Qualitative Methods of Validation

Anne Lazaraton

Abstract

It is not surprising that there are continuing tensions between the disciplinary research paradigms in which language testers situate themselves: psychometrics, which, by definition, involves the objective measurement of psychological traits, processes, and abilities and is based on the analysis of sophisticated, quantitative data, and applied linguistics, where the study of language in use, and especially the construction of discourse, often demands a more interpretive, qualitative approach to the research process. This chapter looks at qualitative research techniques that are increasingly popular choices for designing, revising, and validating performance tests – those in which test takers write or speak, the latter of which is the primary focus of this chapter. It traces the history of qualitative research in language testing from 1990 to the present and describes some of the main findings about face-to-face speaking tests that have emerged from this scholarship. Several recent qualitative research papers on speaking tests are summarized, followed by an examination of three mixed methods studies, where both qualitative and quantitative techniques are carefully and consciously mixed in order to further elucidate findings that could not be derived from either method alone. I conclude by considering challenges facing qualitative language testing researchers, especially in terms of explicating research designs and determining appropriate evaluative criteria, and speculating on areas for future research, including studies that tap other methodological approaches such as critical language testing and ethnography and that shed light on World Englishes (WEs) and the Common European Framework of Reference (CEFR).

A. Lazaraton (✉)
Department of Writing Studies, University of Minnesota, Minneapolis, MN, USA
e-mail: lazaratn@umn.edu

E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language
and Education, DOI 10.1007/978-3-319-02261-1_15

Keywords
Discourse analysis • Introspective techniques • Mixed methods research • Speaking/oral proficiency assessment

Contents

Introduction

In a recent position paper, McNamara (2011) claimed that "the distinctive character of language testing lies in its combination of two primary fields of expertise: applied linguistics and measurement" (p. 435). He further noted that language testers come to the discipline from one of two homes (rarely both): psychometrics and statistics, or applied linguistics, in which a major intellectual preoccupation is the various facets of language in use. Seen in this way, it is not surprising that there are continuing tensions between the research paradigms in which with these disciplines are situated. Psychometrics, by definition, involves the objective measurement of psychological traits, processes, and abilities and most often employs sophisticated, quantitative data collection, analysis, and interpretations. In contrast, the study of language in use, and especially the construction of discourse, often demands a more interpretive, qualitative approach to the research process. While language assessment research remains primarily a quantitative endeavor focused on product (i.e., scores), an important methodological development over the 25 years has been the appearance of qualitative research methodologies to assist in language test design, revision, and validation. This is especially so for performance testing, in which test takers speak and/or write; qualitative research techniques *have* become more prominent in the discipline in order to understand the processes and manifestations of language use in assessment, particularly tests of oral proficiency. In this chapter I trace major historical developments in qualitative language testing research, consider several research traditions that have been employed in such research, analyze a number of published studies that adopt these techniques, problematize the qualitative research endeavor in language testing, and look ahead to its future.

Early Developments

As reported in Lazaraton (2008), from a methodological standpoint, two periods of language testing research characterize the field: pre-1990, when almost all scholarship employed a positivistic, outcome-based framework, and post-1990, after which a great deal of attention was directed to understanding the processes of performance testing. The dividing line between these periods emerged with the publication of Leo van Lier's (1989) seminal paper, in which he questioned the assumed but untested premise that the discourse produced in face-to-face, direct speaking tests involving interlocutors and test takers is, essentially, "natural conversation." He urged the language testing community to investigate not only the construct of oral proficiency but also the processes that underlie its demonstration. A second paper by applied linguists Jacoby and Ochs (1995) explored "co-construction," defined as "the joint creative of a form, interpretation, stance, action, activity, identity institution, skill, ideology, emotion, or other culturally meaningful reality" (p. 171). In response to their description of co-construction and van Lier's call for research on speaking test discourse, language testing researchers have undertaken numerous empirical studies that investigate many aspects of oral (and to a lesser extent, written) proficiency assessment.

