Assessing second language speaking proficiency

Jong, N.H. de

ANNUAL REVIEWS

Annual Review of Linguistics

Assessing Second Language Speaking Proficiency

Nivja H. de Jong

Leiden University Centre for Linguistics and Leiden Graduate School of Teaching, Leiden University, Leiden, Netherlands; email: n.h.de.jong@hum.leidenuniv.nl

Annu. Rev. Linguist. 2023. 9:541–60

First published as a Review in Advance on November 7, 2022

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-030521-052114

Keywords

speaking proficiency, language assessment, interaction

Abstract

In today's global economy, most people all over the world need to speak a second language (L2) for study, work, or social purposes. Assessment of speaking, either in the classroom or as an external exam, is therefore an important task. However, because of its fleeting nature, the assessment of speaking proficiency is difficult. For valid assessment, a speaking test must measure speaking proficiency without construct-irrelevant variance, for instance, due to tasks, raters, and interlocutors. This article begins by bringing together insights from different disciplines to develop a multi-componential construct of speaking proficiency, which includes linguistic and strategic competencies. Because speaking usually takes place in conversation, the ability to take part in interaction, including rapid prediction, is described as part of the speaking construct. Next, the factors that need to be controlled when making a speaking assessment are discussed. Finally, challenges and ideas for future research are briefly described.

1. INTRODUCTION

Speaking is a language skill we use every day. In today's globalized world, most of us also need second language (L2) speaking skills for study, work, and social purposes. However, among L2 teachers, teaching speaking is seen as one of the most daunting tasks (Goh & Burns 2012), especially when faced with large classes. This is no wonder because, for example, teachers would have to give feedback on individual learners' fleeting speech to foster learning. Likewise, speaking is seen as difficult to assess (Fan & Yan 2020), leading to the fact that assessment of speaking has been neglected. For example, speaking was not a compulsory part of one of the major large-scale English tests [test of English as a foreign language (TOEFL®)] until 2005 (Isaacs 2016; for an overview of the history of testing speaking, see Fulcher 2003).

Speaking is seen as difficult to assess because of its transient, context-dependent, and dynamic nature and also because speaking takes place in different forms (monologue, paired conversation, group discussion, etc.). To assess speaking proficiency, multiple factors have to be taken into account, including tasks, interlocutors, scoring criteria, and raters (Fulcher 2015, McNamara 1996).

To assess any ability includes measuring the ability. To measure someone's length, you can use a tape measure, and to measure that person's speed on a 100-m sprint, you can use a stopwatch. In both cases, there is a direct relationship between the numbers on the tape measure or stopwatch and the aspect to be measured. For speaking proficiency, however, there is no direct measurement possible. No tape measure or stopwatch can gauge someone's speaking abilities. This article therefore first describes the construct of speaking proficiency to be measured in assessment. Next, the multiple factors that need to be taken into account are reviewed. The article finishes with current challenges and directions for future research.

2. DEFINING SPEAKING PROFICIENCY

Before defining speaking proficiency, one needs to define what it means to speak. Speaking entails translating thoughts to sounds, mostly while taking part in conversations. This brief description of speaking reflects how two main research fields have influenced thinking about the construct of speaking for language assessment. The psycholinguistic tradition views speaking as an individual cognitive skill, whereas the sociolinguistic-interactional tradition views speaking as a social interactional ability. In a complete definition of speaking proficiency, both views need to be represented.

2.1. Psycholinguistic Approaches to Speaking

In traditional psycholinguistic models of speaking, the speech production process consists of multiple stages (Dell et al. 1997, Levelt et al. 1999). Although models of speech production diverge on which processes in speech production may run parallel to each other and which may interact (Morsella & Miozzo 2002), they agree on the general stages in speech production: preverbal planning, lexical retrieval processes that include retrieving morphosyntactic knowledge to build correct structures and phonological knowledge to plan intelligible sounds, and finally, articulatory planning. During these processes, speakers monitor their progress (Dell et al. 1997, Levelt et al. 1999). To communicate successfully then, a speaker needs to have the processes at each of the stages of speech production run efficiently. Speaking, from the cognitive perspective, can therefore be broken down into several subskills: a skill to conceptualize the preverbal message, a skill to retrieve the correct lexical items quickly along with their morphosyntactic characteristics, a skill to retrieve the appropriate sounds with these lexical items and to plan them as connected speech, a skill to send motor programs to the articulatory muscles to produce intelligible sounds, and finally, skills to efficiently monitor one's speech. In the L2 acquisition and

