Article

LANGUAGE
TESTING

Language Testing
2018, Vol. 35(3) 427–445
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532218772325
journals.sagepub.com/home/ltj
SAGE

# Interactional competence: Genie out of the bottle

**India Plough**
Michigan State University, USA

**Jayanti Banerjee**
Paragon Testing Enterprises, Canada

**Noriko Iwashita**
University of Queensland, Australia

## Abstract

The papers in this special issue provide support for continued scrutiny of interactional competence (IC) as an important component of the speaking construct. The contributions underscore the complex nature of IC and remind us of the multiple factors that affect any construct definition. At the same time, each study offers insights into those factors through their explorations of IC. In this final paper, we first briefly review key findings from the papers that confirm what is already known about IC and that provide new information to our understanding of the construct of IC. After summarizing points of convergence and of divergence, we turn to a discussion of areas that require additional targeted attention and offer four generalizations as starting points for research. In the final section, we take a critical look at the challenges associated with including IC in the speaking construct and the implications of the studies in this special issue for the relationship between IC and proficiency.

## Keywords

Assessment, computer-based testing, context, conversation analysis, interactional competence, nonverbal behavior, speaking construct, task

**Corresponding author:**
India Plough, Michigan State University, Snyder Hall-C210, 302 Bogue Street, East Lansing, MI 48825-1106, USA.
Email: ploughi@msu.edu

## Themes emerging from this special issue

In their analyses of repair sequences, Roever and Kasper highlight the extent of examiner influence. They show that test takers may use language that appears dysfluent as an interactional strategy. However, rather than being evidence of a gap in linguistic knowledge, these dysfluencies may be a response to unexpected behavior by their interlocutor (in this case, the examiner). By conceptualizing and investigating the construct of IC from a conversation analytic perspective, Roever and Kasper achieve an in-depth understanding of how interactions are co-constructed and the significant implications of examiner discourse on that co-construction. A broader consequence of this detailed analysis is Roever and Kasper's (p. 344) conclusion that "language knowledge and IC are separate kinds of competence," confirming a multicomponent model of language proficiency.

Ross reminds us of the possible role of the L1, in this case the ability to employ appropriate socio-pragmatic functions, in the IC of L2 speakers. In his examination of listener responses in the Oral Proficiency Interviews of an L2 Japanese candidate (an L1 English speaker) and an L2 English candidate (an L1 Japanese speaker), Ross notes an oversupply of backchannels in English by the L2 English candidate. In contrast, there was a noticeable lack of backchannels produced by the L2 Japanese candidate. While cautioning against generalizations based on two case studies, Ross suggests that the L2 English candidate may have oversupplied backchannels because of a transfer from Japanese, the candidate's L1, in which backchannels serve multiple functions in discourse. It is interesting to note that Ross does not attribute the L2 Japanese candidate's lack of backchannels to a transfer from English, the candidate's L1. Rather, he suggests that there may have been an examiner-effect. Echoing Roever and Kasper, Ross also highlights the impact of test format and task design on candidates' opportunities to demonstrate the range of their abilities. We return to this critical issue later in the next section. With Ross's micro-analytic study we are able to clearly see the complexity of variables (e.g., L1, L2, examiner behavior, task design) affecting IC, and thus the difficulty in identifying distinct features of the construct that can be assessed.

By focusing on one interactional feature, that of responding, Lam establishes a methodological foundation for future research into IC. Through his 'thick' description of contingent responses (formulating previous speakers' contributions; accounting for (dis) agreement with previous speakers; and extending previous speakers' ideas), Lam clearly distinguishes them from formulaic backchannels (an issue raised by Roever & Kasper and Ross). Our understanding of this interactional feature is expanded, and we see the IC construct beginning to take shape. In addition, we gain insight into the construct of interactive listening. That is, contingent responses are more accurate indicators of listener comprehension and engagement than are formulaic backchannels. As noted by Roever and Kasper, speakers can only move the conversation forward if they have understood what their interlocutor has said.

Through her examination of radiotelephony communication, Kim reintroduces us to the context-dependent nature of the IC construct, particularly in highly specialized domains. Kim notes that professional knowledge (in this case, the lack thereof), which is not accounted for in the descriptors and scale of the International Civil Aviation Organization, was a significant impediment to an effective interaction between a Korean air traffic controller and a Russian pilot. Although all contributions in the special issue emphasize the importance of contextual variables (e.g., location of interaction, examiner discourse style), Kim takes this point further. She argues that highly specific contexts such as air traffic radiotelephony communication are markedly and importantly different from general purposes contexts. Therefore, the definition of IC for air traffic radiotelephony must incorporate the features of this professional context.

Taken together, the papers in this special issue have revealed new layers in the IC construct. We next examine the points of convergence and divergence among the studies with respect to implications for its assessment. Roever and Kapser and Lam argue that preliminaries (e.g., presequences to requests and refusals [Roever & Kasper]) and listener responses (Lam) are indeed identifiable features of IC that can potentially be used as indicators of proficiency. While Roever and Kapser provide evidence that production of preliminaries can be systematically categorized, Lam speculates that the importance of the "relative quality" of contingent responses is such that increased use of the feature does not necessarily mean higher proficiency. That is, contingent responses can vary in complexity from simple formulaic responses to those that transform, conceptually and linguistically, the previous speaker's contribution. In contrast, Ross focuses on 'traditional' backchannels as potential indicators of IC and concludes that this feature cannot be consistently distinguished from other linguistic criteria such as fluency, accuracy, and coherence. Roever and Kasper assert a multicomponent model of speaking proficiency, and, thus remark that interactionally competent speakers who have limited linguistic ability may use these features. Kim stands apart in terms of the focus on professional knowledge. Nonetheless, the inherent effects of context (which includes content knowledge and an understanding of appropriate norms and conventions of a particular context) are referenced either implicitly or explicitly by all contributors.

