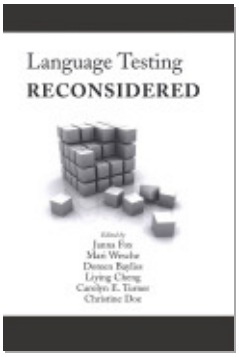
Language Testing Reconsidered

Fox, Janna, Wesche, Mari, Bayliss, Doreen, Cheng, Liying



Published by University of Ottawa Press

Fox, Janna, et al.
Language Testing Reconsidered.
University of Ottawa Press, 2007.
Project MUSE.muse.jhu.edu/book/4459.

6 Qualitative Research Methods in Language Test Development and Validation

Anne Lazaraton
University of Minnesota

and Lynda Taylor
University of Cambridge ESOL Examinations

Abstract

One of the most important methodological developments over the last fifteen years has been the introduction of qualitative research methodologies to support the design, description, and validation of language tests. Many language testers have come to recognize the limitations of traditional statistical methods for language assessment research, and have come to value these innovative methodologies as a means by which both the assessment process and product may be understood.

This chapter introduces readers to several qualitative research methods as they relate to language testing and assessment, namely: Discourse and Conversation Analysis, Observation Checklists and Verbal Protocol Analysis. It focuses specifically on some of the qualitative studies in speaking and writing assessment undertaken by Cambridge ESOL in order to illustrate how outcomes from such investigations can feed directly into operational test development and validation activity. These research methods — together with other qualitative techniques — offer language testers viable solutions for a range of validation tasks.

Introduction

In a state-of-the-art paper published in *Language Testing*, Bachman (2000) argues that the field of language testing has shown ample evidence of maturity over the last 25 years — in practical advances such as computer-based assessment, in our understanding of the many factors involved in performance testing, and in a continuing concern over ethical issues in language assessment. However, in our opinion, an equally important methodological development over the last fifteen years has been the introduction of qualitative research methodologies to design, describe, and validate language tests. That is, many language testers have come to recognize the limitations of traditional statistical methods for language assessment research, and have come to value these innovative methodologies as a means by which both the assessment process and product may be understood.

Our purpose in this chapter is to discuss several qualitative research methods as they relate to language testing and assessment, namely: Discourse and Conversation Analysis, Observation Checklists and Verbal Protocol Analysis. There are various other qualitative techniques which we cannot cover here due to space limitations; for a broader overview the interested reader is encouraged to consult Banerjee and Luoma (1997), Lumley and Brown (2005), McNamara, Hill, and May (2002) and Richards (2003).

We then go on to examine the application of qualitative methodologies in relation to the testing of speaking and writing where such methods have proved particularly fruitful over the past 10 to 15 years. Specifically, we discuss qualitative analysis as a means to gain insights into interlocutor behaviour (the *process* of *speaking* assessment), into test-taker behaviour (the *product* of *speaking* assessment) and into rater behaviour (the *process* of *writing* assessment).

Process and Outcome in Oral Assessment

An examination of the body of research on language testing suggests that it can be grouped in two main periods: pre-1990 and post-1990. The first period was defined by research that was almost entirely quantitative and outcome-based. Construct validation studies, comparisons of face-to-face vs. tape-mediated assessments and analyses of rater behaviour were undertaken largely on the FSI/OPI Oral Proficiency Interview (see Lazaraton, 2002, for a review of this literature). Leo van Lier's seminal 1989 paper on the assumed but untested relationship between oral interviews and natural conversation was to change that, by stimulating an interest in undertaking empirical research into the nature of discourse and interaction that arises in face-to-face oral assessment. Specifically, van Lier called for studies that would even go beyond detailing the oral assessment process, to inform us about the turn-by-turn sequential interaction in oral interviews and how oral test discourse is structured by the participants. Work along these lines would allow us to determine whether or not conversational processes are at work in the oral interview, and thus whether (and how) test discourse resembles non-test discourse.

