underpinning coupled with empirical investigation promotes confidence among stakeholders. Key issues for researchers range from examiners' understanding of, and confidence in, the assessment process, to the aspects of mark schemes and students' work that most influence examiners' judgements and therefore assessment validity. Here, we compare and contrast some well-established methodological approaches to researching some of these issues.

## 'Think aloud' method

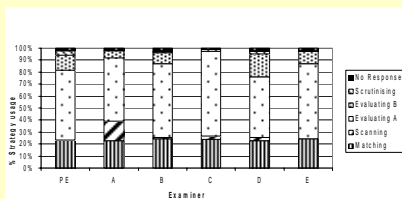
This primarily qualitative research method has a long history and can be used with research participants (e.g. examiners, teachers, students) in person or over the telephone (Greatorex and Suto, 2008). Researchers at Cambridge Assessment have investigated examiners' cognitive marking strategies by inviting 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:

<b>Matching</b>	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.
<b>Scanning</b>	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.
<b>Evaluating</b>	Examiner attends to aspect of the space dedicated to a question. She processes the information semantically, considering the student's response for structure, clarity, factual accuracy and logic or other characteristics given in the mark scheme.
<b>Scrutinising</b>	Examiner uses this only for unexpected or incorrect responses. She 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, she scans and evaluates multiple aspects of the response with the overarching aim of reconstructing the student's line of reasoning or working out what she was trying to do.
<b>No response</b>	Examiner uses this when nothing is written in the answer space allocated to the question. She looks at the space once or more to confirm this, she can then award zero marks.

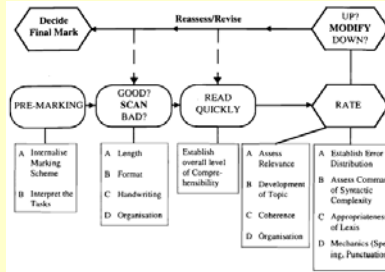
A quantitative analysis of marking strategy usage was conducted. This graph shows the proportions of marking strategy usage for each of six GCSE Business Studies examiners (Suto and Greatorex, 2008 b):



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

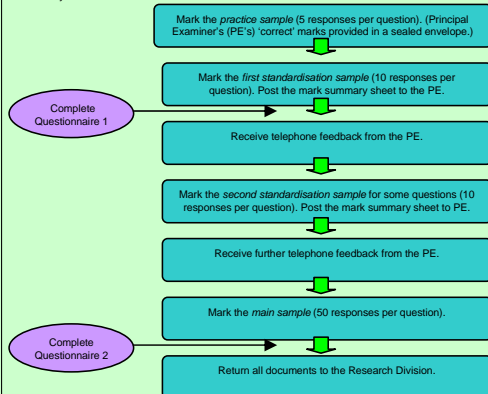
Verbal protocol analysis has also been used to identify reading strategies in the assessment of second language writing (Milanovic, Saville, and Shuhong, 1996). Group interviews, introspective verbal reports, and retrospective written reports with experienced examiners of Cambridge ESOL First Certificate in English (FCE) and Certificate of Proficiency in English (CPE) compositions were conducted. They revealed four discernible approaches to composition marking:

- Principled and Pragmatic Two-Scan Approaches
- Read Through Approach and
- Provisional Mark Approach

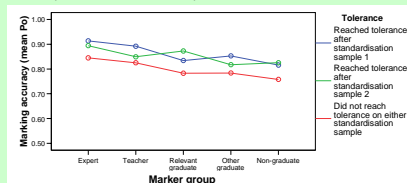


## Experimental marking methods

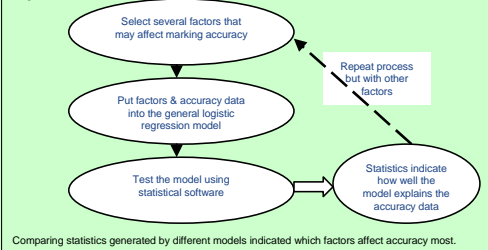
Inviting examiners to mark past students' work in experimental settings enables researchers to explore issues such as inter-examiner reliability and the relative merits of different training and standardisation procedures. Key advantages of experimental marking are that it cannot affect the results of current students, and that the same students' responses can easily be marked by multiple examiners. Suto and Nadas (2009) conducted an experimental marking study of a past GCSE 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:



