Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

Perceptions of Examiner Behavior Modulate Power Relations in Oral Performance Testing

India C. Plough & Pamela S. H. Bogart

To cite this article: India C. Plough & Pamela S. H. Bogart (2008) Perceptions of Examiner Behavior Modulate Power Relations in Oral Performance Testing, Language Assessment Quarterly, 5:3, 195-217, DOI: 10.1080/15434300802229375

To link to this article: https://doi.org/10.1080/15434300802229375

Published online: 23 Jul 2008.

LANGUAGE ASSESSMENT QUARTERLY, *5*(3), 195–217, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434300802229375

Routledge
Taylor & Francis Group

Perceptions of Examiner Behavior Modulate Power Relations in Oral Performance Testing

India C. Plough and Pamela S. H. Bogart
University of Michigan

To what extent are the discourse behaviors of examiners salient to participants of an oral performance test? This exploratory study employs a grounded ethnographic approach to investigate the perceptions of the verbal, paralinguistic and nonverbal discourse behaviors of an examiner in a one-on-one role-play task that is one of four tasks in an oral performance test. Candidates were international graduate students who were prospective teaching assistants. Video recordings of the test were reviewed in separate feedback sessions with the candidates, one participating and one nonparticipating examiner. These sessions were then reviewed by a researcher who was not involved in either the initial test or in the collection of feedback. Findings indicate that the verbal, paralinguistic, and nonverbal behaviors of examiners are indeed salient to participants and nonparticipants in the testing event. In fact, candidates found these behaviors to be meaningful in terms of their comfort level during the test task and in terms of their perception of the realism of the interaction. Drawing on the concept of footing, the various positions or roles taken by participants in an oral exchange, it is argued that these perceptions and interpretations may then serve to reduce the unequal power relationship inherent in oral performance tasks. The study suggests implications for the development of oral performance tests and for the training of examiners.

The results of a high-stakes oral proficiency test can have far-reaching consequences on the individual taking the test and the institution administering it. In the context of international graduate students seeking appointments as teaching assistants at colleges and universities in the United States, negative results on

Correspondence should be addressed to India C. Plough, English Language Institute, University of Michigan, Ann Arbor, MI 48104-2028. E-mail: indiac@umich.edu

these tests have the potential of ending an individual's education in the United States, which in turn affects the fabric and vitality of academic departments in universities around the country. It is the responsibility of test developers to determine those variables affecting test validity and to design tests in such a way that maximizes candidate performance.

These very serious issues have been the motivation in part for an active research agenda at both the applied and theoretical levels examining the nature and composition of the speaking construct in terms of the context in which that speaking takes place and the influence of contextual variables on the realization of that construct. The importance of context has taken center stage with the understanding that, as pointed out by Douglas (2000), "context is not simply a collection of features imposed upon the language learner/user, but rather is constructed by the participants in the communicative event" (p. 43).

The exploratory study described here focuses on the perceptions of examiner behaviors by members of an oral performance test. Particular attention is paid to the candidates and to discovering to what extent the verbal, paralinguistic, and nonverbal discourse behaviors of the examiner are salient to the candidate. This study is grounded in the belief "that the meaning in focused interactions is co-constructed by the participants through situated interpretation of both linguistic and nonverbal behavior as participants manage their talk in interaction" (Jenkins & Parra, 2003, p. 91). The project responds to McNamara's (1996) call over a decade ago:

> The study of language and interactions continues to flourish . . . although it is all too rarely cited by researchers in language testing, and almost not at all by those proposing general theories of performance in language performance tests; this situation must change. . . . We need a research agenda to investigate the significance for measurement of variables that our models may tell us are likely to be important. (pp. 85–86)

We seek to gain insight into one of those variables—that of examiner behavior—and, more specifically, the power relationship that may be altered by examiner behavior as perceived by those participating in or observing the test task. "Power" continues to be an enigmatic construct. For the purposes of this study, we follow a framework offered by French and Raven (1959) that outlines different sources of power. One source of power inherent in oral performance testing is "reward power," which is defined as "the target's perceptions that the agent has the ability to provide him or her with desired tangible or intangible objects" (as cited in Aguinis, Simonsen, & Pierce, 1998, p. 456.) In the testing context, the target is the test candidate and the agent is the examiner, roles that cannot be reversed. The study presented here does not assume that this actual power differential will change. However, we suspect that this inherent reward power might be "masked" through behavior that more appropriately maps onto the role-play task that examiners and candidates are being asked to complete.

The test task, described fully in the Methods section, is an Office Hour Role Play. In this task, the candidate is placed in the role of a person with higher status—that is, a teacher who must give advice to a student, one of the examiners. As pointed out by Bachman and Palmer (1996), "test takers' perceptions of the relevance of the characteristics of a given test task to their TLU [target language use] domains may be different from those of the test developers" (p. 24). Alignment between examiner behavior in a role-play task with the corresponding behavior that is appropriate in the TLU context would contribute to face validity and possibly task authenticity as defined by Bachman and Palmer (1996), who stated that authenticity is "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (p. 23). Thus, an issue to be addressed is to determine the behaviors that candidates are aware of and what their interpretations of those behaviors might be. This study does not endeavor to determine the effect(s) of perceptions on candidate performance or test outcomes. The study is limited to an exploration of which examiner behaviors may be salient to members of the testing event.

BACKGROUND

Van Lier's (1989) seminal article in which he questioned the validity of the oral proficiency interview format provided the impetus for closer investigations into the nature of the interaction in the interview test format relative to the interaction that previous research had indicated occurred in "natural" conversations. In one of the first studies that sought to describe the discourse of oral performance tests, Young and Milanovic (1992) explored the discourse of oral interviews of the Cambridge First Certificate Examination. They quantitatively analyzed the interview discourse for dominance, dominance and contingency, contingency, and goal orientation in terms of the contextual variables of examiner, candidate, theme, task, and gender. Findings indicated that the interview discourse was asymmetrical, that different contextual factors affected participants differently, and that the measures of dominance employed were insufficient in revealing control of the right to speak. This issue of the relative dominance and control of examiner and candidate is a thread that has continued throughout research in this area.

