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

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.

The ‘Think aloud’ method

The primarily qualitative research method has a long history and can be used with research participants (e.g. examiners, teachers, students) in person or via telephones or online (Johnson and Nádas 1999). Research 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 A-level Economics. Here are some typical instructions that have been read out to research participants:

In this study, we are interested in how you think when you are in the process of marking questions. I can be talking to you ‘think aloud’ whilst marking, and will record your thoughts using something called a think aloud protocol. I want to ask you to ‘think aloud’ as if you were marking a paper in front of you, silently whilst you are marking. It may help if you imagine that you are alone.

By ‘think aloud’, I mean that I want you to not only think about everything you would normally say to yourself, but actually say it all out loud. It is not important that you make sense or that it sounds good; I just want to hear you thinking. Please speak clearly so that the recording equipment picks up everything you say. If you are silent for any period of time, don’t feel that you have to think aloud any more. Please try to make as much noise as you can, but try to be as accurate as you think. If you are thinking about if you are doing

Questions of judgment only, whenever evaluating answers of exams or more interim judgments.

Questioning methods

Interviews (in person or 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 (in submission) used the NASA-TLX questionnaire to measure researchers’ marking load. The NASA-TLX questionnaire is a validated method of assessing the mental workload 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:

<table>
<thead>
<tr>
<th></th>
<th>***</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>****</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| +  -2 +           +       +        +     +
| +  -1 + **        +       +        +     +
| | *         | TL    |        |     |
| *   0 * *****     * PE    *        *     *
|       | ***       |       |        |     |
| +   1 + **        +       +        +     +
| +   0 +           +       +        +     +
| | ****      |       |        |     |
| +   4 + *         +       +        +  4  +
| |           |       |        |     |
| |           |       |        |     |
| |           |       |        |     |
|Measr|+Candidates|-Raters|-Tasks  |Scale|
|------------------------------------------|
|   0 |*******************************************************************|
|22             3333 111       4444  2222  555     333         4444 |
|y |      2       1        3       2      4      3   5       4         |
|l |         2  1             3                3       5  4            |
|i |          21                2            43         54             |
|b |         1  2                3          3   4     4   5            |
|P |      1        2       2        33  33                   5         |
|1 |                                                                  5|

Kelly’s Repertory Grid (KRG) method

KRG techniques have been used by Suto and Nádas (2009 a & b) to explore the features of questions and marks that most influence marking accuracy. Studies have focused on examiners for GCSE English Language, GCSE English Literature, GCSE Business Studies, A-level Economics and A-level Marketing. Johnson and Nádas (in submission) developed a list of descriptive factors to represent the marking process. Students were invited to think aloud whilst marking sample assessments and rate the importance of each factor. Suto and Nádas (2009) conducted an experimental marking study of a set of GCSE Business Studies question sets with a number of experienced examiners to assess the relative marking load of questions and the impact of reduced input and output factors on marking accuracy.

Experimental marking methods

Marking examiners to rank past examination: work in experimental settings enables researchers to explore issues such as inter-examiner reliability and the role of marks in different training and standardisation processes. Key advantages of experimental marking are that it (1) affords control over the methods of marking, (2) identifies and remedies the factors that influence marking, and (3) allows for the development of models and hypotheses about marking.

In an experimental marking study of the Certificate in Advanced English (CAE) marking process (Shaw 2002), examiners were asked to rate the importance of various factors to their marking process. They were also required to rank the impact of reduced input and output factors on marking accuracy.

Marking accuracy (mean Po)

Kelly’s Repertory Grid (KRG) method

Ratings of questions for marking accuracy (sample 1)

1 Question: makes a single point
3 Questions: makes a single point
5 Question: makes a single point
7 Question: makes a single point
9 Question: makes a single point

Multi-faceted Rasch methods

As part of a project to modify the Certification in Advanced English (CAE) marking process (Shaw and Suto 2002) used a multi-faceted Rasch approach to determine the impact of reduced input and output factors on marking accuracy. This approach is based on the general logistic regression model explained below.

Rasch (Generalised Discrete) Model

The relative influences of several factors affecting marking accuracy were assessed using logistic regression and comparing the relative loadings of the factors in the generalised logistic regression model fitted to the observed data. In the generalised logistic regression model, the probability of observing an item response is a function of the item characteristic curve and the person characteristic curve. The parameters of the item characteristic curve are estimated using maximum likelihood methods. The parameters of the person characteristic curve are estimated using maximum likelihood methods.

Marking accuracy (mean Po)

Kelly’s Repertory Grid (KRG) method

Multi-faceted Rasch methods

As part of a project to modify the Certification in Advanced English (CAE) marking process (Shaw and Suto 2002), used a multi-faceted Rasch approach to determine the impact of reduced input and output factors on marking accuracy. This approach is based on the general logistic regression model explained below.

Rasch (Generalised Discrete) Model

The relative influences of several factors affecting marking accuracy were assessed using logistic regression and comparing the relative loadings of the factors in the generalised logistic regression model fitted to the observed data. In the generalised logistic regression model, the probability of observing an item response is a function of the item characteristic curve and the person characteristic curve. The parameters of the item characteristic curve are estimated using maximum likelihood methods. The parameters of the person characteristic curve are estimated using maximum likelihood methods.

Marking accuracy (mean Po)

Kelly’s Repertory Grid (KRG) method

Multi-faceted Rasch methods

As part of a project to modify the Certification in Advanced English (CAE) marking process (Shaw and Suto 2002), used a multi-faceted Rasch approach to determine the impact of reduced input and output factors on marking accuracy. This approach is based on the general logistic regression model explained below.

Rasch (Generalised Discrete) Model

The relative influences of several factors affecting marking accuracy were assessed using logistic regression and comparing the relative loadings of the factors in the generalised logistic regression model fitted to the observed data. In the generalised logistic regression model, the probability of observing an item response is a function of the item characteristic curve and the person characteristic curve. The parameters of the item characteristic curve are estimated using maximum likelihood methods. The parameters of the person characteristic curve are estimated using maximum likelihood methods.

Marking accuracy (mean Po)
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 10th 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


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)