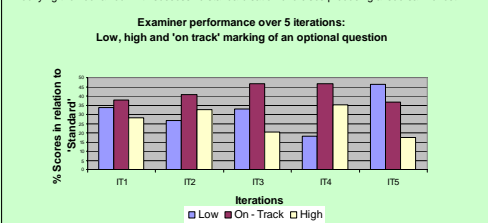
This graph reveals some of the influences of the background experience on marking accuracy on the main sample of students' responses. The mean proportion of exact agreement between each marker and the Principal Examiner (mean Po) has been used as a measure of accuracy. Accuracy is very high in general. 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:



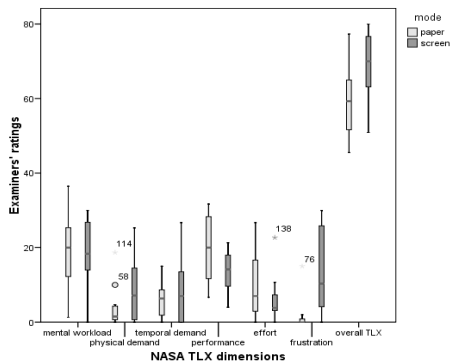
In an experimental marking study of the Certificate of Proficiency in English (CPE), Shaw (2002) observed that an experimental iterative standardisation process of training and successively delivered feedback to examiners did not improve inter-rater reliability (but this was perhaps affected by the fact 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.



## 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 Nadas (in submission) 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 Nadas (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?"

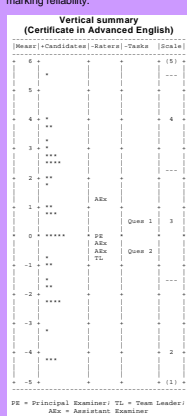
The examiner is usually able to identify a distinguishing feature (a KRG construct), such as amount of writing by the student or number of acceptable answers given in the mark scheme. The researcher and examiner use the feature to create a 5-point rating scale. The examiner then uses the scale to rate 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:

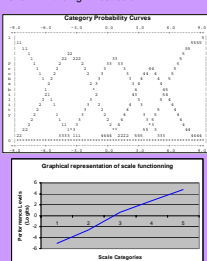
Relative marking accuracy of examination questions	Ratings given by an expert examiner (on a 5-point scale)					Total N examination questions
	1 Question makes a single demand on the student	2	3	4	5 Question makes multiple demands on the student	
Relatively low accuracy	0	1	2	2	3	8
Medium accuracy	1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

## Multi-faceted Rasch methods

As part of a project to modify the Certificate in Advanced English (CAE), Coore and Shaw (2007) used a multi-faceted Rasch approach to determine 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 asterisk). 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.



# Creating research programmes to support the development and validation of qualifications: what are the key assessment issues and what are the key research methods?

A poster presented at the 10<sup>th</sup> Annual Conference of the Association for Educational Assessment in Europe, 5-7 November 2009, Malta

## Abstract

When implementing any new qualification or programme of study, it is important to create a body of research evidence to support and justify any claims of assessment validity including reliability. This is crucial for innovative qualifications with novel structures and/or covering new topics. A theoretical underpinning coupled with empirical investigation promotes confidence among students, teachers and other stakeholders.

In this poster, we review a selection of the most common issues and concerns surrounding the assessment in new qualifications. We conceptualise the robustness of the assessment process as being affected by the personal expertise of assessors on the one hand, and the quality of the assessment instruments on the other hand. Key issues for researchers range from examiners' understanding of, and confidence in the assessment process, to the aspects of students' work and of the mark schemes that most influence examiners' judgements and therefore assessment validity, to setting standards and aligning syllabus content with the needs and preparedness of students.

We compare and contrast some well-established quantitative and qualitative methodological approaches to researching these issues. For example, multiple marking and script judgement studies generate quantitative data that can be analysed using a range of statistical techniques, including multi-faceted Rasch modelling. Verbal protocol data generated through 'thinking aloud' can be analysed quantitatively or qualitatively whereas Kelly's Repertory Grid technique generates purely qualitative data.

We illustrate this review with examples from studies conducted in the context of a range of qualifications used within the UK and internationally. We discuss how these studies have led to a greater understanding of some complex conceptual issues at the heart of test construction.

## 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.
- Greator, 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. & Nadas, 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.
- Suto, W.M.I. and Greator, 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 Greator, 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 Nadas, 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 Nadas, 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., Nadas, R. and Bell (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](mailto:suto.i@cambridgeassessment.org.uk)) or Stuart Shaw ([shaw.s@cie.org.uk](mailto:shaw.s@cie.org.uk)) at Cambridge Assessment, 1 Hills Road, Cambridge, CB1 2EU