Using qualitative discourse analytic techniques (discussed in the next section), several books published around the turn of the century investigated the nature of oral testing talk by the interviewer, the test taker, and, most notably, in the interviewer and the test candidate's "co-constructed" discourse (Lazaraton 2002). From this (and other) earlier work, we have learned that:

- Interviewers, through their talk and behavior (such as supplying answers, simplifying task directions, completing or correcting test taker responses, and rephrasing questions), bring unpredictability into the encounter, thus threatening test reliability.
- Test takers do not always produce the sorts of language or use it in ways that the test developers predict intuitively in test design.
- As a genre, language assessment interviews do "share features with conversations [but] they are still characteristically instances of interviews of a distinctive kind for the participants (Lazaraton 2002, p. 15).
- Pair and group orals, where test takers talk with one or more other test takers (instead of or addition to engaging with the interviewer), have gained popularity for several reasons: they approximate pair and group class activities; the power differential between interviewers and test takers is reduced; and a broader range of speech functions are displayed in peer talk when compared to interviewer-test taker talk.
- Pair and group test talk has been shown to be influenced by gender, personality, proficiency, and acquaintanceship, but the relationship between these variables, discourse produced, and outcome test scores is still not well understood.

In reaction to and with the help of these findings, testing organizations such as Cambridge English (www.cambridgeesol.org) instituted various refinements to their speaking tests, including the development and use of an "interlocutor frame" – an interview agenda – to guide the interviewer and the revision of rating scale descriptors to more accurately reflect the nature of test taker discourse and speech functions produced (see Taylor and Galaczi 2011 on other ways that Cambridge English has engaged in a continuous validation process for its oral assessments).

These efforts are part of an ongoing effort to keep the concept of validity and the process of validation front and center in language assessment research. According to Kane (2012), validation boils down to two questions: "What is being claimed? Are these claims warranted?" based on the evidence provided (p. 4). In a traditional sense, validity claims made by language testers are based on evidence that the assessments they design and use present a true picture of the construct being measured – for example, interactive communication or extended discourse – a task for which qualitative research techniques are ideally suited. On the other hand, an "argument-based" validation approach requires "specification of the proposed interpretations and uses of test scores and the evaluating of the plausibility of the proposed interpretative argument" (Kane, p. 3). A more detailed explication of validation is beyond the scope of this chapter; suffice it to say that "validation is simple in principle, but difficult in practice" (Kane, p. 15). The authors of the studies summarized below utilize qualitative research methods to grapple with test validation concerns for assessment interpretation and use.

Major Contributions

In this section, I first provide background on the most widely used qualitative approach to understanding the process and outcomes of oral testing, namely, discourse analysis, followed by summaries of three recent studies that illustrate some current research in this area. Next, I overview a second qualitative research methodology that some language testers have utilized, introspective methods. Finally, I consider the principles of mixed methods research, a methodological choice that is increasingly prevalent in language assessment (LA) scholarship.

Discourse Analysis

The most widely used qualitative approach to understanding the output of oral performance testing is discourse analysis, which traces its roots to various disciplines – primarily anthropology, linguistics, philosophy, psychology, and sociology – and can be construed broadly as an endeavor with several defining characteristics. Generally speaking, discourse analysis:

- Relies on careful transcription of authentic spoken discourse, a laborious yet fruitful part of the research process

- Accounts for and responds to the importance of context, in its broadest sense
- Produces a rich, deep analysis based on intensive engagement with the data
- Reflects one or more theories of language in use, such as accommodation theory and conversation analysis (CA)