## Areas to address

Although this special issue consolidates what we have learned about IC to date and takes us forward from earlier work (cf. Berry, 2007; Brooks, 2009; Davis, 2009; Ducasse & Brown, 2009; Galaczi, 2008; Nakatsuhara, 2006, 2011; May, 2009, 2011), the discussion is not at an end. A number of areas remain unexplored or incompletely explored, including:

1. the relationship between task and the evidence of IC that is elicited;
2. the role of nonverbal behavior in IC;
3. the effect of assistive technology (such as video-conferencing tools) upon the operationalization of IC; and
4. the extent to which computer-delivered tests can support inferences about IC.

### *Interactional competence and task design*

The data for each paper in this special issue have been elicited in a variety of ways. Kim's data are taken from recorded radiotelephony discourse between an airline pilot and an air traffic controller. Lam's data have been taken from two group discussion tasks, each with a different communicative purpose, that is, description versus decision making. Roever and Kasper investigate data gathered both from classroom role-play tasks and from oral proficiency interviews (OPIs). Ross analyzes data gathered from OPIs (both face-to-face and telephonic). OPIs, though they do not have to be face-to-face, are typically unscripted. OPIs offer the test taker an opportunity to demonstrate their ability to interact with a single interlocutor and to participate in a situational role play. However, as discussed by van Lier (1989) and Johnson (2001), turn taking and topic control are dominated by the examiner in OPIs. Brown's (2003) work on the International English Language Testing System's (IELTS™) speaking test suggests that the manner in which a language proficiency interview (LPI) proceeds, and the opportunities available for test takers to demonstrate their IC, also depends very much on the examiner. Consequently, during an OPI or an LPI, the scope of the interaction and the subsequent score generalization can be (and often is) limited.

This limitation suggests that, if IC is to be assessed, the test design should better ensure that test takers are able to demonstrate their IC. A task-based approach, with its focus on eliciting language "within a well-defined communicative context (and audience), for a clear purpose, toward a valued goal or outcome" (Norris, 2016, p. 232) seems ideally suited to ensuring opportunities to elicit and evaluate IC. The question, however, is which approaches to test format (including task design) are likely to be the most successful and what task features should be incorporated. One of the tasks used by Lam (this issue) requires test takers to participate in a group discussion of their favorite film. The other task requires test takers to work in a group to make a decision. These tasks are like those used in the Cambridge English First (FCE) Exam (Cambridge English, www.cambridgeenglish.org/exams-and-tests/first/exam-format/). In Part Three of the FCE, the test takers are presented with some options to discuss (e.g., the ideas that a town has developed for attracting tourists). They are then required to agree on the best option (decision making). Tasks like those used in Part Three of the FCE as well as tasks described by Lam (this issue) offer interactional opportunities, but are open to variation in difficulty and in the nature of the interaction elicited, depending on the interactional behavior of the other test takers (see Nakatsuhara, 2006).

Tightly delineated tasks present their own challenges. Computer-based tests are well placed to deliver tasks that can elicit specific interactional strategies. For instance, one section of the PTE Professional™ (http://pearsonvue.com/ptepro/) asks test takers to listen to several situations and give an appropriate response. For example, the test taker might be presented with the following situation:

> You borrowed a jacket from your friend and accidentally spilled coffee on it. The coffee left a big stain. Mark wants his jacket back; what would you say to him?

Tasks like this might tap into test takers' IC and, if there are a sufficient variety of different scenarios, might successfully elicit a well-rounded picture of a test taker's IC. However, certain types of interactional strategies (such as negotiation of meaning, response to clarification requests, and response to misinterpretation of the message) might be more easily operationalized in these structured tasks than other strategies (e.g., checking comprehension). In their investigation of IC task design, van Batenburg, Oostdam, van Gelderen, and de Jong (2016) comment that proactive strategies (such as checking comprehension) that occur naturally in communication cannot be operationalized in scripted prompts. This means that tightly scripted tasks risk narrowing the focus of the IC evaluation to the types of IC that can be elicited. In addition, because such tasks lack interaction (i.e., a response from the interlocutor) and co-construction of meaning, we might argue that they do not truly capture IC. Indeed, it is not clear how performance on these tasks can be generalized to real-life contexts. The tasks are so carefully controlled that one might worry that even if a test taker were successful in tasks intended to focus on interactivity, their performance would not be able to be extrapolated to more unpredictable, real-life interactions. There is also the risk of what Bernstein, Van Moere, and Cheng (2010, p. 374) euphemistically describe as "off-construct coaching." Bernstein et al. (2010) use this phrase in relation to sentence-repetition tasks but the danger is equally applicable to tightly controlled IC tasks. Highly structured tasks might be exploited for their predictability and test takers could be provided with interactionally appropriate language chunks to memorize for use in a rote fashion during the test.

Clearly, therefore, there is scope for a detailed exploration of the extent to which specific speaking tasks operationalize IC. For instance, Taylor (2016) argues that the ability to ask questions is crucial for effective communication and is central to interaction. The interactive phase of Grades 7–12 of the Graded Exams in Spoken English (GESE, Trinity College London, www.trinitycollege.com/site/?id=368) explicitly evaluates test takers' competence in asking and understanding questions:

> The interactive task provides the opportunity for the candidate to demonstrate his or her ability to take control through the use of questioning techniques and language functions associated with requesting information, seeking clarification and encouraging further detail.
>
> (Trinity College London, 2009, p. 7)

Two ideas need to be unpacked here: First, the claim that the asking of questions is an important facet of IC; and, second, that a test task can successfully approximate the real-world conditions under which an individual asks questions. It would be interesting to see if the structure of this task operationalizes this facet of IC as intended and whether the test takers' performances can be extrapolated to the real-world contexts in which questions can occur. The Trinity Lancaster Spoken Learner Corpus (which includes a Questions Bank, http://cass.lancs.ac.uk/?page_id=1327) presents an excellent starting point for such investigations.