Examiner discourse came under further scrutiny by Ross and Berwick (1992), who have conducted quantitative studies on the discourse of The American Council on Teaching Foreign Languages' Oral Proficiency Interview, focusing on control (e.g., initiation of topics) and accommodation (e.g., modification of speech) by examiners. They were particularly interested in the extent to which examiner accommodation affects an examiner's perception of candidate proficiency. Results indicated that an examiner's use of elements of control mainly serves an administrative function, regardless of candidate proficiency—these elements "move the test along," so to speak, and do not appear to be related to ratings. However, a majority of final ratings could be predicted based on an examiner's degree of accommodation, leading the researchers to conclude that "accommodation provides a potentially useful metric of proficiency" (p. 170).

Subsequent studies by Ross (1992) and Berwick and Ross (1996) further explored examiner control and accommodation in Oral Proficiency Interview tests. Ross argued again that examiner perceptions of candidate proficiency are reflected in the extent of accommodation behavior. Berwick and Ross concluded that Japanese and American examiners exert control in the interview in different ways, with significant differences in the use of accommodation. This comparative study brought to the forefront the topic of variation in examiner discourse behavior.

Lazaraton (1996) employed conversation analysis to examine the transcripts of Cambridge Assessment of Spoken English interviews for the types of examiner support provided to candidates. Lazaraton identified eight types of support that are also present in native–native speaker and native–nonnative speaker conversations (i.e., nontest), which was interpreted as a positive finding in terms of test validity. However, support was not consistent across examiners, which, as Lazaraton pointed out, raises serious reliability issues. Lazaraton questioned the impact of this support on candidate language use and on ratings of that language.

Brown (2003) has made significant contributions to the area of oral performance testing by focusing directly on the influence of variations in examiner discourse behavior on candidate performance. Using discourse analytic methodology, she examined the transcripts of International English Language Testing System interviews of the same candidate with two different examiners. Brown's findings indicate that differences in the ways examiners provide feedback, structure talk, and formulate questions influence candidate performance, which in turn affects raters' perceptions of candidate proficiency as described in postinterview verbal reports of raters. Focusing on both linguistic features (e.g., lexicon, grammar) and extended discourse, the raters' perceptions of the candidate's performance when interviewed by two separate examiners was noticeably different. These results have led Brown to conclude that the "interviewer [is intimately] implicated in the construction of candidate proficiency" (p. 1).

Jenkins and Parra (2003) added a crucial dimension to this body of work through a comprehensive synthesis of research in multiple disciplines that demonstrated the importance of nonverbal behavior in interaction. In an investigation of candidate behaviors that affect examiner perception of candidate proficiency, Jenkins and Parra expanded the features under focus to include nonverbal and paralinguistic variables. More explicitly than in the studies already discussed, Jenkins and Parra considered the unequal power relationship likely to exist in the oral performance test context and the role of candidate nonverbal behavior. They conducted microanalyses of videotaped oral proficiency interview tests of international teaching assistants. The specific features they examined included kinesic features, such as eye contact and body posture; paralinguistic features, such as voice volume, speed, and nonlexical sounds; and verbal and nonverbal turn-taking and listening behavior, such as head nods and back channel cues. The researchers then interviewed candidates and evaluators and analyzed written comments made by evaluators during the tests. Results indicated that those candidates who employed active listening behavior considered appropriate by North American evaluators, such as frequent eye contact, smiling, forward leaning, head nodding, and back channeling, were able to compensate for weaker linguistic proficiency. That is, "while students who were rated linguistically proficient passed regardless of nonverbal competence, linguistically weaker candidates who employed nonverbal strategies were also successful" (p. 103). These strategies, according to Jenkins and Parra, enabled candidates to reduce the unequal power relationship found in oral performance tests by "[framing] the interview as a discussion or conversation among peers . . . [rather than] as an examination" (p. 90).

More recently, studies by Brown (2006), Ducasse (2006), and May (2006) have confirmed that the kinesic and nonverbal behaviors of candidates are particularly salient in examiner perceptions of candidate performance. In a validation study of the revised speaking portion of the International English Language Testing System, Brown (2006) surveyed and interviewed examiners to discover their opinions of the format, the scales, and the performance of examinees. Among the criteria that examiners recommended including in the rating scale were "engagement, demeanor and paralinguistic aspects of language use, tone, and pitch or naturalness of intonation."

May (2006) emphasized how the increased use of the paired format in oral assessments and the concomitant focus on interaction warrants clear definition and operationalization of the construct of interaction. Using data from a paired format speaking test of English for Academic Purposes, May investigated those features of performance that affect raters' perceptions of candidates' ability to interact. Through extensive analyses of scores given for "effective interaction," rater notes, stimulated recall, and paired (raters) discussions, May discovered that raters consider the use of "appropriate and effective body language" as an essential element of "effective interaction."

In the development of an empirically based rating scale for use in paired oral tests of Spanish, Ducasse (2006) utilized Verbal Protocol Analysis to determine raters' definitions of "effective" interaction and to incorporate their judgments of candidate interaction. Ducasse's findings include the observation that "the 12 experienced raters in [the] study . . . unfailingly noticed body language, supportive listening [which includes body language], topic cohesion and fluency." Furthermore, "when asked to operationalise the construct [interaction] by developing a scale, raters . . . chose [body language] and [supportive listening] first and [topic cohesion] and [fluency] second."

The findings just summarized clearly indicate the prominent role of nonverbal behavior in oral performance testing. An area that warrants closer examination is the saliency of examiner nonverbal behavior to participants of a testing event and the extent to which this behavior may play a role in reducing or magnifying the unequal power relationship inherent in oral performance tests. Following Jenkins and Parra (2003), our study draws on Goffman's concept of "footing" as a theoretical framework within which to work. According to Goffman (1981),

> a change in footing implies a change in the alignment we take up to ourselves and the others present as expressed in the way we manage the production or reception of an utterance. A change in footing is another way of talking about a change in our frame of events. (p. 128)

Goffman proposed the concept of "frame space" (p. 230)—the prescribed discourse and identity constants of speakers in a given communicative event. In the context of oral performance assessment, this frame space might appear to be static and unchanging. However, Goffman suggested that through one's communicative behavior, one can change one's footing and thus alter the frame space. In this study, we are interested in examiner verbal and nonverbal discourse behavior that may affect perceptions of the testing event, the participant roles, and the concomitant power relationships.