Since 1989 there has been a proliferation of studies[1] analyzing aspects of the discourse and interaction in oral interview contexts. Briefly, this work includes:

- analyses of candidate behaviour and
  - proficiency level
  - interlocutor familiarity
  - the role of the L1
- analyses of interviewer behaviour and
  - the role of accommodation

  - cross-cultural influences on behaviour
  - features of *good interviewers*
  - its effect on candidate ratings
  - the effect of gender on discourse and scores
- analyses of interviewer vs. candidate behaviour and
  - their asymmetrical nature
  - how these behaviours differ in conversation
- analyses of test format
- rating scale production
- task analysis
- candidate behaviour in the pair/group oral
- analyses of rater cognition

Many studies in this area have been multidisciplinary in nature and almost all employ transcribed discourse as the basis for analysis; a few of them employ Conversation Analysis (CA), an inductive method for finding recurring patterns of talk-in-interaction and the approach with which much of our own work was undertaken. Though even a cursory overview of CA is beyond the scope of this chapter (but see Atkinson and Heritage, 1984; Pomerantz and Fehr, 1997; Schegloff, Koshik, Jacoby, and Olsher, 2002), it should be noted that CA, in its pure form, is guided by the following analytic principles:

- using authentic, recorded data which are carefully transcribed
- *unmotivated looking* at data rather than pre-stating research questions
- using the *turn* as the unit of analysis
- analyzing single cases, deviant cases and collections thereof
- disregarding ethnographic and demographic particulars of the context and participants
- eschewing the coding and quantification of data

These principles should be kept in mind in understanding the interactive discourse data presented below.

Interlocutor Speech Behaviours on Tests

Although our work on oral interviews was motivated by an inherent interest in the construction and co-construction of oral discourse, the University of Cambridge ESOL Examinations (Cambridge ESOL), whose data are reported on below, were mainly interested in using CA as a means of construct validation;

that is, evaluating the meaningfulness and the appropriateness of interpretations based on oral test scores. Validation questions of particular interest to language testing researchers include: How might the behaviour of the interlocutor impact the oral interview process? Are interlocutors consistent across interviews with different candidates? What sort of consistency is there between interviewers for any given speaking test?

To answer these questions, Cambridge ESOL made available a number of audio cassette tapes and accompanying test materials for two of their general English proficiency examinations: the Key English Test (KET) close to beginner level[†] and the Certificate in Advanced English (CAE) at upper intermediate level.[2] The audiotapes were carefully transcribed using Conversation Analysis conventions (Atkinson and Heritage, 1984; see Appendix 1) and then analyzed on a turn-by-turn basis for patterns of interest (for an in-depth description of collecting, transcribing, and analyzing such data, see Lazaraton, 2002). These analyses showed that the interlocutors *routinely* modified their speech behaviour (and thus test delivery) in what came to be seen as predictable ways. A sampling of the behaviours found are illustrated in the fragments below.

(1) *KET Tape 20, Part 1, Interlocutor E*:

IN: you are from which country.

(2) *KET Tape 21, Part 1, Interlocutor D*:

IN: what do you- um how is London different from the town in which you come from.

One of the most immediately apparent speech modifications made by the interlocutors in both the KET and CAE tests is the rephrasing of interview questions. In (1), this rephrasing (from "Where are you from?") results in an unembedded wh-question; in (2), the rephrasing results in an awkward, if not ungrammatical question. Interlocutors also rephrased wh-questions into "easier" yes-no questions, as in (3) and (4).

(3) *KET Tape 19, Part 1, Interlocutor Y*:

IN: tell me something about your family.=have you got any brothers or sisters?

(4) *CAE Tape 35, Part 1, Interlocutor R*:

CA: and basically we we we deals with um worlduh bank. and of course English is uh (only) (.) way to deal with this.

IN: tsk! %okay%

[†][Ed. note: These tests are linked to the levels of the CEFR. See Alderson, Chapter 2, and McNamara, Chapter 7, for other examples of the relationship between proficiency levels and the CEFR.]

CA: so it is really [( )
5 → IN: [right. so how do you actually use
→ your English.=do you use it on the telephone? or
→ IN: [on faxes or telexes or]
CA: [no. (.) no no | basically reading because.