assessment literature, the model by Levelt et al. (1999) has been adapted to describe processes of L2 production (De Bot 1992, Kormos 2006, Segalowitz 2010), and these adapted models assume that the speech production processes for L1 and L2 are largely the same. They add that, unlike for L1 speakers, many of the processes involved after the preverbal planning stage are not automatized, and L2 learners need to rely more on (slow) declarative knowledge in addition to procedural knowledge (e.g., Kormos 2006). In addition, L2 learners need to tap more into strategies due to a lack of (for instance, lexical) knowledge and due to more limited cognitive resources when automatization falls short (Skehan 1998).

Pickering & Garrod (2004, 2013) review the psycholinguistic literature and state that the traditional models for speaking and comprehension (which have influenced language assessment) do not fit conversational data. In psycholinguistic research and handbooks, models for production and comprehension are separate. This was also reflected, for example, in the organizational structure of the Max Planck Institute for Psycholinguistics (one of the leading institutes in psycholinguistics) in the 1990s through 2000s, with separate departments for production and comprehension. Following this separation in the field, a traditional psycholinguistic model combining production and comprehension assumes that conversation is a serial monologue (Pickering & Garrod 2013). In conversation, interlocutors would alternate between the roles of speaker and listener. As a speaker, they construct a message and translate this to words and sounds; their interlocutor goes through the stages of comprehension while listening to the message and then takes on the role of the speaker to respond with a new message. In conversation, however, there is no clear separation of speaker and listener at any one time: Contributions overlap, with the listener providing verbal or nonverbal feedback to the speaker and the speaker altering their contribution accordingly (Pickering & Garrod 2013). When turns between participants do not overlap, the gap between turns is approximately 200 ms (Stivers et al. 2009), which is the fastest possible reaction time for humans responding with any single action. Because we know that language production is much slower than this—articulating a word from scratch takes 1000 ms (Bates et al. 2003) and producing a short clause as much as 1500 ms (Griffin & Bock 2000)—it follows that, generally, the response in conversation is planned during the middle of an incoming turn (Levinson 2016).

Pickering & Garrod (2013) therefore assume a central role for prediction in language production, comprehension, and dialogue. Their assumption is in line with a recent paradigm shift in cognitive science and psycholinguistics, referred to as the predictive turn (Huettig et al. 2022). Likewise, Levinson (2016) states that in conversation, fast speech-act prediction or recognition is crucial. According to Pickering & Garrod (2013), prediction is important not only for anticipating what the interlocutor will say but also as a forward model for the speaker to efficiently construct their own utterances (see also Hickok 2012). Another finding from the more recent psycholinguistic literature is that, in dialogue, successful speakers align their messages to each other, including alignment in terms of lexis and grammar (e.g., Branigan et al. 2000). This interactive alignment mechanism, as proposed by Pickering & Garrod (2004), leads interlocutors to repeat each other's expressions, making dialogue highly repetitive, in which fixed expressions become dialogue routines. Such repetitions and fixed expressions simplify language processing and make turns in dialogue more predictable.

The conclusion from the psycholinguistic approach is that L2 speakers not only need to possess cognitive skills to produce language (fast) but also need the cognitive skills to rapidly predict and anticipate the current incoming message in a conversation.

2.2. Social-Interactional Approaches to Speaking

The recent emphasis on dialogue in psycholinguistic research is inspired by and in line with the sociolinguistic-interactional research paradigms that view human communication as a social

activity. Conversation analysis is the research field in sociology that specializes in studying social interaction. From such studies, we know that interaction, although dynamic and spontaneous, is sequentially organized in adjacency pairs such as greetings and counter greetings or questions and answers. Also, the transition of turns between two participants in a conversation is organized (Sacks et al. 1974). Importantly, meaning in interaction is coconstructed by the speakers involved, and thus, interactional competence is "distributed across participants" (Young 2011, p. 430). Because interaction is reciprocal, the skill of speaking in interaction must include—in addition to the individual cognitive skills already mentioned—the ability to listen attentively, to design the message for the recipient (recipient design) (Drew 2012), to manage the conversation, and to use appropriate nonverbal behavior. As Clark (2002) put it, in dialogue, a speaker is simultaneously delivering their primary message (the propositional content) and also the collateral message, in which the speaker is informing the listener about their performance and their understanding.

