Language Testing and Evaluation

13

Annie Brown/Kathryn Hill (eds.)

Tasks and Criteria in Performance Assessment

PETER LANG
Internationaler Verlag der Wissenschaften

Tasks and Criteria in Performance Assessment

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn
and Günther Sigott

Volume 13

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Wien

Annie Brown/Kathryn Hill (eds.)

Tasks and Criteria in Performance Assessment

Proceedings of the
28th Language Testing Research Colloquium

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at <http://www.d-nb.de>.

E-ISBN 978-3-653-05394-4 (E-Book)
DOI 10.3726/978-3-653-05394-4

ISSN 1612-815X
ISBN 978-3-631-58416-3

© Peter Lang GmbH
Internationaler Verlag der Wissenschaften
Frankfurt am Main 2009
All rights reserved.

Printed in Germany 1 2 3 4 5 7

www.peterlang.de

Contents

Contributors

The Editors

Annie Brown is Associate Director for Assessment and Professional Development in the Ministry of Higher Education and Scientific Research in the United Arab Emirates. Prior to this she was Senior Research Fellow and Deputy Director of the Language Testing Research Centre at the University of Melbourne. Annie received the ILTA Best Paper Award in 2003 and the Jaqueline A. Ross Award for Outstanding PhD Disseration in 2004. She is a member of the Executive Board of ILTA and the Editorial Boards of Language Assessment Quarterly and Assessing Writing.

Kathryn Hill teaches clinical communication skills in the Faculty of Medicine, Dentistry and Health Sciences at the University of Melbourne. Prior to this she was a Research Fellow at the Language Testing Research Centre, University of Melbourne, and the Australian Council for Educational Research where she was involved in an extensive range of funded research projects in language and literacy testing and language program evaluation. She is currently completing her PhD in Applied Linguistics at the University of Melbourne.

The Authors

Jennifer Balogh holds a Ph.D. in psychology from UC San Diego. She designed and evaluated voice user interfaces at Nuance Communications and coauthored the book Voice User Interface Design. She was manager of Test Development at Ordinate, which specializes in automatic spoken language tests. She is currently an independent consultant.

Ana Maria Ducasse lectures in Spanish at La Trobe University in Australia. Prior to this she taught and managed ESL programs. Teaching languages kindled an interest in second language acquisition, particularly spoken language, and working as an IELTS assessor developed a parallel interest in language testing. These interests came together in her PhD research focus, which examines the assessment of interpersonal communication skills in a foreign language.

Thomas Eckes is Deputy Director and Senior Research Scientist in charge of the Language Testing Methodology, Research and Validation Unit at the TestDaF Institute, Germany. He has published numerous articles in edited volumes and peer-review journals. His research interests include rater effects, many-facet Rasch measurement, polytomous IRT models, construct validity of C-tests, standard setting, computerized item-banking and internet-delivered testing.

Usman Erdösy is a graduate of the Ontario Institute for Studies in Education. He has taught Academic English at the University of Toronto and at York University, and was Testing Manager for the Canadian Academic Language Assessment between 2004 and 2007. He is currently an independent researcher.

Tomoko Horai is currently completing her PhD at the Centre for Language Assessment Research, Roehampton University, London, where she has been undertaking research into the assessment of spoken language. Tomoko has presented her work at a number of international conferences and has published on assessment in Japan and internationally.

Youn-Hee Kim is a doctoral candidate in Second Language Education at the Ontario Institute for Studies in Education of the University of Toronto (OISE/UT). Her research interests include native and non-native teachers' assessment practices, fairness and validity in language assessment, and cognitive diagnostic assessment. She received the IELTS Master's Dissertation Award in 2006. She is currently co-editor of the MLC Bulletin, a seasonal newsletter for students and staff of the Modern Language Center, OISE/UT.

Anish Nair pursued his graduate study in Computer Science at USC. His areas of interest are empirical methods in natural language processing, web search and machine learning. At Ordinate Corporation he worked on algorithms for automatic grading of spoken utterances. He now works on modeling relevance of search advertisements at Yahoo! Inc.