Taking another example, the speaking test for the Examination for the Certificate of Proficiency in English (ECPE, www.michiganassessment.org) is based on an interactionist perspective of second language performance developed by Swain (2001). Within this view, learner factors (e.g., world knowledge; language knowledge) and contextual factors interact to construct and influence acquisition and performance. An interactionist view maintains that dialogues are jointly constructed and distributed across the participants. The ECPE is a paired-format speaking test and comprises five stages, each of which places the test takers in situations where the interactional demands vary both by the characteristics of the task, the relationship between the speakers that is established by the task, and what needs to be accomplished (Bygate, Skehan, & Swain, 2001; Elder, Iwashita, & McNamara, 2002; Pica, Kanagy, & Falodun, 1993). In the third stage of the test, the test takers collaborate to choose between one of two options. The test design purposefully places test takers in a position of "mutual equality, and so the direction the discourse develops will not be pre-ordained and orchestrated by the [examiner]" (Skehan, 2001, p. 169). Later, in the final stage of the test, the test takers present their choice to a decision-maker (such as the chairperson of a committee). The situation is crafted to ensure that the power differential (van Lier, 1989) is key to the task design. Two examiners administer the test, taking turns to perform the role of interlocutor/facilitator. The examiners' roles in stages one and five are minimally scripted but stages two through four are tightly scripted and the examiners are expected to be peripheral to the interaction between the test takers.

The multi-layered nature of the ECPE appears very promising; the test appears to combine both a clear frame for the test takers as well as room for them to demonstrate their IC. However, as in the case of the GESE, the intended interpretation of the ECPE test score depends upon whether these test design elements successfully operationalize IC. This presumption has not yet been fully tested. Banerjee and Plough's (2016) analysis of the test takers' nonverbal behavior during the test (discussed in more detail in the next section) provides some preliminary evidence of its success at eliciting IC.[1] In addition to this work, a corpus of ECPE speaking test performances, organized by intended outcome (e.g., negotiation, explanation, persuasion) and other task characteristics, would be a valuable resource.

## *Interactional competence and nonverbal behavior*

The papers by Roever and Kasper, Ross, and Kim in this special issue reference nonverbal behavior (NVB) as a feature of IC. Roever and Kasper note "[a]ttention to the temporal dimension of interaction highlights the need for the assessment of talk to consider other vocal resources than language … When participants in interaction have visual access, they regularly mobilize a range of semiotic resources (e.g., gestures, gaze, body position and movement, objects and space) to accomplish the activity at hand (e.g., Mondada, 2014)" (p. 349). Kim remarks that all research of radiotelephony communication emphasizes the effect that the unavailability of NVB has on strict adherence to procedures and spoken conventions. Ross speculated that there would be more verbal backchannels in the phone-delivered OPI because opportunities for paralinguistic cues, such as eye contact and head nods, are not available. However, this expectation was not borne out. Nonetheless, together these papers highlight the need for words to fill in the interactional gaps when the visual channel is unavailable. In doing so, each highlights the contribution of NVB in an interaction.

This recognition of the influence of nonverbal behavior on the co-construction of dialogue can be found in work from the 1980s (see Kramsch, 1986). More recent research increasingly provides evidence that NVB is an important component of IC. Jenkins and Parra (2003) specifically targeted the influence of NVB in a high-stakes, local test of English proficiency for prospective international teaching assistants (ITAs). The participants in the study were Spanish-speaking and Chinese-speaking university graduate students. The test (characterized as a 20-minute interview) required each test taker to interact with three trained raters on four different tasks. Each rater provided a rating and a written justification of their evaluation for every section of the test. In order to examine the influence of paralinguistic features and NVB on examiners' ratings, Jenkins and Parra carried out microanalyses of the videotaped interviews. These analyses included: kinesic features (e.g., eye contact, body posture); paralinguistic features (e.g., voice volume, non-lexical sounds); nonverbal turn taking; and active listening strategies (e.g., head nodding, back channel cues). Jenkins and Parra also reviewed the examiners' written comments made during the test and post-test interview comments. Their findings indicated that nonverbal and paralinguistic behaviors played a critical role in whether or not students passed the test. Specifically, students who were rated linguistically proficient passed regardless of their NVB. Students who were rated linguistically weaker but employed nonverbal behaviors associated with active listening (e.g., frequent eye contact, smiling, forward lean, head nodding) "created an impression of … interactional competence" and also successfully passed the test. As a result of this study, evaluator training now incorporates nonverbal behavior; additionally, the scoring rubric was revised to "include listening comprehension and communicative competence … defined as including verbal, nonverbal, and paralinguistic interaction" (Jenkins & Parra, 2003, pp. 102–103).[2]

In a project to develop an empirically based rating scale for use in paired-format oral tests, Ducasse and Brown (2009) confirmed the salience of nonverbal behavior in the evaluation of speaking proficiency. The study consisted of 34 beginning-level learners of Spanish and 12 teacher-raters. The test takers participated in a paired speaking test and the teacher-raters each evaluated three video-recorded tests. The teacher-raters also provided verbal protocols on their evaluations, focusing specifically on the test takers' interactional abilities. Rater commentary on the test takers' interaction was grouped into three general categories: nonverbal interpersonal communication (gaze and body language), interactive listening (comprehension and supportive listening, which may include nonverbal behavior), and interactional management (horizontal and vertical cohesion). Ducasse and Brown (2009) report that the raters considered NVB a contributing factor to the success or failure of an interaction. They also noted that the raters viewed NVB as somewhat culture specific. In the words of one rater: "the girl … uses her hands when she talks. It gives a nice color and is more in tune with the Latin American speech and culture" (p. 434).

Nakatsuhara (2011) incorporated nonverbal behavior in her quantitative analysis of the effects of test takers' levels of extroversion and proficiency in group oral tests with three and four participants. She found that gestures were used in several of the groups of four. In one instance, the individual with the higher extroversion level suggested the turn-taking order by making a circle with her index finger in a counter-clockwise direction. Gesturing with one's hand to an adjacent person to propose turn taking was another form of NVB exhibited in groups of four. In some cases, this same gesture was used to encourage more reticent participants into the discussion and in still other cases it functioned as a means to pass one's turn.