In summary, this exploratory study is based on the findings of previous research that have identified the discourse behavior of examiners and the nonverbal behavior of candidates as critical variables in the construction of an oral performance test. Our study contributes to this research by focusing on examiner behavior as perceived by participants and observers of one test task to determine (a) which examiner behaviors are perceived by participants and observers and (b) what meaning participants and observers associate with these behaviors.

METHOD

Participants

Participants in the study included four prospective international graduate student instructors (GSIs), one examiner, one examiner/researcher, and one researcher. The first examiner, referred to as "Steve," participated as the undergraduate student in an office hour role-play task. The second examiner observed and rated the office hour role-play task but did not interact with Steve or the candidates in this task. The primary role of the second examiner in this study is that of researcher, taking notes on Steve's behavior during the test and then interviewing the candidates after the test. Thus, this participant will be referred to as Researcher 1.

TABLE 1
Participant Profiles

Participant	*Gender*	*Degree Program*	*L1*	*Time in U.S.*
Harry	M	Chemical Engineering	Chinese	4 years
Zelda	F	Biomedical Engineering	Chinese	3 years
Jay	M	Astronomy	Chinese	9 months
Julie	F	Electrical Engineering/ Computer Science	Chinese	3.5 years

Note. L1 = first language.

The final participant in the study is Researcher 2, who had no involvement in either the initial role-play task or in the interview sessions with the candidates. The roles of each researcher are further explained in the Data Collection and Data Analysis sections. Table 1 shows the gender, degree program, first language (L1), and length of time in the United States for each of the graduate student participants.

This was the first time the candidates had taken the test (Graduate Student Instructor Oral English Test; Briggs, 1987, 2003), and all were approved for teaching.

Materials

The materials for this study were videotaped (and transcribed) interviews with candidates, who were asked to watch a videotape of a portion of their oral performance test and provide feedback. This test, which takes approximately 20 to 30 min to complete, includes four tasks—Background Interview, Lesson Presentation, Office Hour Role Play, and Ten Video Questions–designed to represent the multiple contexts in which GSIs may interact.

The focus of our study is the Office Hour Role Play. In this task, Steve "transforms" himself into an undergraduate student visiting the GSI (the candidate) during office hours. All of the issues presented by Steve have been compiled from information gathered from both undergraduate students and graduate student instructors. In this task, then, candidates are asked to take on the role of teacher, which ostensibly is the role with higher power. This task provides an instance where the candidate assumption of power is contextually appropriate in that it corresponds, presumably, with the TLU domain.

Data Collection

A grounded ethnographic approach provides the means to discover the interpretations of an event by all participants of the event. Recommended by

Douglas (2000) as a useful technique for the development of tests of language for specific purposes, the grounded ethnographic approach makes use of features common to all ethnographic research; that is, it seeks to understand an event by examining both natural occurrences and the descriptions of it provided by coparticipants. Based on Erickson (1979), Tyler (1995) was one of the first to adopt this methodology in the field of applied linguistics, using it to investigate the sources of miscommunication during a tutorial session between an undergraduate student who was a native speaker of English and an international graduate student tutor who was a native speaker of Korean. Two native speakers of English viewed the videotaped interaction and independently noted locations of discomfort in the exchange. The original participants themselves then viewed the tape and, in addition to being asked to comment on the locations commented on by the native speakers of English, were asked to pause the tape at any point when they had felt uncomfortable or confused. This methodology allowed Tyler to uncover participants' variable interpretations of the exchanges during the communicative event, revealing "that differences in perceptions of the negotiability of status and role can play an important part in cross-cultural miscommunication" (p. 145).

In our study, a grounded ethnographic approach was thus employed to investigate the participants' perceptions and interpretations of the verbal, paralinguistic, and nonverbal discourse behaviors of Steve. With this approach, a specific interest in candidate perceptions is informed by the perceptions and interpretations of Steve, a nonparticipating examiner (Researcher 1) and an external observer (Researcher 2), as described next.

Researcher 1 asked candidates to return to examine a portion of their test and to provide feedback on ways to improve the test. With the exception of one candidate who returned 2 months after the test, candidates returned within 1 to 2 weeks after taking the test.

The interview protocol consisted of the candidate and Researcher 1 watching the video of the office hour role-play task once all the way through without stopping. Before beginning the second viewing, candidates were asked for their overall impression of or reaction to the role-play. The candidate was then given the remote control and asked to pause the video at points where he or she would like to comment. After the second viewing, if the topics under investigation had not already been addressed without elicitation, questions asked included, Did you believe Steve is a student? Was the issue he asked about believable? Feedback interviews lasted from 15 to 35 minutes. Unlike Tyler's (1995) implementation of this methodology, Researcher 1 did not bring specific features of the interaction to the attention of the candidates, as the focus was on the candidates' perceptions of Steve and their articulation of the behaviors that led them to those perceptions. In fact, concerted effort was made not to highlight particular behaviors or to lead the candidate.

Researcher 1 also interviewed Steve after the feedback sessions with all the candidates had been completed, viewing all four tests in a single session. The purpose of this session was to determine Steve's perception of his own behavior. The interview protocol for this session was similar to that of the interviews with the candidates. That is, the video of an office hour role-play task was watched once all the way through without stopping. Before beginning the second viewing, Steve was asked to focus on his overall impression of or reaction to his own behavior during the role-play. Steve was then given the remote control and asked to pause the video at points where he would like to comment. This videotaped discussion lasted approximately 1 hr.

All videotaped interviews were transcribed by a research assistant who was not involved in the study. Digital video recordings of tests were converted to digital sound files and then transcribed using the transcription tool Sound Scriber. General Michigan Corpus of Academic Spoken English (http://www.lsa.umich.edu/eli/micase/index.htm) transcription conventions were followed.

Finally, as stated, during the actual tests, Researcher 1 had composed a summary of the salient features of Steve's verbal and nonverbal behaviors.