Concerning L2 speaking, research has shown that although some of these conventions in interaction may hold cross-linguistically (Dingemanse et al. 2015, Stivers et al. 2009), this does not mean that L2 learners can put them into action straight away (Pekarek Doehler & Pochon-Berger 2015). Interactional competence (Kramsch 1986) in L2 is therefore part of the speaking construct that progresses as learners develop their speaking skills, and it needs to be measured as part of the construct in L2 assessment, including nonverbal behavior (Plough et al. 2018, p. 429). But if interaction is coconstructed and speakers involved in interaction are jointly making meaning and fulfilling a communicative task (Galaczi & Taylor 2018), how can the language tester award an individual score to one of the speakers (Chalhoub-Deville & Deville 2005, p. 826)? This matter is further discussed below when reviewing the literature on the interlocutor as one of the factors that need to be taken into account in the assessment of speaking. For now, it is clear that speaking proficiency involves cognitive individual skills that include anticipation or prediction in interaction, as well as social conversational skills that show up through joint interaction.

2.3. Models of Communicative Ability

In addition to theories and findings from research fields investigating the psycholinguistics and interactional nature of speaking, models of L2 communicative ability rely heavily on Hymes' (1972) notion of communicative competence. Reflecting most of the literature briefly described in Sections 2.1 and 2.2, models of communicative ability for L2 learners (Bachman & Palmer 1996, 2010; Canale & Swain 1980; Celce-Murcia 2007; Hulstijn 2015) thus combine such theories and explicate how being able to communicate involves many different aspects. For instance, the model that was highly influential in language assessment by Bachman & Palmer (1996, 2010), based on Canale & Swain (1980), is a hierarchical model in which language ability subsumes language knowledge and strategic competence. Language knowledge, in its turn, is depicted as consisting of both organizational (grammatical plus textual) and pragmatic (functional plus sociolinguistic) knowledge (Bachman & Palmer 1996). Hulstijn (2015) explicitly adds speed of processing to the language domain of this model. Hulstijn (2015) argues that the thus-called linguistic cognition part of L2 proficiency, which consists of knowledge and speed in the phonetic, phonological, morphological, syntactical, and lexical/pragmatic domains, should be seen as the core of L2 proficiency, with other aspects such as interactional ability and strategic competence as peripheral. Hulstijn's (2015) core of proficiency, can perhaps be seen as the core from the psycholinguistic, cognitive perspective.

2.4. Speaking Proficiency

In this section, I describe overall speaking proficiency as a psycholinguistic and social-interactional functional ability, primarily based on Bachman & Palmer's (1996, 2010) model of language ability

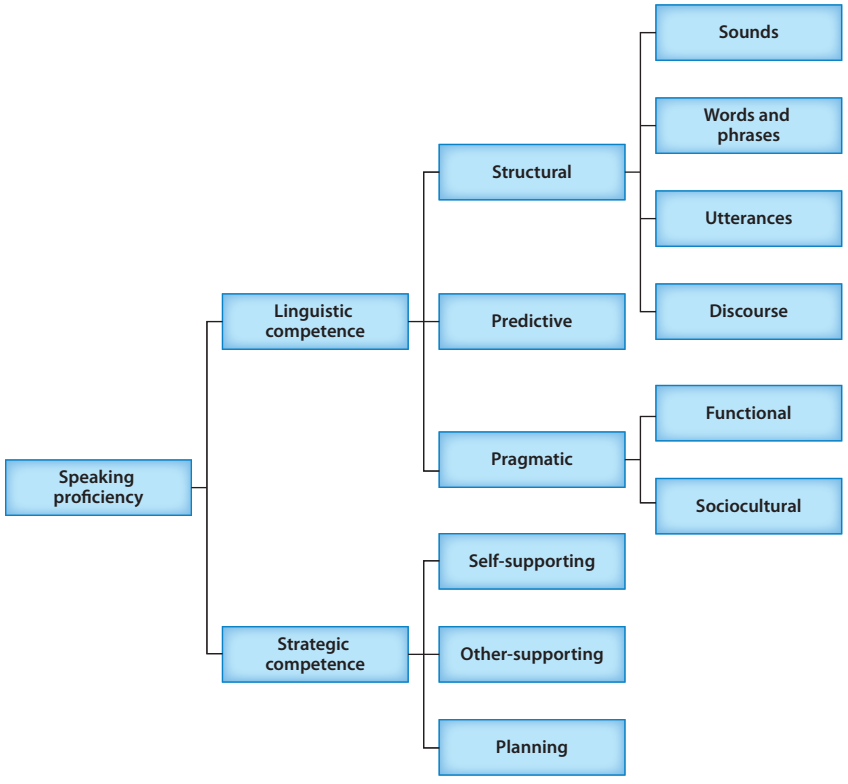

Figure 1

Speaking proficiency components. Figure inspired by Bachman & Palmer (1996, 2010).