May (2011) investigated those features of performance that affected the evaluations by four trained raters of candidates' ability for effective interaction. Analyses of data (rater notes, verbal recalls, rater discussions) revealed that body language, "including eye contact, facial expressions and gestures" (p. 136), was interpreted by raters as contributing to an effective, authentic interaction. However, May voices concern that including nonlinguistic features, such as body language, in an assessment of IC "would entail a consensus as to exactly what constitutes effective body language in a particular context" (p. 140). Referencing McNamara (1996), May suggests that "perhaps this Pandora's box has remained closed for very good reasons" (p. 140). We acknowledge that NVB can be context specific. This makes its evaluation more complex but not necessarily without reward. We suggest that the time has come for empirical investigations of NVB in speaking tests with a view to incorporating it into a definition of IC and thus the speaking construct.

It is necessary to note a distinction made between gesture and nonverbal behavior. Gullberg (2006a) defines gestures as a set of actions that are usually limited to movements of the arms and hands. This definition excludes functional (e.g., drinking) and symptomatic (e.g., scratching) actions as well as posture and proxemics, which "are not communicatively irrelevant but … are not typically part of the message that the speaker intends to convey" (Gullberg, 2006a, p. 104). NVB is a broader category of actions that includes, in addition to those movements encompassed within gesture studies, paralinguistic features (e.g., prosody, voice quality, non-lexical vocalizations); posture; all body movement; glances, gaze, eye contact; and facial expressions. In some sense, therefore, gestures can be thought of as a subset of NVB.

There has been comparatively more research of gesture (rather than NVB more broadly) primarily because Second Language Acquisition (SLA) researchers have pursued investigations grounded in the theoretical frameworks of gesture studies (e.g., Kendon, 2004; McNeill, 1992). Research and debate continue in areas such as a classification system (forms and functions), the relationship of gesture to thought and language, and in the case of L2 learners, the role of gesture in L2 development, and the influence of L1 gestures. Nonetheless, four interconnected characteristics of gestures are particularly relevant to NVB in L2 assessment:

1. Gestures are multifunctional, and there is not a one-to-one correlation between form and function. That is, a single form may serve multiple functions, which may be realized simultaneously.
2. Gestures and speech are intertwined. Although theories propose different descriptions and explanations of the relationship, all agree on the complexity of the interconnection.
3. The functions of gestures can be divided into two general categories: self-directed and other-directed. The former includes those (spontaneous) gestures when one is, for example, searching for a word or organizing one's thoughts. The latter includes those gestures that serve interactional purposes such as turn taking or back-channeling.
4. Gestures exhibit individual variation; yet, at the same time, uniformity of gestures within groups exists.

With the exception of the Jenkins and Parra (2003) study that we have already described, no studies focus specifically on the role of gestures in L2 proficiency testing. Looking at the relationship of gesture to proficiency and the role of culture in performing gestures, research in the L2 classroom and experimental contexts indicates that speakers produce more gestures in their L2 than in their L1 (Gullberg, 2006a, 2012), and that these seem to serve primarily compensatory functions for various linguistic difficulties in vocabulary, grammar, and fluency-related issues. Learners gesture while using circumlocution, to mark time onto space metaphorically, and to hold the floor during word searches. Little research has specifically addressed the relationship between level of proficiency and gesture production. However, initial findings of several studies conducted with speakers of various L1s (Chinese, Dutch, French, Swedish) and L2s (French, English, Swedish) indicate that as proficiency increases, the number of gestures used to identify co-referents decreases (Gullberg, 2006b, 2008; Yoshioka & Kellerman, 2006) and that the usage of beat gestures becomes more target-like, serving meta-pragmatic purposes, rather than simply marking prosodic features (McCafferty, 2006).

The body of work in cross-linguistic comparisons of gestures has yielded a number of very tentative generalizations (Gullberg, 2006a), as follows: The form of gestures and when they are performed is fairly consistent within a culture when the situation and content remain the same. Cultural norms determine conventionalized forms (e.g., the “OK” emblem), speech-associated, spontaneous gestures (e.g., pointing), and the amount of space that is used. The fact that gestures (and other NVB) are so culture-specific poses exceptional challenges, but the work already done in language testing (cf. Ducasse & Brown, 2009; May, 2011) indicates that the role of NVB in the co-construction of discourse cannot be ignored. It affects test takers’ performances in paired or group settings and has an impact on the scores awarded.

A research agenda designed to describe and explain the role of nonverbal behavior in second language testing should minimally address four areas. First, we need to better understand the structural components of NVB and the correspondences between NVB forms and the functions they achieve. In the case of NVB that serves compensatory functions, for example, we need to delimit the scope of ‘allowable’ compensatory forms. As noted previously, Banerjee and Plough (2016) reported very early work in this area. Analyzing the paired-format, face-to-face speaking portion of the Michigan Language Assessment ECPE, Banerjee and Plough independently annotated two tests (four test takers) for all NVB and coded them by function. Following a constant comparative method (Boeije, 2002; Glaser & Strauss, 1967) the annotations were then shared and points of agreement and divergence compared. Where there was disagreement, video recordings were viewed and discussed to reach consensus. The suggested correspondences between the NVB and the function achieved were based on the L1 of the test takers and L2 gestures as well as current L2 testing research (preliminary findings are summarized below). More work of this type is needed in order to develop a more robust theoretical foundation for interactional abilities.

The second area that needs to be addressed is the relationship between nonverbal behavior, task and individual characteristics. The complex interaction of these variables has been identified in a validation study of a paired-format test of intermediate speaking proficiency that has been developed in multiple languages (Plough, 2014). The test consists of three tasks, one of which requires examinees to collaborate to reach a decision. In the development process, test trials were conducted with native speakers – as both examiners and test takers. During the collaboration task for a test trial in Spanish, one pair of test takers simultaneously stood up and changed the orientation of their chairs to be facing each other, where they stayed throughout the task. This supports Ducasse and Brown's (2009, p. 433) empirically derived characterization of nonverbal interpersonal communication, which "includes the flow of … body positioning, that physically support[s] what takes place verbally in the interaction." In the case of Plough's (2014) study, however, this cooperative stance did not occur in all instances of the collaboration task; there were test trials in which examinees did not face each other or one member of the pair did not face the other. More work is needed to tease apart the complex interactions of NVB, task characteristics, and individual variables, which clearly has implications on and informs the third area for investigation.