IN's question in (4) is notable for its complexity (wh- to yes-no to several or-choices added on) and for its non-necessity (that is, the candidate begins his answer at the first possible space after the first rephrasing).

When candidates seemed to be struggling to come up with a word or complete a thought, interlocutors would often supply them, as in (5) and (6).

(5) *KET Tape 24, Part 1, Interlocutor N*:

IN: and do you work in Trieste? or are you a student.
CA: um I am a student at the university.
IN: uh huh and what do you study.
CA: I study pedagogue pedagogue
→ IN: pedagogies
CA: yeah yeah %pedagogies%

(6) *CAE Tape 36, Part 4, Interlocutor H*:

IN: and um Sweden is having some serious environmental problems.
CA: yeah (.) yeah th- we really have that . . . so the fish and the plants? in the sea they can't .hhh! uh(hhh) [they can't breathe [now so this | they died
→ IN: [breathe [breathe yeah yeah]

While the candidate in (5) has "invited" the suggestion by the interlocutor (through an attempts to produce the word), it is not immediately apparent that the candidate in (6) needed or even wanted IN's contribution of "breathe."

Interlocutors also reacted to candidate answers in various ways, as shown in (7) and (8).

(7) *KET Tape 18, Part 1, Interlocutor T*:

IN: tell me Edgard. how long have you been here. in London.=
CA: =yes three months.
→ IN: three months?=
CA: =yes

(8) *KET Tape 18, Part 1, Interlocutor T*:

IN: .hhh and so how often do you go to the cinema.
CA: .hhh the weekend.
→ IN: at the weekends.=
CA: =%mmm%

In Fragment (7), IN repeats CA's response from line 3, perhaps as a confirmation check of understanding. Repetitions also function as embedded corrections, as in IN's turn in Fragment (8). Here, the repetition contains the missing preposition "at."

Other reactions included evaluating and commenting on responses (Fragments 9–11) and drawing conclusions for candidates (Fragments 12–13).

(9) *KET Tape 5, Part 1, Interlocutor K*:

IN: okay uhm Arzu. where do you come from.
CA: .hhh I come from (.) Turkey.
IN: ah. and what town are you from in Turkey.
CA: uh from Istanbul
→ IN: ah. %right% okay. .hhh and do you work...

(10) *CAE Tape 33, Part 1, Interlocutor M*:

CA: and she's twenty years old? (from)? lives: uh:: on her own? in a small flat in the north part of: uh: Italy?
→ IN: wow.
CA: and it's uh...

(11) *CAE (KET) Tape 80, Part 1, Interlocutor F*:

CA1: after in the winter (really) and in the summer (.8) used to: (.8) be a lifeguard (.5) on the beach. (.5)
CA2: mmhmm
→ IN: a lifeguard!
??: %life[guard!%
CA1: [yeah: (.5) sounds good!

(12) *KET Tape 24, Part 1, Interlocutor N*:

IN: and- and how often do you play tennis.
CA: at home? or in Cambridge.
IN: uhm in Cambridge heh [heh

CA: [Cambridge no

→ IN: you don't play.

CA: I don't play tennis.

(13) *CAE Tape 5, Part 1, Interlocutor F*:

IN1: ... what is she like in the class

IN2: hmm hmh hmh hmh

CA: well really she she's very quiet

IN1: quiet.

CA: yeah.

→ IN: and you're noisier.

As a result of these analyses, Cambridge ESOL instituted various means for insuring greater standardization across interlocutors on their speaking tests. Measures included the introduction of an *Interlocutor Frame*, which guides examiner talk and provides test candidates with consistent interlocutor input and support, as well as the implementation of an Oral Examiner (OE) Monitoring Checklist, which is used for monitoring/evaluating examiner performance over time and highlighting training priorities (Taylor, 2005).

But interlocutor talk is only one aspect of the speaking test process that can be examined using discourse analysis (DA). The other is an analysis of candidate output, which is discussed in the next section.