but explicitly adding emphasis on speed of processing (Hulstijn 2015, Van Moere 2012), as well as the predictive or anticipatory skills crucial for successful interaction (Levinson 2016, Pickering & Garrod 2004). **Figure 1** depicts this version of speaking proficiency, which in part is also empirically supported (e.g., Bachman & Palmer 1982, De Jong et al. 2012b, Jeon et al. 2022, Nakatani 2010). The two main components are strategic and linguistic competencies for speaking. The term competence instead of knowledge for these components is chosen to include both knowledge and speed of using that knowledge under the high time constraints of speaking.

Roever & Kasper (2018) argue that interactional competence and linguistic competence are separate kinds of competencies. However, because recipient design, knowledge of discourse structures, and predictive competence also play a role in monologues, and because it is impossible to draw a line between dialogue and monologue (some monologues have more interaction than others, etc.) in the proposed model, interactional competence is not a separate component within speaking proficiency. Instead, in the proposed model, linguistic predictive competence is added within linguistic competence. In the next two sections, I further describe linguistic and strategic competence as components of L2 speaking proficiency.

2.4.1. Linguistic competence for speaking. Following Bachman & Palmer (1996, 2010), linguistic competence entails both structural and pragmatic components. Following new insights (e.g., Pickering & Garrod 2013), predictive or anticipatory competence is added as a component of linguistic competence.

The structural part encompasses the traditional linguistic range from phonology (sounds) to coherence in oral texts (discourse). A speaker needs to know how to intelligibly pronounce the sounds of the L2, which may be different from the sounds in their L1 inventory. Knowledge and processing speed of words, as well as phrases, are depicted as the second component. Phrases or chunks are a necessary component of knowledge because native speakers generally speak in predictable and conventional utterances (Ellis 2001), and therefore, if there is a conventional phrase that is used for a specific message, an L2 speaker will be harder to understand if they use a different (newly generated and grammatically correct) phrase (Pawley & Syder 1983, Pickering & Garrod 2013). In addition to being more readily understood, using lexicalized phrases and chunks helps a speaker's fluency because retrieving ready-made constructions from memory takes less computational power than building the sentence from scratch (Skehan 1998, p. 32).

Even if many spoken utterances tend to be part repetitions, uncreative, and construction-based, a speaker needs to be able to create new utterances, be it through the application of morphosyntactic rules (e.g., Kempen & Hoenkamp 1987) or through association in construction grammar or exemplars (Bybee 2013, Goldberg 2005). Concerning syntax or lexicogrammar specific for speaking, a successful speaker may use utterances that are not complete sentences, such as (elliptical) single-word utterances.

The organizational or structural component with the largest units is labeled discourse and includes the knowledge and skills to apply this knowledge of how oral texts are organized, including conversational structure. For instance, as described above, conversations follow specific rules for when a new topic is introduced, how adjacency pairs are sequentially organized, and how turn-taking is organized.

The pragmatic component within the definition of linguistic competence for speaking is based on speech act theory (Searle 1976), politeness theory (Brown & Levinson 1987), and Grice's (1975) theory of conversational implicature (Timpe-Laughlin & Youn 2020). The functional component, also called pragmalinguistics, refers to the linguistic means necessary to achieve communicative functions. Sociolinguistic knowledge or sociopragmatics refers to the competence to understand and correctly use the social contextual norms and conventions. For instance, speakers need to learn that a request is politely made by literally asking for information ("Do you have more coffee?") and need to recognize the appropriate sociocultural context and corresponding norms that further define the appropriate linguistic forms ("Would you be so kind as to. . ." or "Can you. . ."). As described by Timpe-Laughlin & Youn (2020), Kasper & Ross (2013), among others, have broadened the definition of pragmatic competence to include aspects of competence in organizing interaction. However, in **Figure 1**, these aspects are labeled under the discourse part of structural, linguistic competence, as described in this section.

The predictive competence is not further described here, because currently, it is not clear how speakers achieve the rapid speech-act prediction or recognition crucial for rapid responses in conversation (Levinson 2016, but see Huettig et al. 2022).