Briefly (but see, e.g., Sidnell and Stivers 2013), conversation analysis investigates instances of "talk in interaction" about which the analyst is ideally agnostic at the outset of an investigation. The unit of analysis is the speaker turn, a central analytic construct that drives the careful transcription of the discourse to be studied. As phenomena emerge as interesting, the researcher will focus on collecting single cases that demonstrate the phenomena as well as deviant cases that should, but don't (or that do, but shouldn't). CA is insistent that assertions about participants' backgrounds, gender, and other demographic factors not be assumed as automatically relevant to the discourse being analyzed; the researcher must show how various identities are (co-) constructed, displayed, or withdrawn at a particular point in the talk. Finally, in its pure form, CA resists coding and counting data because the focus is that of a microscope looking at a single case rather than a telescope that captures phenomena at an aggregate level. Nevertheless, in subsequent sections I detail some recent research where CA data are tagged and coded in order to define and delineate constructs or to test them psychometrically. First, however, I summarize three qualitative studies that investigate facets of test discourse in face-to-face speaking tests involving one interviewer and one or more test takers.

The research of Gan (2010) and Luk (2010) centers on the production of test taker talk in group and pair orals. Gan (2010) investigated the nature of speaking test performance in higher- and lower-proficiency test taker groups in his case study of a school-based oral assessment in Hong Kong. Gan was interested in detailing the interactional features that characterized the discussions of two small groups of four secondary-level English as a foreign language (EFL) students. Gan's data consisted of fine-grained transcriptions of group discourse analyzed according to CA principles, which involved a line-by-line, sequential analysis of extended discourse. His findings revealed that the higher-proficiency group produced collaborative talk that was both "constructive and contingent": participants jointly built "opportunities for substantive conversation and genuine communication" (p. 585) with other group members in managing the conversational floor and engaging with their peer's ideas. On the other hand, the lower-proficiency group demonstrated less engagement with each other's ideas and more interactional work devoted to creating and maintaining a helpful, non-threatening discourse environment. The linguistic issues that arose in these discussions served as the basis for collaborative dialogue and assistance, but overall, their talk did not show the "contingent development of topic talk" (p. 585) that characterized the higher-proficiency group. Gan concluded that this sort of research, which focuses on the social, interactive nature of face-to-face speaking test performance, is crucial for making a validity argument about the particular assessment.

Luk's (2010) report on a group oral assessment in Hong Kong considered a different factor, that of impression management, which she defined as "a social psychological notion that describes the process through which people consciously or unconsciously try to control the impression other people form of them so as to achieve a certain goal" (p. 27). In a very comprehensive methods of analysis section, Luk characterizes her research as "applied CA," where discourse analytic findings are supplemented with information from questionnaires and interviews. In her school-based assessment (SBA), 11 groups of four female secondary school students engaged in discussions about a text they read or a film they watched based on teacher-constructed task prompts; the classroom teacher and six students were also interviewed about their experiences. All participants also completed a questionnaire. Her findings fell into three thematic categories, including task management, content delivery, and "converging speech acts" in which conversational sequences are constructed (such as question-answer). Luk found the interactions were characterized by highly ritualized openings and closings, orderly turn taking, negotiation of meaning avoidance, and responses to fill dead air, among other features. In the end, Luk observed a "strong desire on the part of the students to maintain the impression of effective interlocutors for scoring purposes rather than for authentic communication" (p. 25). That is, participants' interactions were performative and, at times, even "collusive," where test takers rehearsed their responses beforehand. She suggests that test designers reexamine the validity of group oral assessment when test takers only speak with each other because of the possibility of planned performances.

One of the most intriguing recent papers on oral language assessment is Norton's (2013) research on speaking test talk that goes beyond previous work on the test participants to include a third, largely unexplored factor: the presence of an interlocutor frame and various testing materials. Norton employs the post-structuralist concepts of intertextuality and interdiscursivity to analyze speaking test data from two Cambridge English exams, the First Certificate of English (FCE) and the Certificate of Advanced English (CAE). Her goal was to understand the identities that speaking test participants construct by means of talking with another person and dealing with interlocutor frames and other test materials while also accounting for "the myriad of other 'voices'" present in these interactions, such as test designers who develop the assessment (p. 309). As such, her paper represents a critique of some current testing practices that were employed as solutions to problems raised in earlier investigations.

