



CAMBRIDGE ASSESSMENT

A critical review of some research methods used to explore rater cognition

Irenka Suto

E-mail: suto.i@cambridgeassessment.org.uk

A paper presented at the Annual Conference of the American Educational Research Association, New Orleans, LA, USA, 8-12 April, 2011.

Part of a symposium entitled: Rater Cognition and its Importance for Score Validity: Global Perspectives and Findings

Abstract

Objective

Internationally, many educational assessment systems rely on human raters to score examinations. Human scoring is the norm in most European countries, including England, and interest in its merits and disadvantages is growing in the United States. Using human raters tends to facilitate curriculum fidelity and the assessment of a range of sophisticated educational constructs, arguably strengthening assessment validity. However, it can introduce subjectivity into the scoring process, engendering threats to reliability. The objective of this presentation is to examine the key research techniques used in England to explore this potential trade-off. Self-report and experimental methods are reviewed critically with the aim of informing decisions on their applicability in American and other new contexts for rater cognition research.

Theoretical framework

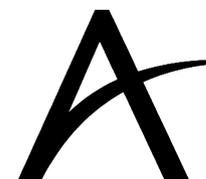
Dual processing theory of human cognition distinguishes two qualitatively distinct but concurrently active systems of operations: System 1 'intuitive' thought processes, and System 2 'reflective' thought processes. This theory is used: (i) to categorise examination questions according to the task demands placed on raters, and (ii) to develop the concept of personal expertise among raters. The relationship between task demands and personal expertise is presented in a model of optimal scoring reliability. For any given scoring task, reliability can be optimised by enhancing personal expertise or by reducing task demands. All other influential factors can be grouped according to which of these two improvement routes they contribute to.

Methods and techniques

Multiple components of the optimal scoring reliability model have been validated in empirical studies of rater scoring. Self-report methods including Kelly's Repertory Grid,



UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate



CAMBRIDGE ASSESSMENT

concurrent think aloud, and the NASA task load index have yielded important insights into: (i) the features of examination questions and student responses that most influence judgment complexity, (ii) the cognitive strategies raters adopt, and (iii) rater performance in screen-based and paper-based scoring environments. These methods' strengths and weaknesses are contrasted with those of multiple scoring studies, which have generated correlation statistics on scorers' background characteristics and scoring reliability.

Results

Both qualitative and quantitative data are presented to support decisions on the categorization of scoring task types and their assignment to raters of differing expertise. The value of such data is evaluated in terms of the methods used to obtain it. Critical concerns relate to the limits to the scale of studies imposed by self-report methods, the subjectivity in associated analyses, and the validity of generalising findings beyond immediate research contexts. These concerns contrast with those surrounding the limits to causal argument and in-depth explanation imposed by correlation statistics.

Significance

This investigation has significance for designers, developers and administrators of examinations. Arguably, the two pre-eminent goals of rater cognition research are prediction and control. Assessment professionals need both to predict levels of scorer reliability in given situations, and to anticipate the effects on reliability of manipulating controllable factors. The presentation concludes with a discussion of the place of rater cognition research in the wider context of assessment validity.