The equivalence of direct and semi-direct speaking tests
STUDIES IN LANGUAGE TESTING . . . 13
Series editor: Michael Milanovic and Cyril J. Weir

Also in this series:

An investigation into the comparability of two tests of English as a Foreign Language: The Cambridge–TOEFL comparability study
Lyle F. Bachman, F. Davidson, K. Ryan, I-C Choi

Test taker characteristics and performance: A structural modelling approach
Antony John Kunnan

Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem
Michael Milanovic, Nick Saville

The development of IELTS: A study of the effect of background knowledge on reading comprehension
Caroline Margaret Clapham

Verbal protocol analysis in language testing research: A handbook
Alison Green

A multilingual glossary of language testing terms
prepared by ALTE members

Dictionary of language testing
Alan Davies, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, Tim McNamara

Learner strategy use and performance on language tests
James Enos Purpura

Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida
Antony John Kunnan

Issues in Computer-adaptive testing of reading proficiency
Micheline Chalhoub-Deville

Experimenting with uncertainty: Essays in honour of Alan Davies
A. Brown, C. Elder, N. Iwashita, E. Grove, K. Hill, T. Lumley, K.O‘Loughlin, T. McNamara

An Empirical investigation of the componentiality of L2 Reading in English for academic purposes
Cyril J. Weir, Yang Huizhong, Jin Yan
The equivalence of direct and semi-direct speaking tests

Kieran J. O’Loughlin
Contents

Series Editor’s note vii
Preface ix
Acknowledgements xi

Chapter One
Introduction 1
  Rationale 1
  Methodological approach 2
  Structure of the book 3
  Direct, semi-direct and indirect speaking tests 4
  Establishing the equivalence of direct and semi-direct tests 6
  Research findings 12
  Conclusion 31

Chapter Two
The access: test 33
  Background to the test 33
  Development of the test 34
  The oral interaction sub-test 35
  Conclusion 45

Chapter Three
Methodology 47
  The instrumental case study approach 47
  Research design of the study 55
  Summary 61

Chapter Four
Trial 1: The analysis of test scores 62
  Overview 62
  The December 1992 trial 62
  A preliminary study 65
  The current study 69
  Conclusion 79
Chapter Five
Trial 1: The analysis of candidate language output
  Overview
  Rationale
  Methodology
  Discourse Study A: The qualitative analysis
  Discourse Study B: Lexical density
  Conclusion

Chapter Six
Trial 2: The study of test processes
  Introduction
  The June 1994 trial
  Method
  Findings
  Conclusion

Chapter Seven
Trial 2: The analysis of test scores
  Introduction
  Analysis with anchored item difficulty estimates
  Analysis with unanchored item difficulty estimates
  Bias analysis
  Conclusion

Chapter Eight
Summary and conclusions
  Overview
  Summary of findings
  Implications of findings

References
Appendices
Subject index
Author index
Series Editor’s note

This volume addresses the issue of spoken language assessment looking in particular at the equivalence of direct and semi-direct oral interviews. Kieran O’Loughlin’s work is based on the development and validation of the spoken language component of the access: test designed in the early 1990s for migrants to Australia. It is an important language testing project in the Australian context and was funded by the Commonwealth Department of Immigration and Ethnic Affairs. While the project as a whole brought together experts from a number of Australian universities, the oral test was developed by a team at the University of Melbourne. This volume is of particular significance and interest to the language testing community because it takes a multi-faceted view of the investigation of test comparability. While much research of this sort has tended to look only at quantitative data, largely correlational analyses, O’Loughlin taps into a range of different types of evidence and attempts to explore the process of construct validation in oral assessment to a depth that is rarely found.

The assessment of spoken language ability is a topic of enduring importance in the work of the University of Cambridge Local Examinations Syndicate (UCLES) given that UCLES assesses the spoken language ability of about 800,000 candidates around the world every year. The issue of semi-direct versus direct assessment of speaking has continued to be a topic of interest at UCLES and we have found that O’Loughlin’s work makes a valuable contribution to our understanding. His work closely reflects our own interests particularly in the area of the qualitative analysis of oral interview interaction.

The importance of oral assessment and the need to better understand the complex issues and interactions that underlie performance in this particular context have long been a topic of debate at UCLES. As early as 1945, Jack Roach, an Assistant Secretary at UCLES at the time, was writing on the topic in his internal report entitled ‘Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students.’ Indeed, in his book Measured Words (1995), Bernard Spolsky considers Roach’s work to be ‘probably still one of the best treatments in print of the way that non-psychometric examiners attempted to ensure fairness in subjective traditional examinations’. Roach’s work is addressed in more detail by Cyril Weir in a volume currently being prepared for this series that focuses on the revision of the Certificate of Proficiency in English (CPE).
Over the last ten years or so, a considerable amount of work has taken place at UCLES in order to gain a better understanding of oral interview interactions, processes and rating scales both in quantitative and qualitative studies. Working internally or with colleagues at universities in the UK, USA and Australia, numerous studies have been carried out. Amongst other things, projects have looked at:

- discourse variation in oral interviews;
- rating scale validation;
- interlocutor frames and how their use by examiners can be described and monitored;
- interlocutor language behaviour;
- a methodology to allow test designers to evaluate oral assessment procedures and tasks in real time;
- comparisons of one-to-one and paired oral assessment formats;
- test takers’ language output;
- the development and validation of assessment criteria.