Don Salting holds a Ph.D. in Linguistics from Indiana University. He was an assistant professor at North Dakota State University before moving to Silicon Valley. He was previously a test developer for Ordinate Corporation, but currently works for NetBase.com where he does NLP and computational lexicography.

Robyn Spence-Brown holds the position of Senior Lecturer in the School of Languages, Cultures and Linguistics at Monash University, Australia, where she teaches undergraduate Japanese language and studies, and post graduate units in Applied Linguistics. Robyn's research interests cover Language Assessment, Second Language Acquisition, and Language Education in school and tertiary settings.

Introduction

We are pleased to introduce this volume of selected papers from the 28th Language Testing Research Colloquium, held in July, 2006, at the University of Melbourne, Australia. The papers selected for this volume share a common theme – that of ‘performance’. Not only do they focus on performance assessments – of second or foreign language speaking and writing – but they also focus on the performance of one or both of the participants – the candidate (or candidates) and the raters – and the means whereby they construct that performance - the tasks and the assessment criteria.

These papers all recognise that test behaviour, be it the production of test discourse or the production of test scores, is a performance. The common theme that they share is to help us understand better the performance, and the factors surrounding it, so that we can ultimately ensure that our tests are more authentic, and more valid, representations of the construct we wish to measure. The papers draw on diverse contexts, including not only large-scale tests but also classroom assessments, and include languages such as German, Japanese, Spanish, in addition to English.

In a study of rater performance on a test of German as a foreign language, Eckes problematises the notion that raters, through training and regular practice, can be brought into line with one another. He found that given a set of analytic rating criteria, they differed in the importance they attached to different criteria, the ease with which they were able to apply them, and the confidence they had in their accuracy of application. Using a classification approach to rater variability, he found not only that raters were heterogeneous, but that they fell into distinct types. Eckes argues that analyses such as his can be used to improve rater training, for example by redirecting particular rater types’ attention to criteria not adequately represented within their characteristic scoring profile.

Kim’s study is also concerned with rater behaviour. She investigated the rating behaviour of native-speaker and non-native-speaker judges on a speaking test involving three distinct task types. While she found no difference between the two groups’ in terms of their internal consistency and severity, and no bias with regard to the three test tasks, she did find that they differed in the way they calibrated the scale levels, and in the features they attended to on the different tasks. Ultimately, however, Kim argues that non-native speakers should be considered as acceptable as native-speakers as judges of second or foreign language performance.

Continuing the theme of ‘difference’, Spence-Brown investigated students’ approaches to an authentic assessment task which formed part of the course assessment in a university Japanese language program. Drawing on interviews with the students, in addition to an analysis of the task discourse, Spence-Brown argues that the students approached the task in different ways as a result of different orientations to the course and the assessment, and that, despite the best of intentions, students orientations can, ultimately, subvert the teacher's intentions in setting the task, change intended positive washback into negative washback, and invalidate the outcomes of the assessment.

Horai investigated the impact of changes to speaking task performance conditions on scores and on measures of accuracy, complexity and fluency. She found that of four conditions (original, no planning, reduced support, and reduced response time), the removal of planning time was the most critical, although this effect was less noticeable for low proficiency learners than for high and borderline (intermediate) students. Horai draws implications from the findings for test developers, researchers and teachers.

The study by Salting et al also examined the impact of task variables, including lexical, morphological, syntactic features, on item difficulty, this time in an automated assessment of spoken English. The task required candidates to listen to and reorder three phrases to form a grammatical sentence. The authors found that item length and lexical frequency were the strongest predictors of item difficulty. The finding that syntactic and morphological factors did not correlate significantly with item difficulty supports, the authors argue, the implicit nature of syntactic priming.

Erdősy takes up the notion of indigenous assessment in the context of an undergraduate history course. He shows, through a mixture of quantitative and qualitative analyses, that the criteria applied to course assessments while not made explicit are not capricious, but derive from the discourse, spoken and written, of the course itself. Such indigenous criteria, while clearly valid in the specific context in that students are led to know them during the course, are nevertheless, he argues, of little value beyond that due to their context specificity. The way forward for more general assessments (such as that of academic language proficiency), he suggests, is to generate abstract criteria from such indigenous criteria, which can then be applied in a range of contexts.