Third, the evaluation of NVB, specifically the criteria for evaluating the effectiveness and the appropriateness of the nonverbal behavior as well as the relationship between NVB and existing linguistic criteria, must be addressed. The L2 gesture studies, the early work of Bailey (1982, 1984), Lazaraton (1996), Ross and Berwick (1992), and Berwick and Ross (1993), and the more recent studies summarized above provide a beginning to the development of the scope of appropriate and effective NVB usage. Of note is the fact that NVB has regularly emerged as an influential variable in speaking contexts in which individual factors such as first language, age, and personality characteristics as well as contextual factors such as conversational topic and venue differ in significant ways. It is not a feature that can be disregarded as merely idiosyncratic to a particular situation or individual but one that consistently affects and is affected by features that all play critical roles in the dynamics of social interaction.

Rater training, of course, is essential. Even when criteria have been developed, issues of different interpretations among raters and the subsequent difficulty of reaching consensus on the effectiveness and appropriateness of the verbal and nonverbal communicative characteristics inherent in language and culture cannot be overstated. Indeed, individual variation in the interpretations of the same NVB is further complicated by the fact that individual variation exists in the production of NVB in one's first language. Nevertheless, the fields of language testing and SLA have a long history of addressing variation in performance (cf. Ortega, 2014; Ross, 2012). Drawing on our methodological and conceptual advances, we suggest that a 'scope of NVB' be developed, just as ranges of performance are created for linguistic performance.

A final area to be addressed with respect to the role of nonverbal behavior in second language testing is the relationship between NVB and proficiency. Preliminary findings from the Banerjee and Plough (2016) study mentioned previously suggest that higher rated speakers tend to exhibit more NVB (e.g., head nods as listener; beat gestures to accompany speech), which has been interpreted as increased engagement. Additionally, the use of physical space and of direct eye gaze was greater among higher rated individuals, which has been interpreted as higher levels of comfort and confidence. It is not entirely clear how these findings compare to the gesture studies by Gullberg (2006b, 2008), and Yoshioka and Kellerman (2006); these studies suggest that more gestures might be indicative of lower (rather than higher) proficiency. It is important to note, however, that the data in the Banerjee and Plough study (2016) are from participants with the same L1 (Greek) using the same L2 (English). It is possible that, for these data, the frequency of gesturing is subject to individual variation, and that the acceptable use of space is determined by cultural norms. That is, it remains an empirical question if the NVB can be attributed to an influence of the L1.

## Interactional competence and the electronic mediation of speaking assessments

Traditionally, face-to-face speaking tests have required the test takers and the examiners to be in the same physical location. However, with improvements in technology, it has become possible to envisage the assessment of speaking via video conferencing. If it were possible to deliver speaking tests that were equivalent regardless of the way in which they were delivered (i.e., the same physical space or via video conferencing) then test takers might be tested without having to travel long distances to a location where there are examiners. Additionally, it might be possible to match test takers with available examiners (regardless of physical location) and deliver tests in a shorter time frame. Central to the viability of this opportunity, however, is the concept of equivalence of test administration and what that entails.

Nakatsuhara, Inoue, Berry, and Galaczi (2016, 2017) have investigated the use of video-conferencing technology to deliver the IELTS Speaking Test. They compared test taker and examiner behaviors in both the standard test delivery (where the test taker and examiner are in the same room) and video-conferencing modes, analyzing the test takers' scores and linguistic output, the examiners' test management and rating behaviors, and both the test takers' and the examiners' perceptions of the two delivery conditions. The analysis of the test takers' scores indicated that lower proficiency and older test takers were more likely to perform worse on the test delivered via video conferencing. The analytic criteria of fluency and pronunciation were the most likely to be affected by the delivery mode. Overall, however, the study established that there were no statistically significant differences in the scores achieved in the different delivery modes.

Nakatsuhara et al.'s (2016, 2017) analysis of test takers' linguistic output focused on the production of language functions. Five language functions were used differently in the two delivery modes. Clarification questions were used more frequently during the video-conferencing mode; this was attributed to poor sound quality. More test takers elaborated on their opinions in the video-conferencing mode; this was attributed to a misreading by the test takers of the examiners' nonverbal cues that they had spoken enough. The interactional functions of comparing, suggesting, and modifying/commenting/adding were used significantly more frequently during the standard delivery mode; this might be attributed to the comment by some test takers that it was not as easy to relate to the examiner during the video-conferencing mode of the test. Indeed, the test takers reported that the standard mode was easier and that they were able to better understand the examiner during this mode of the test. Many test takers commented that they were able to understand the examiner better because they could clearly see their facial expressions and posture.

Turning to the data from the examiners with respect to their exam management behavior during the two delivery modes, Nakatsuhara et al. (2017) report that the examiners altered their behavior for the video-conferencing mode. The examiners were much more sensitive about their speech articulation; they typically spoke more slowly and enunciated more deliberately for the video-conferencing mode. The examiners were also conscious of the effect of gestures and body language but appeared unsure about how to remain natural during the video-conferencing delivery mode. It is interesting to note that concerns about how to behave appeared to take attentional resources from the act of rating. One examiner commented that, during the standard delivery mode, they were able to pay more attention to a test taker's language because they were not pre-occupied with managing the technology. The examiners also felt that poor sound quality during the video-conferencing mode resulted in their feeling less certain about the scores that they awarded.