In 1998 UCLES EFL established, within its Validation Department, a dedicated unit to drive research in the area of performance testing, which essentially covers the assessment of speaking and writing.

It should also be noted that the next volume in this series, *A qualitative approach to the validation of oral language tests*, by Anne Lazaraton also makes a valuable contribution to the assessment of spoken language ability. Both O’Loughlin’s and Lazaraton’s volumes underline UCLES commitment to furthering understanding of the dimensions of spoken language assessment.
Preface

This study investigates the equivalence of direct (live) and semi-direct (tape-mediated) speaking tests. The issue is explored through a comparison of live and tape-based versions of the speaking component of the access: test, a four-skill English language test designed for prospective non-English speaking background (NESB) skilled migrants to Australia. The access: test was developed between 1992 and 1994 by a consortium of Australian universities and Adult Migrant Education Program (AMES) providers under the aegis of the National Centre for English Language Teaching and Research (NCELTR), Macquarie University. The project was funded by the Commonwealth Department of Immigration and Ethnic Affairs (DIEA). The access: test was used in overseas test centres from January 1993 until May 1998 when it was replaced by the International English Language Testing System (IELTS). The oral interaction sub-test was designed by the Language Testing Research Centre (LTRC) at the University of Melbourne, Australia in 1992. The two versions of the speaking sub-test were used interchangeably in overseas test centres and candidates normally had no choice about the version which they would undertake. It was important, therefore, that candidates’ final results should not be adversely affected by the particular method used to assess their oral proficiency.

Much previous comparability research in language testing has relied primarily on concurrent validation which focuses on the equivalence between test scores. However, in this book, it is argued that examining the relationship between test scores provides necessary but insufficient evidence as to whether the same language constructs are being tapped in different tests. This provided the rationale for a study which employed a wide range of very different types of evidence (including test taker language output, feedback from the various ‘players’ in the test process such as test developers, interviewers, test takers and raters as well as test scores) and analyses (both qualitative and quantitative) in order to investigate the equivalence of the direct and semi-direct versions of the access: oral interaction sub-test. In so doing, it demonstrates the need to examine language tests from multiple perspectives in order to obtain deeper insights into what they appear to measure and thus provide richer, more comprehensive evidence for construct validity of which concurrent validity is only one (albeit important)
component. In turn, it is argued, this approach provides a more solid and therefore more valid basis from which to draw conclusions about test equivalence.

The various types of data used in the study were gathered from two separate trials conducted in Melbourne, Australia (December 1992 and June 1994) where test takers undertook both the live and tape-based versions of the oral component of the access: test.

The statistical analysis of test scores from the two versions in the December 1992 trial were contradictory insofar as conflicting results were obtained from the different measures used to measure their equivalence. However, the most rigorous measure indicated a lack of equivalence.

A comparative study of test taker language output under the two test conditions was then undertaken using audio recordings obtained from the December 1992 trial. While broad qualitative analyses of a range of discourse features suggested that there were important similarities between nearly all of the tasks on the two versions, a more detailed quantitative analysis of lexical density suggested that all tasks on the live version were characterised by a significantly higher level of interactivity than on the tape version. This result suggested the possibility that different oral skills were being tapped in the live and tape versions despite the attempt to equate them at the design stage, i.e. interactive versus monologic speaking ability.

The examination of test processes in the June 1994 trial provided further evidence in support of this conclusion but suggested that the apparent lack of equivalence between test scores obtained in the two versions could also partly be explained by the impact of a number of contaminating factors other than oral proficiency on the measurement process. These factors included the quality of the interaction between candidate and interlocutor on the live version, the adequacy of preparation and response times on the tape version, candidates’ level of comfort with the two versions as test environments and the existence of rater bias in relation to individual candidates on both versions.

Finally, the analysis of test scores and especially the band levels obtained by candidates in the June 1994 trial confirmed the apparent lack of equivalence between test scores reported in the December 1992 trial.

Overall, these findings suggested that the live and tape-based versions of the oral interaction sub-test could not be safely substituted for each other primarily because they were drawing on different components of the oral proficiency construct but also because the measurement process appeared to have been insufficiently constrained so as to yield a satisfactory level of reliability across the two formats.
Acknowledgements

I am grateful to the following organisations for permission to reproduce previously published material in this book: firstly, Arnold Publishers for an article entitled ‘Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test’ appearing in Language Testing (1995) 12, 2: pp. 217–37 (see Chapter 5) and secondly, NCELTR publications, Macquarie University, for a chapter entitled ‘Test taker performance on direct and semi-direct versions of the oral interaction module’ published in access: issues in language test design and delivery (1997): pp. 117–46 (see Chapters 6 and 7).

I dedicate this book to Don Hay, and to all of the other unsung heroes of the AIDS pandemic – the living and the dead.

J'ai toujours pensé que j'étais étranger à cette ville et que je n'avais rien à faire avec vous. Mais maintenant j'ai vu ce que j'ai vu, je sais que je suis d'ici, que je le veuille ou non. Cette histoire nous concerne tous.

La Peste, Albert Camus