Ducasse's study also involved the generation of indigenous criteria, for an assessment of interactional ability in a paired candidate task in a beginners' Spanish course. Her aim, however, is not only to generate a scale for use in the specific course context, but to further our understanding of the construct of 'interaction' in second language tests by providing empirical evidence of how one set of language teachers operationalise it. This study into rater orientations in a paired candidate task complements the growing body of research investigating candidate discourse in paired and group orals.

The Editors
Abu Dhabi and Melbourne, July 2008

Raters as scale makers for an L2 Spanish speaking test: Using paired test discourse to develop a rating scale for communicative interaction

Ana Maria Ducasse

This paper reports on the development of an evidence based rating scale to rate peer-peer L2 communicative interaction. The scale was based on experienced judges' comments on video-taped student samples, filmed during operational paired candidate tests of beginner level Spanish. In the study, six experienced teacher / raters generated criteria for the assessment of communicative interaction using a modified version of the Empirically-based, Binary-choice, Boundary-definition (EBB) method (Turner and Upshur, 1996), originally used to develop assessment rubrics for writing samples. Three main features of paired candidate interaction emerged as critical in defining the boundaries between levels of interactional skill: non-verbal interpersonal communication, interactive listening and interactional management. These features were subsequently incorporated into sample-based rating scales. The study advances our understanding of the significant features of spoken interaction, and also demonstrates how empirically-grounded scales can be developed for interaction.

Introduction

Since the 1980s, paired and group orals tests have been increasingly common as a way of reflecting in testing the emphasis on communicative language teaching in the classroom. Research into these new paired speaking tests originally concentrated on the effect on test scores caused by pairing candidates with different characteristics. Subsequently, the discourse produced in these paired tests was explored, and these studies have been followed more recently by rater verbal protocol studies to shed light on the process of rating pairs.

This paper focuses on rating criteria for paired speaking tasks, and more particularly how they are arrived at by scale makers. There has been a wide range of research into scale development in other contexts from various perspectives. Of particular relevance to this study is the use of student samples to derive empirically-based rating scales. Until now student samples have been used to develop criteria for writing, for monologic speaking tasks and for fluency scales, but not for paired tests involving peer-peer interaction samples.

This study was motivated by a practical need to comprehensively rate peer-peer interaction, in recognition of the fact that interaction among participants in a task plays a central role in generating discourse (Swain, 2001). If interaction is central, more research needs to be carried out into effectively incorporating it into rating scales. To identify the skills involved, the study looks at the point of intersection between the manner in which paired candidates manifest attributes of interaction and the way in which raters attend to those attributes.

The approach used is one that empirically derives scales by using teams of scale makers to define levels of performance by noting the salient differences between samples of paired L2 students performing a paired task in an oral test. Rating scales developed with teams of scale developers from student samples are not new. In a recent study Turner and Upshur (2002) used teams of raters to derive rating criteria from the same set of student samples.

Background to the study

Two strands of research provide the background to this study. One strand is on the development of rating scales, in particular data-based scales. The other strand concerns rating spoken interaction, in particular between peers.

Developing empirical rating scales

Rating scales usually mark out a series of levels, each of which is accompanied by descriptors that include characteristics of the performance expected at that level. The sample of candidate discourse used to assign a score is understood to derive from underlying language abilities or the construct being tested.

As reported in Turner and Upshur (2002), rating scales have been criticised for producing scores with low validity and reliability. Problems they cite involve:

- the ordering of scale criteria may be inconsistent with the findings of second language acquisition (SLA)
- criteria may be irrelevant to tasks and content
- criteria may be incorrectly grouped at different levels
- scales may lead to raters making false judgments because of relative wording

Improving the rating criteria could improve the problems with reliability listed above (Hamp-Lyons, 1991; North, 1995, 2003; North and Schneider, 1998). Scale development methods are basically divided in two types: intuitive and evidence based methods. Although the intuitive method is by far the most common way of arriving at rating scales using prior knowledge and consensus among experts, the evidence-based empirical method, which works *from* language output samples *towards* the descriptors, is the method chosen for this study. A rating scale based on what raters observe and notice during peer-peer interaction might address problems with reliability. It answers calls from the literature, such as that of Chalhoub-Deville (1997), who cautions that theory alone is insufficient to produce task specific scales, and Fulcher (2003), who directly calls for empirically developing rating scales.