Analyses of Candidate Output on Speaking Tests

Another set of Cambridge-commissioned work dealt with language output on the First Certificate in English (FCE) Speaking Test, which took place as follow-up to the FCE Revision Project, 1991–1996. However, because the focus of this research was solely on the features of candidate language, conversation analysis, which analyzes dyadic interaction, could not be used. For this research, a broader approach to discourse analysis was considered a viable option for understanding candidate speech production within the context of an oral examination.

The research question that guided these studies was: What is the relationship between the task features of the four parts on the FCE speaking test and the candidate output in terms of speech production? A corpus of live FCE data from 1996 test administrations was studied to determine if the features of speech that were predicted as output and which were to be evaluated by the rating criteria were actually produced by the candidates. The hypothesized speech functions, described in the then-current FCE materials, were used as a starting point and were modified or supplemented as the data analysis progressed.

As reported in Lazaraton and Frantz (1997), candidate output in Part 2, where candidates are required to produce a one-minute long turn based on pictures, showed the most deviation from what was actually predicted. The FCE materials hypothesized that candidates would engage in *giving information* and *expressing opinions through comparing and contrasting*. While these speech functions did occur in the data analyzed, candidates also engaged in *describing, expressing an opinion, expressing a preference, justifying* (an opinion, preference, choice, life decision), and *speculating*. In Fragment (14), the candidate spends most of her time speculating about the feelings of the people in each picture, as she was directed, but does not compare and contrast. Here is how Lazaraton and Frantz analyzed the response.

(14) *FCE 1996—Candidate 43 Examiner 377, Part 2*:
(Task: Couples: I'd like you to compare and contrast these pictures saying how you think the people are feeling)

1. yeah (.2) from the first picture I can see .hhh these two (.)
description
2. people they: seems not can:: cannot enjoy their .hhh meal (.)
speculation
3. because these girl's face I think she's: um (think) I think
justification **speculation**
4. she's: .hhh (.2) an- annoyed or something it's not impatient
5. and this boy: (.2) she's also (.2) looks boring (.2) yeah I I
speculation
6. think they cannot enjoy the: this atmosphere maybe the: .hhh
speculation
7. the:: waiter is not servings them (.) so they feel so (.) bored
8. or (.5) or maybe they have a argue or something like that (1.0)
speculation
9. yeah and from the second picture (.8) mmm::: this: rooms mmm:
description
10. looks very warm (.) and uh .hhh (.2) mmm thse two people? (.)
11. they also canno- I think they are not talking to each other
speculation
12. .hhh they just (.) sit down over there and uh (.5) these
description
13. gentleman just smoking (.) yeah and this woman just look at her
14. finger

These results proved useful to FCE test developers in understanding the relationship between assessment task design and candidate output. The list of 15 speech functions generated from the study also helped to inform the assessment criteria and rating scale descriptors for the FCE and other speaking tests.

Since the time of the previous study (Lazaraton and Frantz, 1997), Cambridge ESOL has developed other means to analyze the nature of candidate output in its speaking tests. One of the most promising avenues is the Observation Checklist (OC) developed by Saville (2000) and Saville and O'Sullivan (2000) and reported on in Weir and Milanovic (2003; see Appendix 2). This approach complements the use of fine-tuned transcripts as shown above, which require both expertise and a great deal of time to produce and analyze. As an instrument that can be used in real time, the OC allows for a larger number of performances to be scrutinized, thus providing more information for test development and interpretation. The features on the checklist were derived from spoken language, SLA, and the assessment literature, and can be characterized as *informational*, *interactional*, and *management of the interaction*. Based on piloting, revision, and a mapping of the checklists onto transcriptions of candidate talk, it was concluded that they were working well.