For example, past work has shown that the interviewer can be a positive, neutral, or negative factor in the discourse test takers produce; certain interviewer practices may lead to unreliable outcomes. As a result, testing agencies such as Cambridge English instituted a set of interlocutor frames that dictate what interviewers can say, so that "unscripted questions or comments should always be kept to the absolute minimum" (UCLES, 1996; as cited in Norton, p. 316). Comparing the written form of the interlocutor frame with the actual talk interviewers produced, Norton found their discourse contained "numerous deviations" from the dictated format. By including this additional, unscripted material, interviewers displayed a sort of "hybrid identity" of both teacher and examiner because "certain interviewers find it difficult to identify themselves as institutionalized, unindividuated, noninteractive subjects" (p. 316).

A further problem arose with test taker discourse: when candidates are told to say as much as they can in a set period of time, unless they are test savvy, they may confuse the need to produce a ratable sample of language with their need to be truthful. This makes it difficult to distinguish "cannot talk" vs. "will not talk" candidates who respond truthfully rather performatively by engaging verbal behaviors that may be appropriate in conversation but are detrimental to ratings of test talk. Norton's evidence strongly suggests that task design itself is implicated in co-construction of performance and must be accounted for as such; she concludes that it is "intrinsically problematic . . . to impose such a framework to elicit language for assessment purposes when the framework itself may limit participation in speaking tests in ways which cannot be easily predicted" (p. 325). She recommends that testing organizations ensure that all candidates understand assessment criteria, including the desirability of initiating topics and expanding on answers to produce a sufficient sample of language for rating purposes.

Introspective Techniques

While these discourse analytic studies looked at the language produced in the speaking test context, another qualitative research has explored the cognitive processes in which raters engage when assessing language production in performance tests. Often this work utilizes *introspective methods*, which aim to generate usable data on cognition during or after a particular task. Sasaki (2014; see also Green 1998) describes such techniques, including *think-alouds*, where participants articulate their thoughts while engaged in a task, and after-the-fact *recalls* which can be stimulated with a memory aid or collected alone. Sasaki contends that analyses of these verbal protocols are ideal for complementing, rather than merely supplementing, more quantitative analytic techniques. In other words, such inquiry "can also contribute to knowledge accumulation in the LA [language assessment] field by adding a harvest of studies with nonpositivist perspectives that are quite different from those that have hitherto prevailed in the field" (p. 16).

Three such studies are illustrative. In research looking at how oral proficiency ratings may be influenced by rater accent familiarity, defined as "gained through having learned the first language (L1) of the test takers as an L2 in the past" (p. 770), Winke and Gass (2013) asked 26 trained raters to assess Internet-based Test of English as a Foreign Language (TOEFL) speaking test samples and then reflect on their rating processes. The raters, who were native speakers (NS) of Chinese, Korean, or Spanish, engaged in 20–30 min stimulated recalls while viewing their rating sessions. The authors analyzed the recalls using an analytic inductive approach, which generated a total of eight themes, three of which were related accent familiarity: the test takers' L1, the test takers' accent, and the raters' heritage status. They concluded that "although sensitivity to test-taker accents seemed to occur naturally in the rating process, findings suggest that when raters have learned or know, to varying degrees, the test takers' L1, they tend to orient themselves to the speech in a biased way, compromising test reliability" (p. 762).

Rater thought processes were also analyzed by May (2011), who employed stimulated verbal recalls to understand the interactional features that were salient to raters of 12 paired speaking tests. Along with rater notes and data from rater discussions, her analysis indicated that a number of features were noticed, including interpreting and responding to another's message, working together, and adding to the authentic interaction taking place. May's concern was that it is difficult to evaluate an individual's performance in a co-constructed interaction; as a result, paired orals are not necessarily a panacea to the problems encountered in more traditional interviewer-test taker assessments. In other words, pair and group oral assessments have potential validity problems of their own.