Nakatsuhara et al. (2017) conclude that the video-conferencing mode can be considered a "parallel alternative" (p. 15) to the standard delivery mode. They attribute some of the differences between the modes to resolvable technical hiccups and others to familiarity with the delivery mode alternatives. The test takers and examiners were all more familiar with the standard delivery mode. Arguably, practice in using video-conferencing would make all participants more comfortable with it as a test delivery mode. However, more research is needed to better understand the effect of the test delivery mode on the interaction between the test taker and the examiner. Nakatsuhara et al.'s (2016, 2017) language function analysis indicates that there were differences in the interactions as well as in how a test taker's language proficiency might be evaluated. The question then arises of whether IC might be operationalized differently in the two delivery modes.

Here we turn to the wider literature on the use of video-conferencing for meetings, collaborative work, and telemedicine. O'Conaill, Whittaker, and Wilbur (1993) compared the features of talk during face-to-face meetings and those meetings held by video conference. They used two different delivery mechanisms for the video-conferencing meetings: an ISDN connection and another much more stable system (the now defunct London Interactive Video Education System [LIVE-NET]). They analyzed the meeting discussions for: backchannels, interruptions, overlapping speech, explicit handovers in turn taking, the total number of turns, turn length, and turn distribution. The ISDN delivered video-conferencing meetings were interrupted by transmission lags but the video and audio-quality of LIVE-NET delivered meetings was stable. The instability of the ISDN channel had a marked effect on the meeting interactions. The listeners interrupted less frequently and were less likely to anticipate turn endings. Both video-conferencing meetings were distinguished from face-to-face meetings in the use of back channels and the management of turn taking. Listeners in video-conferencing meetings were less likely to produce back channels and speakers handed over turns formally by asking a question or inviting a specific person to speak.

O'Malley, Langton, Anderson, Doherty-Sneddon, and Bruce (1996) compared performance on a collaborative task through face-to-face and videoconferencing. The task was an information gap activity (a 'map task') not dissimilar to the kinds of tasks used in the second language classroom. The participants had to share a route, negotiating the information that they had in common and the information that they did not share. O'Malley et al. (1996) found that the video-conferencing mediated performances were less efficient and effective (as measured by successful completion of the task) than the face-to-face interactions. They partially attribute this to transmission delays which, even when they are slight, resulted in the participants using far more verbal checks and back channels. This interfered with turn taking. O'Malley et al. (1996) also found that participants in the video-conferencing mediated task looked at their interlocutor far more than when the participants were in the same room. One of their hypotheses for this relative overuse of gaze is that the video-conferencing mode is relatively ineffective in conveying visual information. As a consequence, visual cues might be missed or misinterpreted and therefore demand more attention. A later study by Sanford, Anderson, and Mullin (2004), replicated the task and more carefully varied the video-conferencing technology for sound and picture quality. They found that the nature of the interaction (turn-taking management and the establishing of mutual knowledge) as well as the success of the task were affected by the video-conferencing technology.

Even setting aside flaws in connectivity (which we must assume will reduce with time and technological advances) the nature of the interaction is clearly altered by the medium (face-to-face or videoconferencing) and verbal markers of IC are therefore employed very differently. What emerges is that *social presence* is of key importance. Social presence is defined as the extent to which the other person in the communication is perceived to be real and the consequent sense of there being an interpersonal relationship (Short, Williams, & Christie, 1976). The video-conferencing medium impedes each participant's feeling that the other person is 'real' and their ability to accurately read their interlocutor's behavior. This results in qualitatively different interactional behaviors and also strongly suggests that IC will need to be described and evaluated differently depending on a speaking test's mode of delivery.

### Interactional competence and computer-based tests

Despite the scope of this special issue, it has not addressed the question of how IC might be assessed in a computer-based test. Computer-based tests offer many advantages in the assessment of speaking, including standardization of input, a greater variety of tasks and intended audiences, and a guaranteed range of language functions elicited. Computer-based tests also have the potential to resolve the conundrum May (2011) leaves on the table at the end of her study of interaction in a paired speaking test. This study aimed to identify the features of test takers' performances that raters find salient when evaluating interactional effectiveness. May (2011) found that, although raters were able to assess the test takers separately for some aspects of interaction, there were other aspects of interaction that were mutually co-constructed (e.g., cooperative behavior) for which test takers could not be given individual assessments. May (2011) suggests that, given our current understanding of the IC construct, it is perhaps best to award test takers with a joint IC score. A joint score best represents the fact that the interaction has been co-constructed but is difficult to implement in high-stakes testing contexts, where test takers need to receive individual scores. In such contexts, the computer-based delivery of tasks that tap IC could 'isolate' a test takers' IC in the sense that, since the test taker is responding to a standardized, computer-delivered prompt, the IC that they demonstrate can be attributed to them alone.

However, as noted earlier, computer-based tests currently lack interactivity, which means that certain aspects of the IC construct cannot be operationalized. This means that any computer-based evaluation of IC would, because of the current constraints of the technology, offer a narrow interpretation of the IC construct. For instance, computer-based tests cannot elicit a test taker's grasp of interaction management functions such as topic development or indicating comprehension through backchannels and other responses

(Galaczi, 2010). Indeed, even though interlocutors are a potential source of construct irrelevant variance, they also have more potential to capture a wider spectrum of evidence of a test taker's IC. To date, computer-based tests have not been able to harness the less predictable aspects of real-life language situations (the 'unexpectedness' factor) such as sudden changes in circumstances, or the negotiation of meaning when a word or the intention of the interlocutor has been misunderstood (or perhaps simply not heard).

Technology progresses apace, however, and the interactional constraints currently present in computer-based testing are on track for being removed (if not in the short run, then certainly in the medium-term) as evidenced by the advances in the interactivity now being developed in the teaching and assessment of reading comprehension. Li, Shubeck, and Graesser (2016) describe the potential of an intelligent tutoring system, AutoTutor (http://ace.autotutor.org/IISAutotutor/index.html), for use in learning-oriented language assessment. AutoTutor was originally developed to support the learning of students taking an introductory course on computer literacy but its uses have since been expanded to include learning academic subjects as well as language learning (specifically reading comprehension). At the moment, students have to type their responses to the questions posed by the tutor avatar. The avatar's interaction with the students comes in the form of brief positive or negative verbal feedback (e.g., "great" or "oops, you did not get it!"), pumps (probing for more information), hints (suggestions to direct the student), and prompts (specific, leading questions). They are adaptive and have been shown to successfully promote learning gains.