To summarize, then, discourse analysis, both in its fine-tuned CA form and its rougher functional analysis guise, is a tool that allows for a deeper understanding of the nature of talk in oral assessment contexts, which was for too long overlooked in the test validation process. As Lazaraton (2004) remarks:

> Conversation analysis has much to recommend it as a means of validating oral language tests ... Perhaps the most important contribution that CA can make ... is in the accessibility of its data and the claims based on them. That is, for many of us ... highly sophisticated statistical analyses ... are comprehensible only to those versed in those analytic procedures ... The results of CA are patently observable, even if one does not agree with the conclusions at which an analyst may arrive. As such language testers who engage in CA of test data have the potential to reach a much larger, less exclusive readership. (p. 65)

Of course, CA/DA is not without its shortcomings—theoretical, conceptual, and methodological—and has its detractors. It does require expertise, practice, and time, and its results demand sufficient space in publication outlets (often times more than is normally allowed). It is also unclear how this sort (in fact, many sorts) of qualitative research is to be judged—what are the criteria for evaluation? Clearly, this is a pressing issue for discussion, not just for language testing but for applied linguistics in general (see Lazaraton, 2003, for more on this topic).

It will be clear from the discussion so far that oral assessment is a complex business involving many different facets (e.g., task, test-taker, interlocutor,

rating criteria), all of which need to be well understood by test developers if the speaking tests they design are to be valid and reliable measuring instruments.

Process and Outcome in Writing Assessment

The same is true for the assessment of writing — another type of performance assessment involving multiple "facets" that make up a complex and interactive system (for more discussion of the multi-faceted nature of performance assessment see Milanovic and Saville, 1996). We turn now in this chapter to focus on writing assessment and more specifically the contribution of qualitative methodologies to our understanding of rater behaviour.

Understanding Rater Strategies in the Assessment of L2 Writing

Specialists in writing assessment have noted how far assessment criteria may be interpreted differently by different audiences depending on factors such as background, training, attitudes, and expectations (Hamp-Lyons, 1991). This phenomenon presents a challenge for test developers because rater consistency — both within and between raters — is considered an essential feature of good quality writing assessment. Our understanding of the processes by which raters arrive at a judgment about the quality of a performance (whether written or spoken) remains partial; getting inside a rater's head in order to access and observe the judgment process is not a simple matter! However, qualitative research methods offer a range of tools for exploring what happens during the rating process.

One of the most productive methodologies is Verbal Protocol Analysis (VPA).[†] A protocol is a "verbal report," or set of utterances, collected under special conditions and constituting a body of qualitative data for detailed analysis. While discourse analysis focuses on language content and structure, VPA involves using the language content and structure to look beyond the surface representation and to make inferences about the cognitive processes underlying certain behaviours. The assumption underpinning VPA is that "information that is heeded as a task is being carried out is represented in a limited capacity short term memory, and may be reported following an instruction to either talk aloud or think aloud" (Green, 1998, p. 7). The methodology comprises a number of distinct phases, and studies that use VPA can select from a range of different procedural variations — talk aloud/think aloud; concurrent/retrospective; non-mediated/mediated (for a fuller explanation of the phases, and the advantages/disadvantages of procedural variations, see Green, 1998, or Cohen, this volume).

[†][Ed. note: For an extensive discussion of verbal reports, see Cohen, Chapter 5.]

Table 6.1: Coding categories

Category 1: Marking behaviours — each covering a set of activities	
A1	Rater's initial reaction to script — overview (e.g., *This is laid out well*)
A2	Rater comments on their approach to reading the script (subcategories i–v)
A3	Rater comments on assigning a mark to the script (subcategories i–v)
A4	Rater comments of personal nature (subcategories i–vii)
B1	Rater comments on arrangements of meaning (subcategories i–iii)
B2	Rater comments on appropriate use of language (subcategories i–v)
B3	Rater comments on technical features (subcategories i–viii)
B4	Rater comments on task realization (subcategories i–vii)
Category 2: Evaluative responses of the rater	
3	positive response
2	neutral response
1	negative response
0	query
Category 3: Metacomments	
This "catch-all" category covered comments made by raters on their own rating technique, and miscellaneous comments that could not easily be coded using the other codes above, e.g., "I'm having some problem deciding what to do here."	

VPA has often been used in studies exploring rater strategies to address research questions such as: What approaches do raters adopt in the rating process? What features do they pay attention to? How do they use the available assessment criteria? How do they arrive at a judgment? What distinguishes 'good' raters from 'poor' raters? Do raters adjust their marking behaviour according to the level of the script? A study to explore some of these questions was conducted by Cambridge ESOL in relation to one of its proficiency tests of L2 writing at upper intermediate level (Milanovic and Saville, 1994). The study is described in some detail below to illustrate how VPA was used.