Finally, rater cognition is also of interest in writing assessment. Li and He (2015) utilized primarily qualitative, introspective methods in their study of an analytic and a holistic scale used by nine raters of ten essays produced for the Chinese College English Test. The authors focused on how the rating scale type appeared to influence rating strategies and the textual features focused on by raters. Li and He used think-aloud protocols, questionnaires, and semi-structured interviews to collect their data. They found that holistic scales led to more interpretation strategies; judgment strategies were more prevalent with the analytic scale. Overall, the authors suggest that holistic scales force raters to focus on more limited set of text features and to adopt essay comparisons strategies. Additionally, the lack of detailed descriptors in the holistic scale led to rater difficulties in defining and assessing the construct, leading to less reliable and valid scoring.

Work in Progress

A more recent scholarly trend, increasingly apparent in the last 10 years, is the emergence of mixed methods research (MMR) in language assessment. MMR involves the conscious, principled mixing of quantitative and qualitative methodologies and analyses, and in language assessment, to arrive at a more comprehensive understanding of performance test factors in test talk, rater cognition, and the perspectives and beliefs of test stakeholders. The roots of MMR can be traced to the concept of "triangulation" in qualitative research (see Creswell 2014, an essential source on MMR); as a research strategy, triangulation involves one or more of the following: obtaining multiple sources of data, including multiple groups of participants, and employing multiple research techniques. According to Turner (2014), a basic premise of MMR is that qualitative and quantitative researches are not incompatible, but complementary in their strengths and weaknesses. Although mixing methods has been going on for a long time, it wasn't until around 2003 when a set of guiding principles started to take shape. Nevertheless, despite "increasing evidence of [LT] research employing both qualitative and quantitative approaches, [but] specific articulation of employing an MMR is still rare" (Turner 2014, p. 4). Both Turner and Brown (2014a) articulate various design types for MMR, each of which is a permutation of temporal elements – concurrent and sequential – in terms of data collection and analysis and of research goals as exploratory or explanatory.

However, different authors array these factors using sometimes dissimilar vocabulary, so it is not always a simple manner to compare research designs in the absence of informative visuals, as the two studies described below include.

One of the first published MMR research papers in language testing is Kim's (2009) examination of the differences in proficiency judgments between native speaker and nonnative speaker (NNS) teachers of English. The impetus for his research was that previous quantitative analyses of rater behavior and English language background were not sufficiently fine-grained. A total 12 Canadian and 12 Korean teachers of English rated semi-direct speech samples (where test takers speak into a recorder rather than with a person) from ten college-level English as a second language (ESL) students performing a total of eight speaking tasks. Rater behavior was analyzed using multifaceted Rasch; rater comments on student performance were analyzed qualitatively. Teacher comments were open coded, resulting in 19 recurring criteria. Kim's findings indicated that both groups of teachers showed good internal consistency in their ratings and a similar harshness pattern. It was in the comments about evaluative criteria where the rater groups exhibited notable differences. The Canadian teachers produced a larger number of comments, and they were more detailed and elaborate than those from the Korean teachers, although both groups were most concerned with vocabulary, pronunciation, and overall language use. Kim cautiously interprets his findings regarding the qualitative differences he detected, hypothesizing that nonnative speakers (NNS) aren't always trained to assess details of performance and may come from different evaluation cultures. In any case, what is most notable is the degree to which Kim carefully explains his research design, sampling, and analytic techniques using accessible terminology and provides a very helpful diagram of the research procedures.