Shore, Wolf, O'Reilly, and Sabatini (2017) describe two scenario-based assessments (SBAs) in reading that have been developed for K–12 students. The Global, Integrated Scenario-Based Assessment (GISA) was developed for the summative assessment of reading proficiency and the English Learner Formative Assessment (ELFA) was developed for 12–14-year-olds as a classroom-based, formative assessment of reading comprehension. GISA presents students with thematically related reading tasks (selected, for example, to help them to decide whether their neighborhood should have a community garden) that have been carefully sequenced "both to *model* skilled performance and to *gather evidence* on what parts of a more complex task students can or cannot do" (p. 237). ELFA also frames the students' reading within a scenario, explaining the source of the reading passages (e.g., a magazine), and the writers' intentions. The students then read texts in sequence and answer questions based on their reading. The texts and tasks have been designed to provide teachers with information about the reading subskills that students have not yet fully mastered.

AutoTutor and the SBA approaches being developed at Educational Testing Service (ETS) are extremely promising. It is now possible to imagine a computer-delivered, scenario-based speaking test that perhaps harnesses the principles of online gaming and could very profitably elicit evidence of IC at various levels of task difficulty. Until speech recognition and on-the-fly processing of speech are incorporated into these systems, they can only support the teaching and evaluation of text-based skills. However, artificial intelligence (AI) devices, such as Amazon Echo and Google Home, are fast infiltrating homes across North America if not elsewhere in the world. Online language learning tools such as Duolingo and Babbel are also becoming more widely used. It is, therefore, only a matter of time before the ability to interact with an AI interlocutor is available.

## Final thoughts

Even as we come to the end of this paper and this special issue we are aware that there are layers of this figurative onion that remain to be unpeeled. Three topics, in particular, stand out.

The first is nonverbal behavior. We suggest that nonverbal behavior and interactional competence are intrinsically connected, particularly if, as Kendon (2004) maintains, "we are to have a full understanding of how utterances within the context of an interaction are intelligible for the participants" (p. 3). We, therefore, must embark on disentangling the complexities of NVB in terms of its forms and functions; the interactions among NVB, tasks and individual characteristics; the criteria by which to evaluate it; and creating rating scales and training examiners in their use.

The second is operationalization. The extent to which specific speaking tasks operationalize IC still needs to be established. Additionally, it is necessary to ensure that different tasks provide test takers with equal opportunities to demonstrate IC. We also still need to answer the question of whether we should explicitly test for certain features of IC or whether IC should be viewed more globally. Computer-based or computer-mediated testing formats offer many advantages, among them standardization of administration and providing testing opportunities to individuals unable to travel to testing locations. Yet, we must come to a better understanding of how technology impacts the operationalization and assessment of IC. We must do this in the full acknowledgement that face-to-face communication forms only a proportion of our spoken interactions. The remainder are conducted over the telephone or via videoconferencing. Consequently, technology-mediated speaking tasks are no longer outliers in the target language use domain.

Here, we note that with IC comes the question of theoretical stance and strict adherence to one perspective. We suggest that the field must sincerely engage in a discussion of the contributions that each research study along the theoretical continuum makes to an understanding of the speaking construct. Much would be lost if one assumes that IC can only be investigated by adopting a co-constructionist view of language and discourse (Jacoby & Ochs, 1995). Similarly, one need not wholly embrace a psycholinguistic-individualistic perspective at the other end of the continuum for purposes of test validation. Indeed, as the papers in this issue show us, a grounded ethnographic approach to discourse analysis and conversation analysis offer us relatively comprehensive descriptions, concise definitions, and quantification of interactional features.

Finally, we return briefly to the data presented by the researchers in this special issue to suggest possible implications for the relationships between proficiency and IC. Roever and Kasper provide evidence from their examination of preliminaries that production of this interactional feature does increase with proficiency level. Lam, cautioning against generalizing based on an exploratory study, speculates that test takers produce more well-developed contingent responses as proficiency increases. Again, however, Lam emphasizes that quality of response (e.g., appropriateness) cannot be overlooked and that his study was not intended to investigate developmental levels. The domain experts in Kim's study make a distinction between professional knowledge and linguistic proficiency. To a certain extent, IC compensated for limited proficiency. In contrast, lack of knowledge of the norms and conventions of the specific context (i.e., IC) was criticized even when level of proficiency was not negatively evaluated. To these initial findings, we add Ross's data in which backchannels as indicators of IC could not be isolated as a distinct criterion. In light of the studies in this special issue, the interaction of IC with other linguistic competences at different proficiency levels is a subject deserving investigation.

At this point, we are obligated to ask: What do we gain by invoking IC as part of the speaking construct? We would contend that context is central to the speaking construct, that is, knowing what to say to whom, when and how to say to it. This, of course, is the classic definition of pragmatics. So, the question then becomes what distinguishes IC from pragmatics? Although the two are interconnected, we would maintain that they are distinct competences. Both integrate various competences (e.g., grammatical, textual) for meaningful and purposeful communication. However, IC is *necessarily* about building and maintaining relationships, an aspect of the co-constructed nature of speech (see Young, 2011).

We conclude by acknowledging that we have explored topics that have long posed significant challenges for the field of language assessment when we endeavor to create tests from which valid inferences of test taker proficiency can be made. Additionally, we realize that we have merely scratched the surface of these topics. Nonetheless, it is our hope that the studies presented in this special issue shed new light, and scratch a different part of the surface, on these topics.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Note

1. The data analyzed for this research study was provided by Michigan Language Assessment.
2. See Jenkins and Parra (2003, p. 91) for a review and discussion of the research examining a "language threshold level beneath which compensatory strategies or pragmatic skills do not make up for linguistic problems" within testing and classroom contexts.