A dataset of 20 composition scripts produced by test-takers in a live administration of the Certificate in Advanced English (CAE) was selected. The writing task itself and the assessment criteria were carefully analyzed to hypothesize likely rater strategies. A group of 20 raters received pre-training and on-site training in VPA techniques and were then asked to rate the set of 20 scripts, recording their marks on a 0–5 scale and verbalizing their thoughts onto audio-cassette throughout the rating process: in other words, the main procedure adopted was "think-aloud," "concurrent," and "non-mediated" (Green, 1998, pp. 4–7). Raters were also asked to write a short retrospective report and

to complete a questionnaire that would provide supplementary data for analysis. The audio-taped verbal reports were transcribed and their linguistic content reviewed impressionistically in order to develop a framework of coding categories; the aim was to identify and label groups of rater activities or processes represented within the transcripts, which would constitute a workable coding system. Three overarching coding categories emerged, each subsuming a number of main group (e.g., A1, A2, etc.) and subsidiary categories (e.g., Ai–v) as shown in Table 6.1.

Each verbal protocol transcript was then segmented into separate units, with the unit for analysis defined as a clause, phrase, sentence, or group of sentences that identify and mark a boundary; each segment therefore represented a different activity/process. An example of a segmented protocol is shown in (15).

(15) *Transcript A/01*:

001 Beginning with the note to Malcom. Dear Malcom. Thank you for your last letter. I am really sorry that this paper got everything wrong and that you have got to bear the consequences. I hope people still talk to you. Please find enclosed a copy of the letter for which you were asking. I hope it meets your expectations. I did not put too much emphasis on all facts which were wrong but I tried to stress you have not been the mugger. /

002 Not too bad this note. /

003 Dear Sirs, Your article 'Handbag Thief Caught' from Wednesday May 27 1992. I address to you referring to your paper's article mentioned above. As I read it I was quite astonished and I would like to draw your attention to some misunderstandings which occur in this article. You give the impression that it was Mr Malcom Taylor who attempted to steal the woman's handbag. This is completely wrong. It was me, a German, not an American who accompanied Mr Taylor at the evening in question. Therefore please allow me to state shortly what has really happened. Mr Taylor and I were on our way home from the cinema as we saw a young man who attempted to snatch the handbag of a woman passing by. It was me who tried to help the woman. As a result I suffered a cut to my face. Fortunately Miss Erskine has not been injured in this incident but unfortunately the thief managed to escape. Please note that Mr Taylor's reputation suffered from your incorrect report I want to kindly ask you to publish this letter to put things right. I may allow me to thank you in advance. I hope I could help you in this matter. Write soon. /

004 He says this is completely wrong, so is this! /

005 He says he suffered the cuts, that's novel, /

006 the tenses are all over the place and it's a bit heavy. /

007 Malcom's letter is better, the letter to Malcom is better than the one to the editor. I quite like some of it, /

008 'I was quite astonished'. I quite like that but he gives the impression that he was the thief./

009 'You give the impression that it was Mr Malcom Taylor who attempted to steal the woman's handbag. This is completely wrong. It was me....' /

010 I had to go on reading, he hasn't really done the task that he should. /

011 I would give him 2. /

Each segment of a transcribed protocol was then assigned an appropriate code (i.e., Segment 001 = A2i, 002 = A4v3); all the data were coded as objectively and unambiguously as possible, and inter-coder reliability was estimated. Segmentation and coding of the transcript data made it possible to analyze frequency counts for main group and subsidiary categories by individual rater as well as across the whole rater group. It was also possible to subdivide scripts into *high-*, *middle-*, and *low-scoring* subsets and to explore group/individual differences in rater behaviour in terms of the main group and subsidiary coding categories. Results of the analyses suggested that *better* scripts elicit from raters attention to details such as register, style, layout, and content features. However, as performance quality declines, raters focus less on these features and pay more attention to composition elements such as spelling, grammatical accuracy, task understanding, and task completion (Milanovic and Saville, 1994).