Data Analysis

Following Jenkins and Parra (2003), the specific features examined included kinesic features, such as eye contact and body posture, which we refer to as nonverbal behavior; paralinguistic features, such as voice volume, speed, and nonlexical sounds; and verbal discourse features such as back channel cues. Researcher 1 and Researcher 2 independently reviewed videotapes and transcripts of the feedback sessions. Researcher 2 also reviewed the notes on examiner behavior that Researcher 1 had taken during the test. Note that the purpose was not to analyze the tests directly; rather, this study focuses on an analysis of perceptions of examiner behavior. Included are the researchers' summaries of candidates' commentary on examiner behavior, comments of the nonparticipating examiner, and observations of the examiner himself.

RESULTS AND DISCUSSION

Our discussion of which examiner behaviors are noticed and what meanings are associated with those behaviors begins with a summary of each candidate's observations of Steve's verbal and nonverbal behaviors. Then, we summarize Steve's report of his own behavior; finally, we outline the researchers' convergent interpretations of Steve's verbal and nonverbal behavior. We end by synthesizing multiple perceptions of the office hour role-play event and then discuss the implications of these findings on oral performance test development and examiner training.

Candidate Observations

In our discussion of the candidates' commentary on the role-play task, we begin with the initial question posed to all participants—that is, "What is your overall impression or reaction to the role play?" Recall that one of our primary concerns was to determine those behaviors or features that individuals actually notice, so it is informative to note what participants chose to focus on in response to such a general question. In fact, responses to this initial question ultimately formed three main categories of responses that became apparent when analyzing all candidate observations: First, candidates commented on their own language performance; second, candidates noted the realism of the topic content of the role-play; and third, candidates remarked on the realism of Steve's verbal, paralinguistic, and nonverbal discourse behaviors.

Candidates' self-evaluations.

Beginning with the first category, two candidates initially focused on their own performance rather than that of the examiner.

Example 1: Harry. Feedback session. [After first viewing]
[Situation: Steve asks about the content of an upcoming midterm exam.]
[5:26]
1. Rsch: So what do you think? Do you have any overall impression?
2. Harry: Yeah, I think I spoke too fast.
3. Rsch: Okay.
4. Harry: Yeah.
5. Rsch: Anything else?
6. Harry: Uh, and some, sometimes I, I feel like I do this, [puts hand to face] mm, (I was xx)
7. Rsch: Oh.
8. Harry: I have to appear more confident about the ques- or more, (fact) responsive. [5:48]

As can be seen in Example 1, Harry's observations highlight both his linguistic self-awareness and his sensitivity to nonverbal cues in communication. In line 2, he first remarks on a specific pronunciation issue (speaking too quickly), and then in lines 6 and 8, he comments unfavorably on his nonverbal behavior—putting his hand to his head—stating that this may portray a lack of confidence.

Example 2. Julie. Feedback session.
[Situation: Steve complains about group members not doing any of the work on a final project.]
[3:41]
1. Rsch: Okay, let me rewind it now. So what are your thoughts about it?

2. Julie: Um, I don't know if, uh, like, if my answer is understandable, so . . . like, I, I mean uh, because uh the, uh, I was listening to this again, (for) some of my pronunciation is not that clear, [Rsch: mm] so I don't know if it's understandable, I mean, uh, so [4:14]

Julie's linguistic self-awareness is evidenced in Example 2, line 2. Her first reaction to viewing the role play is to find fault with her own pronunciation.

Candidates' evaluation of topic content: Realism. In response to the initial question ("What's your overall impression?"), Zelda provided the second category of response type to emerge—a focus on the topic content of the role-play.

Example 3. Zelda. Feedback session.
[Situation: Steve says that he saw some students cheating on the final exam.]
[2:00]

1. Zelda: S- so, what –
2. Rsch: What was your impression of that? What's your reaction, what, what did you
3. Zelda: yeah, seems I mean, seems that I just, like, a shock when I hear the question
4. Rsch: uhuh
5. Zelda: and didn't have – seems to me the time to, to think about that. [2:21]

As can be seen in Example 3, lines 3 and 5, Zelda states that her reaction was one of shock at the question and she did not feel that she had time to formulate a response.

Example 4: Jay. Feedback session. [After first viewing.]
[Situation: Steve did the wrong problem set for homework.]
[4:13]

1. Rsch: So what's, wh- what's your overall impression of the office hour?
2. Jay: Uh-
3. Rsch: I mean, what's your reaction to it?
4. Jay: I think it's, it's kind of real. [Rsch: mhm] I mean, it's, um, it's l- it's like the, you know, the, the real, real, uh, real student, almost like. [Rsch: oh] so [Rsch: uhuh] and, and, his role is quite like the student's and I, uh, I mean –
5. Rsch: I- so you, you believed that he was a student.
6. Jay: Yeah, right.
7. Rsch: Wh- w- ho- why, do you think?
8. Jay: Uh, from his, uh, it's like, l- uh, the words from, uh, I mean . . . the questions [Rsch: mhm] he asked i- is, is, I mean, was very,

[Rsch: uhuh] uh, every student will, will, will have, m- may have, I mean, [Rsch: uhuh] and uh, and . . . and th- that's it. [5:14]

Jay focuses on the topic as well. He is the first candidate to mention the realism of the role play. In Example 4, line 4, Jay thought the test was "kind of real." He appears to link the realism of the role play to Steve with his observation that Steve was "like the real student, almost like." He then proceeds, in line 8, to attribute this realism to the content of Steve's questions.

At some point in each interview, the candidates were explicitly asked if they believed Steve was a student. All four candidates responded in the affirmative. Although this question was designed to elicit observations about Steve's behavior, again, the density of comments associating believability with the content of the questions was noticeable.