Youn (2015) is a second exemplary inquiry employing mixed methods to develop a validity argument for assessing second language (L2) pragmatics. His research used discourse data from open role-plays to inform task design and rating criteria development and also analyzed rater performance with FACETS. A total of 102 students and four native speakers engaged in role-plays, discourse from which was transcribed and analyzed using CA methods in order to "back[ing the] valid task design and sound rating criteria assumptions" (p. 203). Five "interaction-sensitive, data-driven" rating criteria were derived from this analysis: contents delivery, language use, sensitivity to situation, engaging with interaction, and turn organization, each of which Youn illustrates with a relevant data fragment that depicts how the rating criteria were derived. In Youn's words, "the CA findings helped examine a degree of authenticity and standardization of the elicited performances along with detailed descriptions for rating criteria" (p. 203); "the mixed methods generated convincing backing for the underlying assumptions of the evaluation inferences" (p. 218). Like Kim's work described above, Youn presents two informative diagrams and figures: the first represents the evaluation inference schematically, and the second depicts the study design.

Finally, mixed methods were also used by Zhao (2013) to develop and validate rubric for measuring authorial voice in L2 writing. Four raters assessed authorial voice in 200 TOEFL iBT writing samples using a preliminary rubric containing

11 features rated on a 0–4-point scale. In the development phase, the author used principal components analysis of the resulting ratings to generate a set of construct dimensions to inform the creation of the scoring rubric. She also collected and analyzed think-aloud protocols and interview data "to supplement the quantitative analysis and provide additional evidence on rubric reliability, applicability, and construct validity" (p. 205). The qualitative data were then used to create the final authorial voice rubric used in the validation phase of the study. Her findings indicated that both the quantitative and qualitative data supported a three-dimensional conceptualization of voice. Zhao concludes that a rater's thoughts and feelings about the overall quality of voice in a writing sample are less a matter of the *quantity* of individual voice elements in the text; more meaningful to raters is *how* they are used.

As promising as mixed methods research appears to be for language assessment, there are challenges associated with the approach that testers must consider.

Problems and Difficulties

There are indeed unique concerns in MMR research that must be accounted for. For one, it is not always clear how research questions should be formulated and ordered or prioritized: Separately? Sequentially? Overarching? And what is sampling process? How should MMR be evaluated? And who has the expertise to engage in both qualitative and quantitative researches (Turner 2014, pp. 10–11)? In any case, as language assessment research designs have become more complex, it is even more important for scholars to include visuals that represent a sometimes nonlinear research process; this requirement should be taken more seriously in presenting and publishing MMR research.

Along with the inclusion of schematics and visuals, it is essential for researchers to explicate clearly the framework being employed. Triangulation as a research strategy is a very good one, but simply including multiple methods in an investigation is not the same as engaging in rigorous mixed methods research. "MMR uses a specific logic, especially the *fundamental principle of MMR*, i.e., 'the research should strategically combine qualitative and quantitative methods, approaches, and concepts in a way that produces complementary, strengths and nonoverlapping weaknesses'" (Johnson et al. 2007; as cited in Brown 2014a, p. 9; emphasis in original). Brown astutely notes that "if the qualitative methods and quantitative methods are simply used simultaneously or sequentially, with them not interacting in any particular ways, the research might be more aptly labeled multimethod research" (p. 9).

More broadly, evaluative criteria for qualitative and mixed methods language testing research must be developed and/or refined. As a research community, language testers have a long history of evaluating positivist, quantitative research according to established criteria such as validity (the current intellectual preoccupation; see Kane 2012), reliability, replicability, and generalizability. We also have some familiarity with their qualitative counterparts of dependability, credibility, confirmability, and transferability, but there is still misunderstanding of and debate about how to weigh and describe these criteria – or if they are even the right criteria. Enter mixed methods, an even more complicated endeavor. Brown (2014a) maintains that evaluative criteria for both quantitative and qualitative researches have thematic parallels: *consistency* captures both reliability and dependability; validity and credibility are both concerned with *fidelity*; *meaningfulness* characterizes generalizability and transferability; and *verifiability* subsumes both replicability and confirmability (p. 119). While this heuristic is certainly helpful in setting out the correspondences between quantitative and qualitative research, it is uncertain how, and to what degree, language testers understand these meta-concepts, much less the central and complex issue of validity in mixed research, "legitimation," which "is to MMR what *validity* is to quantitative research and *credibility* is to qualitative research" (Brown 2014a, pp. 127–128).

