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

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

The 'think aloud' qualitative research method has a long history and can be used with research participants in a variety of ways. Research at Cambridge Assessment focused on a self-report method for the收集 cognitive marking strategies. The researchers believed that 'think aloud' could help them identify those features that are related to accuracy. If the dataset is too small to be analysed, researchers can use the feature rating data together with marking accuracy data to predict an examiner's rating. 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 'think aloud' study is thereby conducted.

Subsequently, the researcher uses the feature rating data together with marking accuracy data to predict an examiner's rating. 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 'think aloud' study is thereby conducted.

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 Nkitié (2009) used the NASA-TRL questionnaire with telephone interviews and focus groups to collect the initial data on marking strategies of examiners during paper-based and on-screen marking of GCSE English Literature essays. This graph shows the examiner's ratings of different aspects of their marking experience:

<table>
<thead>
<tr>
<th>Measurment</th>
<th>+ Candidates</th>
<th>- Raters</th>
<th>- Tasks</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE ABCDE</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Ex = Assistant Examiner</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2                3     2        4          5     4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>1                *              4         45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>1      2         2      3      3        44       5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1        2       2        33  33                   5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For a project to modify the Certificate in English (CIE) behaviours and on-screen marking of GCSE English essays (Suto and Gagne, 2008 b):

A quantitative analysis of marking strategy usage was conducted. This graph shows the proportion of marking strategy usage for each of six GCE Business Studies examinations (Suto and Gagne, 2008 b):
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

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)