Example 5: Julie. Feedback session. [During second viewing.]
[Situation: Steve complains about group members not doing any of the work on a final project]
[6:07]
1. Rsch: Was he a be- a believable student?
2. Julie: Uh, yes, I think so.
3. Rsch: Why? What, what did he do or say that made you think, yeah, he could, he could be a student?
4. Julie: Um, because, because I know students usually, they become, they complain about these kind of problem a lot. [Rsch: yeah, yeah] Uhuh. Actually, I'm not sure if my answer is uh uh is a right answer or not, because I, I don't know how to deal with this situation so I just come up with some uh . . . [6:39]

The belief that it was Steve's topic or question that lent credibility to Steve as an undergraduate student and to the role play was also voiced by Julie. For example, in Example 5, line 4, Julie comments that students usually complain about these kinds of problems.

Example 6: Zelda. Feedback session. [After first viewing.]
[Situation: Steve says that he saw some students cheating on the final exam.]
[2:45]
1. Zelda: Yeah, I think, I think this ki- kind of, um, very like the real, real one
2. Rsch: Why? [Zelda: uh?] What made it, what made it real?
3. Zelda: So, because the, I think the, the, I forgot his –
4. Rsch: Steve.
5. Zelda: Steve, yeah. I think, I think he, his way to, to, to talk, [Rsch: mhm] act as a, a student really have a question [Rsch: mhm] for that, but . . . but

seems I, a little w-, just try to get a answer [Rsch: right] quickly! [Rsch: right, yeah] That's a little . . . how say that, um . . . [3:20]

[Discussion of the meaning of "impression."]
[4:00]

6. Rsch: Yeah? Wh- why? [Zelda: yeah. huh?] What did he say, or what did he do, that, that made you believe, yeah, he's a student?
7. Zelda: Because the question is kind of very, it's a, student's question, [Rsch: uhuh] because I, I was GSI before, so some peo- [Rsch: oh, okay] student, they they, they ask such a question. [Rsch: yeah, okay] Yeah, so I think the question is very real. [4:26]

A comment about the content of Steve's questions also appeared later in the feedback session with Zelda, as shown in Example 6, line 7, when she says that "the question is kind of very, it's a, student's question . . . so I think the question is very real."

Finally, as can be seen in Example 7, lines 6, 8 and 10, Harry points out that although his students might ask more detailed questions, Steve's questions are very similar. That is, content appropriateness, the second category of responses, led all candidates to find Steve believable as an undergraduate student.

Candidates' evaluation of steve's behavior: Realism. Candidates also attributed Steve's believability to his speech style and body movements, which ultimately formed the third and final category of responses. Some candidates associated these verbal and nonverbal behaviors with a certain attitude and showed some frustration in finding the right words to articulate the "feeling" that they received from Steve.

Example 7: Harry. Feedback session. [After first viewing.]
[Situation: Steve asks about content of upcoming midterm exam.]
[6:49]

1. Rsch: Okay. So you felt like it was a real office hour, or…?
2. Harry: Yeah, real office hour.
3. Rsch: Did, did Steve seem like a real student?
4. Harry: Yeah, exactly.
5. Rsch: What, what did he do or say that made him seem like a, real student?
6. Harry: Hmm . . . I think, yeah, uh . . . yeah, he looks for, he looks like a student as, uh [Rsch: yeah] the speed and the tune he's asking is quite like nor- like some of my (like, real) students, [Rsch: okay] uh, some, some ways.
7. Rsch: Okay. So the questions he was actually asking, students ask you that.
8. Harry: Uh, the questions he ask, he ask is very general, but [Rsch: mhm] uh my students ask more detailed but uh, [Rsch: ah, okay] still it's kind of, you know, it, it's (xxx)

9. Rsch: Right, right.
10. Harry: Students similar, similar.
11. Rsch: Okay. [7:44]

Continuing with Harry, in Example 7, line 6, Harry first states that the “speed and tune” of Steve's speech is like some of his students. We are uncertain of exactly what Harry means by “speed and tune”; however, we may speculate that “speed” may indicate rhythm of speech, pace, or linking and elision, and that “tune” may indicate overall voice pitch or specific intonation patterns. More significantly, it appears that Harry has called on paralinguistic features of Steve’s language in an attempt to explain why Steve seemed like a real student to him.

Example 8: Harry. Feedback session. [During second viewing.]
[Situation: Steve asks about content of upcoming midterm exam.]
[10:37]

1. Rsch: Right here. You mentioned that you think, that you thought Steve looked like a student. Is it the way he’s dressed?
2. Harry: Um . . . yeah, the way he’s dressed and, uh yeah, the way he’s dressed, he’s dressed, and . . . and, you know, he looks kind of humble. I’m not sure exact word describe this.
3. Rsch: Humble?
4. Harry: Yeah, is that true? Is that, h- humble a good word? So, I don’t use this word very often. [11:09]

[Discussion of the meaning of “humble.”]
[11:32]

5. Harry: Because – for some reason, I thought, I don’t feel like it’s a exam, because uh, the way he’s playing is like a real, he’s really asking my question, [Rsch: oh?] he’s asking my questions, and he really wants to know the answer, [Rsch: ahh] so I feel that –
6. Rsch: What does he do? that makes you feel that?
7. Harry: Mmm . . . I think, the eye contact?
8. Rsch: Eye contact?
9. Harry: Yeah.
10. Rsch: Okay.
11. Harry: Well, it’s hard to say, it’s a very, you know, [Rsch: yeah] it’s, it’s a subtle feel- feeling, [Rsch: right] but, but I feel comfortable with that. . . . Maybe the (arms) like this *(forearms on legs with hand on face)* and (xx) eye, eye contact
12. Rsch: Okay, so he was leaning forward, and the eye contact.
13. Harry: Yeah, seemed like, you know, he’s confused, (but some xx) and he really want, want my help. [12:39]

Later in the feedback session with Harry, in Example 8, the researcher returns to the subject of Steve being perceived as an actual student, and Harry states in line 2 that Steve looks "kind of humble," but he is uncertain of the exact word to describe the impression he has of Steve. Harry states that the test did not "feel like . . . an exam" in line 5 and attempts to explain why he felt that Steve was an actual student. Harry introduces the topic of Steve's eye contact in line 7. In line 11, he hesitates, saying that it is "hard to say" what it is about Steve and that it is a "subtle feeling but [he] feel[s] comfortable with that." Harry concludes in lines 11 and 13 by showing that Steve's eye contact and his positioning of his forearms on his legs with a hand on his face make Steve appear somewhat "confused" and sincere about wanting help from Harry. Recall that Harry (Example 1) expressed awareness of his own linguistic performance and of the impact of his nonverbal behavior; in the excerpt just discussed, we clearly see Harry's awareness of these behaviors in others.