Future Directions

One area that continues to be fertile ground for language assessment research is in understanding the consequential validity of language tests. Surprisingly, impact studies of "their uses, effectiveness, and consequences" (Shohamy 2001, p. xvi) are not plentiful (but see, e.g., some of the papers in Shohamy and McNamara 2009; and O'Loughlin 2011, on the use and interpretation of IELTS scores in university admissions). Should critical language testing (CLT) be considered one type of qualitative research? It seems clear that a positivist paradigm, where objectivity and generalizability are valued goals, is not really consistent with the subjective, "lived" experiences that CLT would tap. Shohamy (2001) claims that numbers are symbols of "objectivity, rationalism, . . . control, legitimacy and truth" (p. x) and their power lies in the fact that they can be challenged only by using different numbers to counteract them; testers "own" the numbers. From this perspective, CLT and quantitative inquiry may well be incommensurable, but this position is unsupported by evidence and is only personal conjecture at this point.

Other research approaches, such as ethnography, are ideal for shedding light on classroom-based assessment (CBA) practices. For example, Hill and McNamara's (2012) ethnography of one primary and one secondary Indonesian as a foreign language classroom in Australia focused on assessment processes, especially in terms of the evidential, interpretive, and use dimensions and scope of CBA. The authors collected and analyzed "a diverse range of data" that established the processes in which classroom teachers engage, the materials from which they gained assessment information, and their views on language learning and assessment. Their data also shed some light on assessment from the learners' perspectives. A different sort of classroom was the locus for an ethnographic study by Tsagari (2012), who examined First Certificate in English (FCE) test preparation courses in Cyprus with the aim of explicating the "details of teachers' instructional behaviors and . . . descriptions of classroom practices" (p. 37) in order to understand the potential washback of the courses. Fifteen classroom lessons totaling 24 h of observation data across three schools along with other supplementary information were analyzed and then presented as data fragments. Tsagari's findings point to both positive and negative washbacks in the FCE test preparation classes. The amount of work dedicated to reading, listening, speaking, and writing was seen as positive impact, while the reading test format, the limited genre writing, and test-wise listening strategies narrowed the authenticity and applicability of the strategic practice in which students engaged.

Additionally, although there has been frequent talk about the role of World Englishes (WEs) in language assessment (see Brown 2014b), to date there is not much empirical inquiry that takes up WEs in a systematic way. Harding (2014) suggests that a "guiding principle of new research on the communicative competence construct must be a focus on "adaptability" . . .to deal with different varieties of English, appropriate pragmatics, and fluid communication practices of digital environments" (p. 194). In the global English context, Harding argues that the research agenda must include "the development and validation of language tests that specifically assess a test-taker's ability to deal with diverse, and potentially unfamiliar varieties of English. These tests would use as their basis a different range of skills and abilities including: ability to tolerate different varieties of English" (p. 194). He notes that paired and/or group speaking assessments, which may require lingua franca interaction, could provide such evidence; "discourse data yielded from tasks of this kind (complemented by stimulated recall performed by test-takers) could be analyzed with a view to locating points at which these abilities are tapped in these interactions" (p. 195).

One final area where qualitative research is underrepresented relates to the Common European Framework of Reference (CEFR). McNamara (2014) points out that while there are "a plethora of studies on applications of the CEFR in various contexts... few of these studies are critical in any important sense. Most are overwhelmingly and unquestionably technist and functionalist" (p. 228).