## ORCID iD

Jayanti Banerjee https://orcid.org/0000-0002-8175-0887

## References

Bailey, K. M. (1982). *Teaching in a second language: The communicative competence of non-native speaking teaching assistants*. Dissertation Abstracts International, *43*, 2812A3441A.

Bailey, K. M. (1984). The "foreign TA problem." In K. M. Bailey, F. Pialorski, & J. Zukowski-Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 3–15). Washington, DC: National Association for Foreign Student Affairs.

Banerjee, J., & Plough, I. (June, 2016). *Behavior in speaking tests: A preliminary model of interaction*. Work-in-progress presented at the annual Language Testing Research Colloquium (LTRC), Palermo, Italy.

Bernstein, J., Van Moer, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377.

Berry, V. (2007). *Personality differences and oral test performance*, Frankfurt am Main, Germany: Peter Lang.

Berwick, R., & Ross, S. (1993). Cross-cultural pragmatics in oral interview proficiency strategies. Retrieved from https://eric.ed.gov/?id=ED366173.

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity*, *36*(4), 391–409. Retrieved from http://doi:10.1023/A:1020909529486.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341–366.

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

Bygate, M., Skehan, P., & Swain, M. (Eds.) (2001). *Researching pedagogic tasks: Second language learning, teaching, and testing*. London: Pearson Education.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, *26*(3), 423–443.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing*, *19*(4), 347–368. Retrieved from http://doi:10.1191/0265532202lt235oa.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.). *Computer-based assessment of foreign language speaking skills: CBA 2010* (pp. 29–51). Luxembourg: Publications Office of the European Union.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Gullberg, M. (2006a). Some reasons for studying gesture and second language acquisition (Homage à Adam Kendon). *International Review of Applied Linguistics*, *44*, 103–124.

Gullberg, M. (2006b). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. *Language Learning*, *56*(1), 155–196.

Gullberg, M. (2008). A helping hand? Gestures, L2 learners, and grammar. In S. G. McGafferty & G. Stam (Eds.). *Gesture. Second language acquisition and classroom research* (pp. 185–210). New York: Routledge.

Gullberg, M. (2012). Gesture analysis in second language acquisition. In C. Chapelle (Ed.). *The encyclopedia of applied linguistics*. Retrieved from http://doi:10.1002/9781405198431.wbeal0455.

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

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*(1), 90–107.

Johnson, M. (2001). *The art of non-conversation: A reexamination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372.

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

Li, H., Shubeck, K. T., & Graesser, A. C. (2016). Using technology in language assessment. In J. Banerjee & D. Tsagari (Eds.). *Contemporary second language assessment* (pp. 281–297). London: Bloomsbury Academic.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397–421.

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

McCafferty, S. (2006). Gesture and the materialization of second language prosody. *International Review of Applied Linguistics in Language Teaching*, *44*(2), 197–209.

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

McNeill, D. (1992). *Hand and mind: What the hands reveal about thought*. Chicago, IL: University of Chicago Press.

Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics*, *65*, 137–156.

Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in paired speaking tests. *Research Notes 25*, University of Cambridge ESOL Examinations, 15–20. Retrieved from www.cambridgeenglish.org/images/23144-research-notes-25.pdf.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483–508.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2016). Exploring performance across two delivery modes: Face-to-face and -conferencing delivery – a preliminary comparison of test taker and examiner behavior. *IELTS Partnership Research Papers*, *1*. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-andresearch/research-reports.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly*, *14*(1), 1–18.

Norris, J. M. (2016). Current uses of task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244.

O'Connaill, B., Whittaker, S., & Wilbur, S. (1993). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, *8*(4), 389–428.

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., & Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, *8*(2), 177–192.

Ortega, L. (2014). Trying out theories in interlanguage: Description and explanation over 40 years of L2 negation research. In Z. H. Han & E. Tarone (Eds.). *Interlanguage: Forty years later* (pp. 173–202). Amsterdam: John Benjamins.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction research. In G. Crookes & M. Gass (Eds.), *Tasks and language learning integrating theory & practice* (pp. 9–34). Clevedon: Multilingual Matters.

Plough, I. (June, 2014). *The local informing the global*. Poster Presentation. Language Testing Research Colloquium 2014. VU University Amsterdam, Netherlands.

Ross, S. J. (2012). Claims, evidence, and inference in performance assessment. In G. Fulcher & F. Davidson (Eds.). *The Routledge handbook of language testing* (pp. 223–333). New York: Routledge.

Ross, S. J., & Berwick, R. (1992). The discourse of accommodation in oral proficiency examinations. *Studies in Second Language Acquisition*, *14*, 159–76.

Sanford, A., Anderson, A. H., & Mullin, J. (2004). Audio channel constraints in video-mediated communication. *Interacting with Computers*, *16*, 1069–1094.

Shore, J. R., Wolf, M. K., O'Reilly, T., & Sabatini, J. P. (2017). Measuring 21st-century reading comprehension through scenario-based assessments. In M. K. Wolf & Y. G. Butler (Eds.). *English language proficiency assessments for young learners* (pp. 234–252). New York: Routledge.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London and New York: Wiley.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language, teaching, and testing* (pp. 167–185). London: Pearson Education.

Swain, M. (2001). Integrating language and content teaching through collaborative tasks. *The Canadian Modern Language Review*, *58*(1), 44–63.

Taylor, C. (May, 2016). *Revisiting the speaking construct: Multiple perspectives*. Paper presented at the EALTA Conference, Valencia, Spain.

Trinity College London. (2009). *Exam information: Graded examinations in spoken English (GESE)*. London: Trinity College London.

van Batenburg, E. S., Oostdam, R. J., van Gelderen, A. J. S., & de Jong, N. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Language Testing*, *35*(1), 75–100.

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

Yoshioka, K., & Kellerman, E. (2006). Gestural introduction of ground reference in L2 narrative discourse. IRAL – *International Review of Applied Linguistics in Language Teaching*, *44*(2), 173–195.

Young, R.F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*, Vol. 2 (pp. 426–443). London: Routledge.