Example 9: Zelda. Feedback session. [During second and third viewing.]
[Situation: Steve says that he saw some students cheating on the final exam.]
[6:09]

1. Rsch: Do you think he looks like a student?
2. Zelda: Yeah, yeah, yeah, his, he ways to speak, yeah.
3. Rsch: What do you mean, his way to speak?
4. Zelda: Um, like uh, when he come in, and then, um . . . (xx) let me think . . .
5. Rsch: Do you wanna see it again?
6. Zelda: Yeah. Just the first part.
7. Zelda: So, to here (*pauses tape)*, I think he, like a student, from the, coming and into here question, I feel he's a student. Like, his way to speak.
8. Rsch: Like what? (0:05)
9. Zelda: I don't know, I, I don't, can't say that, just, [Rsch: yeah?] just a feeling, [Rsch: okay] feel, yeah, feel he's, the way acting like a student. (0:05) um (0:06)
10. Rsch: Here's the play but- (0:11)
11. Zelda: I, I feel like uh, when friends talk to each other, they like uh, just say whatever. [Rsch: mhm] But this, like a student to GSI, they stay still with some, um, like you are our teacher or like [Rsch: mhm] such (feeling), [Rsch: mhm] like uh with a little respect, [Rsch: mhm] those (feeling).
12. Rsch: Is he showing respect?
13. Zelda: . . . mmm, yeah.
14. Rsch: Okay.
15. Zelda: (I feel definitely like that.)
16. Rsch: How? How does he do that? (0:03)
17. Zelda: Um. (0:05) . . . *(restarts tape)* [8:39] Ah, ah, yeah, Steve looks like a student also. [9:57]

18. Rsch: Yeah?
19. Zelda: Yeah, he's, he, because, like uh he walks in, [Rsch: uhuh] yeah, when he, um, it's like student walking, and student chatting with instructor, not a friend chatting, [Rsch: okay] or not a professor chatting. [Rsch: okay] So, so, s- yeah, I think student chatting with instructor, they have different, different way, not - not like friends, [Rsch: uhuh] they just uh directly say something, he will um have something to say, but still a little hesitate, and then s- but – (the)-
20. Rsch: So he hesitates a little?
21. Zelda: Yeah.
22. Rsch: Okay. [Zelda: (xx)] And the way he's dressed?
23. Zelda: Yeah, yeah, the way dressed, and sit there, he's uh the way like a student, he's doing like the gesture, yeah.
24. Rsch: The way his, his gesture?
25. Zelda: Yeah, like a student. *(Puts her hand to her face.)* The professor won't do that. [Rsch: oh] The examiner won't do that.
26. Rsch: Okay. So what else in his, you mentioned his gestures. So you, you say he has his hand here. [11:09] (*restarts tape)* [12:20]
27. Zelda: Mm. (0:08) I, I, I'm bad in expressing what I'm thinking. (0:10) Mm. His t- Yeah, his tone also. His tone. [Rsch: oh, okay] Yeah. [Rsch: okay] Zelda: (Like) a student's tone. [12:49]

Returning now to Zelda, in Example 9, when asked if she thought Steve *looks* like a student, in line 2, after saying yes, she immediately returns to his speech behavior, stating that it is "his . . . way to speak," which is reminiscent of Harry's explanation for Steve's believability as a student. When pressed to explain, Zelda says in line 4 "let me think," watches part of the tape again and then concludes in line 7 that she "feel[s] he's a student. Like his way to speak." The researcher continues to try to get a description and in line 9, Zelda says that she cannot say; she just feels that the way he is acting is like a student. After several pauses, in line 11, Zelda comments on Steve's lack of movement (". . . they stay still . . .") and then in lines 11, 13, and 15, says that she feels that Steve is showing her respect. In line 16, the researcher then asks her "how does he do that?" After mentioning his appearance, again, in line 17, she moves the discussion back to Steve's speech behavior in line 19 stating that the way he is "chatting" is the way a student would with an instructor, "not like friends." In lines 19 and 21, she also characterizes his speech as being hesitant. Even when the researcher returns to his appearance in line 22, Zelda quickly agrees and then in lines 23 and 25 focuses on his gestures, specifically his hand being placed on his face in lines 23 and 25. After stating in line 27 that she's having difficulty expressing herself, she comments on his "tone" being "like a student's tone."

In sum, in attempting to articulate their impression of the Office Hour Role Play, beyond (negatively) evaluating their own verbal and nonverbal performance, candidates commented on the realism of the task in terms of the task content and in terms of examiner behavior. They attempted to describe those features of test task and of examiner behavior that made the event realistic for them. In doing so, they supported the proposal made by Bachman and Palmer (1996), who underscore the importance of authenticity and state that it is not a simple one-to-one correspondence between test task and real-life task but rather includes "characteristics of test takers, the TLU domain, and the test task" (p. 39).

Examiner 1 Self-Observations

We turn now to Steve's observations and explanations of his own behavior. At the very beginning of the feedback session, Steve paused the video and stated, "Lots of 'likes' and 'ums' and 'stuff.' I intentionally do that to try to sound like an undergraduate. Just to try to make it more authentic. I think I'm reasonably good about putting them in very natural positions." That is, the evaluator is consciously adopting what he considers to be the speech style of a typical undergraduate. Later in the feedback session, he stated that he believed, however, that the candidates in any test "[are] controlling the formality of the register" and that if they treat the role play as just an interview portion of the test, then he finds himself becoming more formal in speech register by using, for example, fewer contractions and fillers.

Steve also initially commented on his body posture:

> Another thing that I notice watching it through the first time is how I pretty much am sitting in the same position with my hand on my face almost the whole time and I think that I usually do that and that's more comfortable for me. I'm not sure if that's the best authen– I mean I don't think it's specifically *in*authentic but that's more just because that's how I often am. I mean you see me like that in meetings and stuff all the time.