McNamara further criticizes much CEFR research because it fails to even mention English as a Lingua Franca (ELF); more broadly, "the lack engagement with the larger question of the role and function of the CEFR" (p. 229) is indicative of the lack of engagement with larger sociopolitical issues that are endemic in language testing. We can hope that such inquiry relies, at least in part, on qualitative research techniques that hold much promise for delving more deeply into test impact and consequences. If qualitative research techniques can be properly combined with more traditional quantitative methodologies that lend themselves capturing the scope of language assessment phenomena, all the better.

Cross-References

- ▶ Critical Language Testing
- ▶ Criteria for Evaluating Language Quality
- ▶ Methods of Test Validation

Related Articles in the Encyclopedia of Language and Education

Beatriz Lado; Cristina Sanz: Methods in Multilingualism Research. In Volume: Research Methods in Language and Education

Alastair Pennycook: Critical Applied Linguistics and Education. In Volume: Language Policy and Political Issues in Education

Li Wei: Research Perspectives on Bilingualism and Bilingual Education. In Volume: Research Methods in Language and Education

References

Brown, J. D. (2014a). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.

Brown, J. D. (2014b). The future of World Englishes in language testing. *Language Assessment Quarterly, 11*(1), 5–26. doi:10.1080/15434303.2013.869817.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Los Angeles: Sage.

Gan, Z. (2010). Interaction in group assessment: A case study of higher- and lower-scoring students. *Language Testing, 27*(4), 585–602. doi:10.1177/0265532210364049.

Green, A. (1998). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly, 11*(2), 186–197. doi:10.1080/15434303.2014.895829.

Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*(3), 395–420. doi:10.1177/0265532211428317.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28*(3), 171–183. doi:10.1207/s15327973rlsi2803_1.

Kane, M. (2012). Validating score interpretations and uses: Messick Lecture. *Language Testing, 29*(1), 3–17. doi:10.1177/0265532211417210.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217. doi:10.1177/0265532208101010.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Language testing and assessment 2nd ed., Vol. 7, pp. 197–209). New York: Springer.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly, 12*, 178–212. doi:10.1080/15434303.2015.1011738.

Luk, J. (2010). Talking to score: Impression management in L2 oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly, 7*(1), 25–53. doi:10.1080/15434300903473997.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127–145. doi:10.1080/15434303.2011.565845.

McNamara, T. (2011). Applied linguistics and measurement: A dialogue. *Language Testing, 28*(4), 435–440. doi:10.1177/0265532211413446.

McNamara, T. (2014). 30 years on – Evolution or revolution? *Language Assessment Quarterly, 11*(2), 226–232. doi:10.1080/15434303.2014.895830.

Norton, J. (2013). Performing identities in speaking tests: Co-construction revisited. *Language Assessment Quarterly, 10*(3), 309–330. doi:10.1080/15434303.2013.769549.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146–160. doi:10.1080/15434303.2011.564698.

Sasaki, M. (2014). Introspective methods. In A. Kunnan (Ed.), *Companion to language assessment.* Wiley. doi:10.1002/9781118411360.wbcla076.

Shohamy, E. (2001). *The power of tests: A critical perspective on the use of language tests*. New York: Pearson.

Shohamy, E., & McNamara, T. (2009). Language assessment for immigration, citizenship, and asylum. *Language Assessment Quarterly,6*(4). doi:10.1080/15434300802606440.

Sidnell, J., & Stivers, T. (Eds.). (2013). *The handbook of conversation analysis*. West Sussex: Wiley-Blackwell.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 171–233). Cambridge: Cambridge University Press.

Tsagari, D. (2012). FCE exam preparation discourses: Insights from an ethnographic study. *UCLES Research Notes, 47*, 36–48.

Turner, C. (2014). Mixed methods research. In A. Kunnan (Ed.), *Companion to language assessment*. Wiley. doi:10.1002/9781118411360.wbcla142.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508. doi:10.2307/3586922.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762–789. doi:10.1002/tesq.73.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199–225. doi:10.1177/0265532214557113.

Zhao, C. G. (2013). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing, 30*, 201–230. doi:10.1177/0265532212456965.