In another study of rater decision-making processes during holistic marking of writing, 16 raters rated 40 compositions from two of the Cambridge ESOL exams at intermediate (FCE) and advanced Certificate of Proviciency in English (CPE) levels (Milanovic, Saville, and Shuhong, 1996). After an initial training session, three types of data were collected from the raters: a retrospective written report (i.e., after rating each composition raters noted what had gone through their minds in reaching a judgment); a concurrent verbal report (i.e., raters audio-recorded a verbal report while rating a limited number of compositions); and a group interview (i.e., a 30-minute structured interview). Data from the audio-taped concurrent reports and group interviews were transcribed; along with the retrospective written reports, these were reviewed impressionistically to develop a suitable scheme for coding the transcripts. Analyses of the coded data resulted in findings relating to two dimensions: firstly, the broad approaches taken by raters to the process of marking, and secondly, the details of their approach, i.e., the particular composition elements raters claim to focus on and the relative weight they claim to attach to these elements.

The four broad rating approaches identified in this study were: 1) principled two-scan/read, 2) pragmatic two-scan/read, 3) read-through, and 4) provisional mark. In terms of their detailed approach, raters appeared to focus on 11 composition elements: length, legibility, grammar, structure, communi-

cative effectiveness, tone, vocabulary, spelling, content, task realization, and punctuation.[3]

Results from VPA studies such as those we have described can play an important role in informing test developers' decisions on design of writing prompts/tasks and selection of writing assessment criteria, especially at different proficiency levels; such studies also inform and improve procedures for rater training and standardization. VPA was also used, for example, at the end of the IELTS Writing Revision Project (2001–2005) to check raters' interpretation and application of the new assessment criteria and to confirm how the revised rating scale was functioning (Falvey and Shaw, 2006).

Conclusion

It has become increasingly apparent that the established psychometric methods for test validation are effective, but limited, and other methods are required for us to gain a fuller understanding of the language tests we use. Two decades ago Cohen (1984) and Grotjahn (1986) advocated the use of introspection techniques as a means of gathering information to feed directly into the test development and validation process. Since then introspection techniques, such as verbal protocol analysis, together with other qualitative research methods, such as conversation and discourse analysis, have been increasingly used by applied linguists and language testers to explore a range of test-related issues.

We have chosen in this article to focus specifically on some of the qualitative studies in speaking and writing assessment undertaken by Cambridge ESOL in order to illustrate how outcomes from such investigations can feed directly into operational test development and validation activity. It should be noted, however, that many qualitative studies have been conducted with speaking and writing tests other than the Cambridge examinations. Conversation analysis, discourse analysis, and verbal protocol analysis — along with other qualitative research methods — now offer language testers viable solutions for these validation tasks.

Notes

[1] See Lazaraton (2002) for details on and citations for these studies.

[2] University of Cambridge ESOL Examinations is a non-teaching department of the University of Cambridge and offers language proficiency examinations in English. The general English exams referred to in this chapter are linked to the levels of the *Common European Framework of Reference* published by the Council of Europe (2001) in the following way: KET (A2), PET (B1), FCE (B2), CAE (C1) and CPE (C2).

[3] Space restrictions prevent a fuller description of these results; see Milanovic, Saville, and Shuhong, 1996, for more information.

Appendix 1

Transcription Notation Symbols

(adapted from Atkinson and Heritage, 1984, pp. ix–xvi)