Steve's observation is interesting in the light of the candidates' perceptions of his body language. It is interesting that he does not consciously adopt a specific pose for the role play, and yet it was one of the initial features that he noticed about himself, and one of the features specifically described by a candidate.

Researchers' Observations and Interpretations[1]

Although Steve and the candidates did not consistently and explicitly state their observations in terms of power or dominance, our own are reported and

[1] Recall that Researcher 1 noted examiner behavior during the actual test. Researcher 2 reviewed these notes. Both researchers independently interpreted these notes and the candidate and examiner commentary from the feedback sessions.

interpreted in terms of the power relationship that may be altered as a result of examiner behavior. First, Steve rests his elbows on his legs and places a hand on his cheek; as a result, his head is physically lower than that of the candidates. This is a prototypical posture of non-aggression and almost submissiveness.

Example 10: Steve and Zelda. Role-play.
[2:09]
1. Zelda: Come in, please.
2. Steve: Hey [ ], how's it going?
3. Zelda: Good, how are you?
4. Steve: Good. *Um, I wanted to* ask you about the final last week.
5. Zelda: Mhm.
6. Steve: *Um, I think* some of the people in our section were cheating on it.
7. Zelda: Oh really?
8. Steve: Yeah.
9. Zelda: So how, how are you doing on the exam?
10. Steve: *Um, I th- I think* I did all right. *Um, I mean,* I hope, *you know,* the curve doesn't get screwed up *or anything*. [2:35]

Second, features that were not initially the focus of investigation became apparent. As shown in the italicized text throughout Example 10, Steve's speech is frequently marked with false starts, fillers, and hedges, forms that have been categorized as serving, for example, to mitigate the assertiveness of a statement and, therefore, as markers of powerless speech styles (e.g., Coates, 1996; Culberson, 2002; Hirschman, 1994; Holmes, 1996; van Baalen, 2001). Steve's speech does not typically contain these features as he notes himself during the feedback session: "Lots of 'likes' and 'ums' and 'stuff.' I intentionally do that to try to sound like an undergraduate. Just to try to make it more authentic." In addition, it is notable that early back channels (Yngve, 1970; as cited in Schiffrin, 2001) are not observable in Steve's speech, consistent with the features of a speech pattern of an individual of lower status. Prosodic features of Steve's speech were also noticeable to the nonparticipating evaluator. Steve's speech was frequently marked with rising intonation at the end of statements, which may indicate uncertainty or a desire for confirmation when introducing a topic or acknowledging a candidate response, grounded in a low-pitch flat intonation pattern common among male undergraduate students. The deference communicated by the rising intonation segments stands out more in this speech pattern than in one with frequent pitch modulation for emphasis or other communicative functions.

CONCLUSION

This exploratory study sought to gain insights into the verbal, paralinguistic, and nonverbal behaviors of an examiner as perceived by members of the testing event. Two questions were posed: (a) Which examiner behaviors do participants and observers report as salient, and (b) what, if any, meaning do they associate with these behaviors? To lay the groundwork for future research that would examine which, if any, examiner behaviors affect candidate performance, particular attention was paid to the observations of the testing candidates.

The grounded ethnographic approach adopted for this study allowed us to examine the convergence and divergence of the commentary and interpretations by all participants in the event. The main area of divergent observations revolved around task content. Observations of examiner behavior converged, and, given our interest in the candidate, conclusions are therefore organized according to the candidates' commentary, which can be grouped into three main categories. The implications of their observations on assessment are included with the summary of each category.

First, of significance is the fact that all the candidates initially commented on the content of Steve's issue or question. Topic content resulted in Steve seeming more like an undergraduate student and in the task seeming more real for the candidates. In contrast, neither Steve nor the nonparticipating examiner commented on task content in their feedback on the testing event. Although content validity is certainly at the fore of the test development process, this explicit commentary on task content from candidates underscores the imperative to return continually to the target language use domain and its participants as the source of test task content.

Second, all participants commented to varying degrees on Steve's nonverbal behavior. Candidates provided relatively explicit comments on Steve's body position and/or lack of movement and his eye contact, indicating that this lent credibility to his being an undergraduate student. Even though not all candidates could articulate the exact attitude that they "felt" from Steve or associate a specific posture or behavior with that attitude, these nonverbal behaviors were salient to the candidates, who connected them with an impression that Steve did seem like an undergraduate student. This awareness of nonverbal behavior supports previous studies that have shown other aspects of examiner behavior as crucial variables in the co-construction of the testing context. Our study reveals the meaningfulness of examiner nonverbal behavior from the perspective of the candidate in that context.

As has been noted by other researchers working within this same vein, this finding has direct implications on the training of examiners. Although numerous studies have found that an examiner's perception of candidate performance is influenced by the nonverbal behavior of candidates, examiners must be made equally aware of the influence of *their* nonverbal behavior on the testing event and ultimately the possible effect on candidate performance. Training programs for examiners must include a component that raises examiner attentiveness to and appreciation for the sociocultural aspects of oral performance tests. For large-scale tests at least, it would be practical to provide examiner trainees with video samples of actual tests that include both TLU consistent and TLU inconsistent examiner behavior for the trainees to analyze. For small-scale tests, the training program can potentially be more elaborate. For example, examiner trainees can participate in role-play tasks based on TLU situations that would be observed by trainers and other trainees.

Finally, two of the four candidates, Steve, and the nonparticipating examiner commented at some point in their reflections on Steve’s speech behavior. The candidates agreed that the speech pattern adopted by Steve makes him more plausible as an undergraduate student. Recall, however, that Steve believes candidates themselves control whether he maintains this register, a register that we have suggested is a style marking less power. He finds himself switching to what he calls more of an interview style in response to candidates who adopt this (formal) framing behavior. As test developers and examiners, then, we have a decision to make. As we know, discourse is co-constructed; if a candidate, for whatever reason, does not step into the pretend world of the role play and frames the event as a test, do we force the issue—and in a certain sense, exert power—and continue using the speech patterns associated with a different speech event? After all, we are asking candidates to suspend reality and to participate in a speech event—a role play—that may be difficult for them in any language. Given the very limited feedback from candidates in our study, we *can* say that the informal or powerless speech style adopted by Steve does not go unnoticed. Candidates reported feeling comfortable and respected. We suggest that these experiences are indicative of a context in which the power relationship has temporarily shifted and the relatively powerful position of the examiner is not at the forefront.