The development of evidence-based scales for rating paired orals is further motivated by the fact that this format has been included comparatively recently into test batteries. There has been less time to research the peer-peer construct. It is difficult to gauge theoretically what features might be salient to raters in peer-peer interaction. It has been said that assessment that takes into account salient features of a task can improve measurement (Pollit and Hutchinson, 1987) but taking salient features into account can be difficult if such features have not been shown empirically to be salient from a rater perspective.

Rating paired orals

Different aspects of peer-peer interaction, in a group or in a pair, are interesting to testers. Features researched so far that have been empirically observed in paired discourse involve the number of functions produced (Lazaraton, 2002; Taylor, 2001) and conversation management skills (Dimitrova-Galaczi, 2004). These aspects have been qualitatively described and validated but have not been used as evidence to build data-based scales. There still remain, however, other unobserved, and until recently undescribed, features of interaction that make scoring interaction in groups or pairs difficult. Politt and Murray (1996) ask:

- Should comprehension be assessed as part of oral proficiency?
- Should a proficiency battery test language production or language interaction or both?
- Should the oral test be one of communicative success or linguistic ability?

Comprehension, language production versus interaction, and communicative versus linguistic success are issues unexplored for the pair format from a *rater* perspective.

Of the studies carried out so far, a number have investigated the difficulty for scales and scale makers to adapt to the paired and group context: Nunn (2000) tackles the problem of designing rating scales for small group interaction during classroom activities as distinct from paired oral tests. The study acknowledges that for groups and rating scales "the considerable difficulties of reliability and validation need to be fully understood and the facile extrapolations about how students can perform in real life should be avoided" (Nunn 2000: 178). Nevertheless, despite the recognition of a difficult problem it is suggested that teachers recognise that "the question is not whether to do it but how to do it as fairly and efficiently as possible" (Nunn 2000: 178). The solution offered is to use the same scales for teaching, learning and assessment. How one develops these scales empirically still remains unresolved, regardless of the scope of the intended application.

In a more recent validity study on a university group oral test (Van Moere, 2006) the greatest variability was found in the person by occasion interaction: the people in a group are most likely to affect each others' performance, which is expressed as "the more intangible interpersonal factors in the way group members react to each other" (Van Moere, 2006:436). The 'intangible' remains so far unexplained in the peer-peer testing context and these interpersonal factors need to be described and captured in a scale to reduce variability.

In contrast, Bonk and Ockey (2003) in a many facet Rasch analysis of a second language group oral discussion task found that "rater and scale reliability were achievable under real testing conditions even when the discourse was largely uncontrolled". We argue that rater training and scale relevance is the key to turning Van Moere's (2006) 'intangible interpersonal factors' which characterize paired oral communication into a "reliability...achievable under real testing conditions" (Bonk and Ockey, 2003). This can be achieved in two ways: by focusing empirical scale development on candidate output and by including features in scales that scale makers attend to while rating to facilitate rater training. These issues have been addressed in only a handful of studies so far, and these have focused on interviews not on peer-peer interaction.

Orr (2002) analyses verbal reports given by raters on the decision making process during the rating of the UCLES First Certificate of English (FCE). Thirty two raters completed verbal reports (Green, 1998) on two separate pairs of candidates performing the paired task from the FCE under test conditions. In that study Orr reports most compromising results. Raters were firstly found to apply different standards because they vary in severity, and secondly they were found to focus on rating criteria in different ways. (This has also reported been reported in Brown (2000) and Meiron (1998)) Lastly, raters were found to vary in the amount of non-criterion information they noticed for each candidate. Included in the non-criterion information heeded while rating the paired interaction was the amount of non-verbal communication, for example eye contact and body language. The results have serious implications for the validity of the paired oral: the raters had varying perceptions of the performance but how the raters vary was not obvious in the scores. This makes it difficult to understand what FCE speaking test scores represent.