1. **unfilled pauses or gaps:** periods of silence, timed in tenths of a second by counting "beats" of elapsed time. Micropauses, those of less than 0.2 seconds, are symbolized (.); longer pauses appear as a time within parentheses: (.5) is five tenths of a second.
2. **colon (:):** a lengthened sound or syllable; more colons prolong the stretch.
3. **dash (-):** a cut-off, usually a glottal stop.
4. **.hhh:** an inbreath; **.hhh!** — strong inhalation.
5. **hhh:** exhalation; **hhh!** — strong exhalation.
6. **hah, huh, heh, hnh:** all represent laughter, depending on the sounds produced. All can be followed by an (!), signifying stronger laughter.
7. **(hhh):** breathiness within a word.
8. **punctuation:** markers of intonation rather than clausal structure; a period **(.)** is falling intonation, a question mark **(?)** is rising intonation, a comma **(,)** is continuing intonation. A question mark followed by a comma **(?,)** represents rising intonation, but is weaker than a (?). An exclamation mark **(!)** is animated intonation.
9. **equal sign (=):** a latched utterance, no interval between utterances.
10. **brackets ([ ]):** overlapping talk, where utterances start and/or end simultaneously.
11. **percent signs (% %):** quiet talk.
12. **caret (^):** a marked rising shift in pitch.
13. **arrows (> <):** the talk speeds up, **arrows (< >)** — the talk slows down.
14. **psk:** a lip smack, ***tch*** — a tongue click.
15. **underlining or CAPS:** a word or SOund is emphasized.
16. **arrow (→):** a feature of interest to the analyst.
17. **empty parentheses ( ):** transcription doubt, uncertainty; words within parentheses are uncertain.
18. **double parentheses (( )):** non-vocal action, details of scene.

Appendix 2

Observation checklist for speaking test validation

(from Weir and Milanovic, 2003, p. 453)

Informational functions

Providing personal information	• give information on present circumstances • give information on past circumstances • give information on future plans
Expressing opinions	express opinions
Elaborating	elaborate on or modify an opinion
Justifying opinions	express reasons for assertions s/he has made
Comparing	compare things/people/events
Speculating	speculate
Staging	separate out or interpret the parts of an issue
Describing	• describe a sequence of events • describe a scene • describe a person
Summarising	summarise what s/he has said
Suggesting	suggest a particular idea
Expressing preferences	express preferences

Interactional functions

Agreeing	agree with an assertion made by another speaker (apart from 'yeah' or non-verbal)
Disagreeing	disagree with an assertion made by another speaker (apart from 'no' or non-verbal)
Modifying	modify arguments or comments made by another speaker or by the test-taker in response to another speaker
Asking for opinions	ask for opinions
Persuading	attempt to persuade another person
Asking for information	ask for information
Conversational repair	repair breakdowns in interaction
Negotiating meaning	• check understanding • indicate understanding of point made by partner • establish common ground/purpose or strategy • ask for clarification when an utterance is misheard or misinterpreted • correct an utterance made by another speaker which is perceived to be incorrect or inaccurate • respond to requests for clarification

Managing interaction

Initiating	start any interactions
Changing	take the opportunity to change the topic
Reciprocating	share the responsibility for developing the interaction
Deciding	come to a decision

Example of the observation checklist used to analyse functional content across the 4 parts of a paired speaking test
(Test 1 FCE, 1998-99, Candidates A and B)

	OPERATIONS		Task 1		Task 2		Task 3		Task 4	
			A	B	A	B	A	B	A	B
INFORMATIONAL	Providing personal information	Present	√	√	√	√	√	√	√	√
		Past		√						
		Future	√	√						
	Expressing opinions		√	√	√	√	√	√	√	√
	Elaborating		√	√	√	√	√	√	√	√
	Justifying opinions		√		√	√	√	√	√	√
	Comparing		√		√	√	√	√		
	Speculating		√	√	√				√	√
	Staging		√		√					
	Describing		√	√	√	√				
	Summarising									
	Suggesting						√		√	√
	Expressing preferences		√	√	√	√		√		
INTERACTIONAL	Agreeing						√	√	√	√
	Disagreeing						√	√		
	Modifying						√		√	
	Asking for opinions						√	√		
	Persuading						√	√		
	Asking for information									
	Conversational repair						√	√		
	Negotiating meaning	Check understanding					√			
		Indicate understanding					√	√		
		Establish common ground								
		Ask for clarification		√					√	
		Correct an utterance								
		Respond to requests for clarification								
Managing interaction	Initiating						√			
	Changing									
	Reciprocating						√	√		
	Deciding									

This page intentionally left blank