Of course, there do exist TLU domains in which the relative power of TLU participants and the unequal power relationship present in testing contexts (e.g., graduate student instructors must frequently interact with faculty) correspond. Indeed, the framing behavior of an examiner should align with the TLU context. As with the aforementioned nonverbal behaviors, test specifications must clearly describe the speech behaviors that are appropriate for a given TLU speech event.

With those generalizations made, we need to point out several limitations of the study described here, the first of which is that any generalizations are tenuous at best. Given the limited number of participants, the difficulty candidates had in finding the words to articulate their impressions of Steve, and the fact that not all candidates consistently commented on the same behaviors, one needs to be cautious in drawing any conclusions. We need to ask to what extent the observations reported here are unique to the few individuals participating in the study. For example, all candidates are Chinese; it may be that there are fundamental differences in the manifestation and interpretation of discourse behaviors across cultures. In addition, with this methodological approach, researcher expectancy is a significant concern; therefore, one needs to be cautious in categorizing and assigning particular meanings to the reflections and observations of those interviewed.

Nevertheless, we believe that this exploratory study demonstrates a need for larger scale investigations into the effect of examiner nonverbal behavior in oral performance testing. With an understanding of the range of examiner behaviors that are meaningful, particularly to candidates, future research can then investigate the effects of those perceptions on candidate performance. Subsequent studies should focus on critical variables not addressed in this study. For example, the effect, if any, of the gender of the examiner on candidates' perceptions of evaluator behavior warrants attention. In addition, one would want to observe an examiner who does not adopt the verbal and nonverbal behaviors highlighted here and then conduct similar interviews with the candidates. Such a comparative study would allow us to examine the relationship between different nonverbal behaviors and the distinct perceptions of those behaviors. Furthermore, in terms of methodology, in addition to the rather indirect questioning applied in our study, asking candidates explicit, direct questions regarding the various power relationships in a testing context may be very instructive. Finally, we would want to determine the extent to which types of power relationships that are unique to oral performance testing (and absent in corresponding TLU contexts) might interfere with or be irrelevant to valid assessment of candidate performance. The ultimate goal is to determine the effect of the power relationship, established in the testing context through the verbal, paralinguistic, and nonverbal behavior of examiners, on candidate performance.

ACKNOWLEDGMENTS

A version of this article was presented at the 14th World Congress of Applied Linguistics, Madison, Wisconsin, July 2005, by the first author.

We thank Carson Maynard and Sunny Park for the hours spent transcribing; Aaron Ohlrogge for his assistance with data collection; and Sarah Briggs, Barbara Dobson, Nick Ellis, Theresa Rohlck, and John Swales for insightful discussions on the study. We also thank four anonymous reviewers for their feedback. Any shortcomings remain our own.

REFERENCES

Aguinis, H., Simonsen, M. M., & Pierce, C. A. (1998). Effects of nonverbal behavior on perceptions and power bases. *The Journal of Social Psychology, 138,* 455–469.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Berwick, R., & Ross, S. (1996). Cross-cultural pragmatics in oral proficiency interview strategies. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: selected papers from the 15th Language Testing Research Colloquium* (pp. 34–54). Cambridge, UK: Cambridge University Press.

Briggs, S. L. (1987/2003). *Graduate Student Instructor Oral English Test.* Ann Arbor: English Language Institute, University of Michigan.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1–25.

Brown, A. (2006). *Validating the revised IELTS speaking test.* Poster presented at the Language Testing Research Colloquium, July 2006, University of Melbourne, Australia.

Coates, J. (1996). *Women talk, conversation between women friends.* Oxford, UK: Blackwell.

Culberson, R. (2002). *Perceptions of assertiveness as a function of tag questions*. York: York College of Pennsylvania. Retrieved from http://www.ycp.edu/besc/Journal2002/paper%201.htm. Accessed May 16, 2005.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge, UK: Cambridge University Press.

Ducasse, A. M. (2006). *An empirically-based rating scale for 'interaction' in a paired oral test.* Paper presented at the Language Testing Research Colloquium, July 2006, University of Melbourne, Australia.

Erickson, F. (1979). Talking down: Some cultural sources of miscommunication in interracial interviews. In A. Wolfgang (Ed.), *Nonverbal behavior: Applications and cultural implications* (pp. 99–126). New York: Academic.

French, J. R. P., & Raven, B. H. (1959). The bases of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 150–167). Ann Arbor, MI: Institute for Social Research.

Goffman, E. (1981). *Forms of talk.* Philadelphia: University of Pennsylvania Press.

Hirschman, L. (1994). Female–male difference in conversational interaction. *Language in Society, 23,* 427–442.

Holmes, J. (1996). Hedging your bets and sitting on the fence: Some evidence for "hedges" as support structures. *Te Reo, 27,* 47–62.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal, 87,* 90–107.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13,* 151–172.

May, L. (2006, July). *"Effective interaction" in a paired candidate EAP speaking test.* Paper presented at the Language Testing Research Colloquium July 2006, University of Melbourne, Australia.

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing, 9,* 173–186.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14,* 159–176.

Schiffrin, D. (2001). Discourse markers: Language meaning, and context. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *Handbook of discourse analysis* (pp. 54–75). Oxford, UK: Blackwell.

Tyler, A. (1995). The coconstruction of cross-cultural miscommunication: Conflicts in perception, negotiation, and enactment of participant role and status. *Studies in Second Language Acquisition, 17,* 129–152.

van Baalen, I. (2001). Male and female language: Growing together? *Historical Sociolinguistics and Sociohistorical Linguistics*. Retrieved June 2, 2005, from http://www.let.leidenuniv.nl/hsl shl/van%20Baalen.htm van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23,* 489–508.

Yngve, V. (1970). On getting a word in edgewise. *Papers from the 6th Regional Meeting, Chicago Linguistic Society*, pp. 567–578.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14,* 403–424.