2.4.2. Strategic competence for speaking. Strategic competence for speaking consists of both verbal and nonverbal communication strategies (Dörnyei & Scott 1997) to compensate for breakdowns in communication. These can be broadly categorized as self-supporting and other-supporting strategies (Van Batenburg et al. 2018). For instance, if a learner struggles to find the correct lexical item or phrase, a speaker can choose to abandon the message entirely (which would not lead to communicative success), to simplify the intended message, or to use a word such as "thingamabob" as a substitute. Another strategy is needed when learners find themselves short of processing time and need to buy time while speaking, for instance, by repeating a phrase or using "uhm." These self-supporting strategies are directly related to the psycholinguistic perspective of speaking as an individual skill (Dörnyei & Kormos 1998, Kormos 2006). In addition to such

self-supporting strategies, other-supporting strategies are also needed when successful interaction between speakers is considered: Speakers need to be able to listen to their interlocutor with attention; to align their messages to the interlocutor's needs; and to aptly respond to clarification requests, indications of incomprehension, and misunderstanding of the message (Bygate 1987).

Bachman & Palmer (2010), in their model, mention three general metacognitive skills applicable to all language abilities, as well as in broader realms of cognition, summarized as planning in **Figure 1**. The first is goal setting, which involves deciding what one is going to do; the second is assessment, involving evaluation of the situation and one's abilities to cope with the situation; the third is planning and involves deciding how to use what one can. All mentioned metacognitive and strategic skills need to be known (knowledge) and put into action quickly (processing) under the high time constraints of natural speech, hence the term strategic competence.

To summarize, a proficient speaker knows how to react aptly, efficiently, and intelligibly to a communicative task in different sociocultural settings, managing problems that may naturally occur during communication. A task can be any task in which spoken production is involved, among which may be asking for information at a store, requesting help from friends to move house, retelling facts from a news article during a conversation with coworkers, or explaining graphs at a finance meeting. This means that the above-described linguistic and strategic competencies need to be put to use simultaneously for any communicative task at hand.

3. FACTORS IN SPEAKING ASSESSMENT

Before describing the different factors and processes involved in speaking assessment (**Figure 2**), I briefly describe the two main purposes of assessment. At the end of a lesson series, a teacher may want to determine the extent to which their students met the intended learning outcomes of those lessons and give an assessment that counts toward the students' final grades. Or, at the end of the second lesson of that series, a teacher may want to find out what students did and did not understand so that in lesson three the teacher can best meet students' learning needs. In the first case, the purpose of the assessment is summative, and in the second example, the

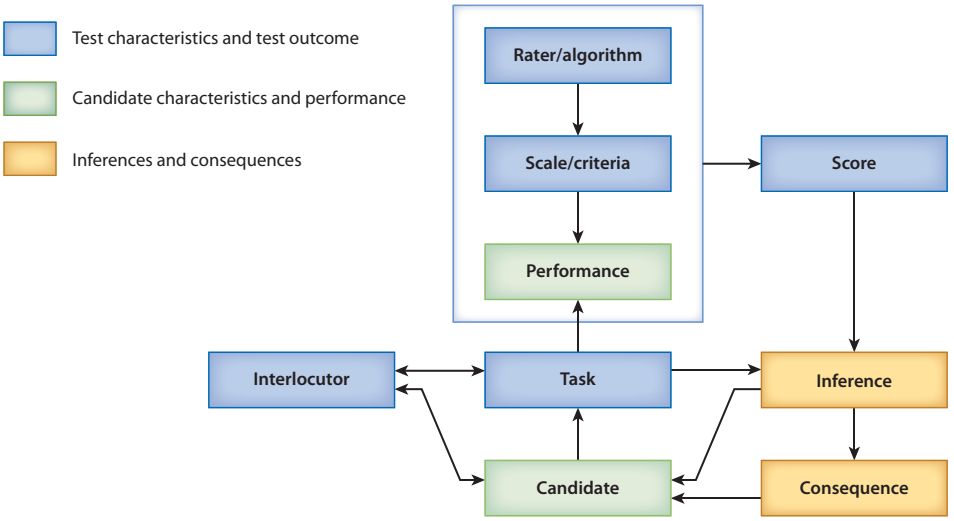

Figure 2

Processes and factors in the assessment of speaking. Figure adapted with permission from McNamara (1996, p. 86).

Table 1 Main differences between formative and summative assessment

Characteristic	Formative	Summative
Timing	During learning process	After learning process is completed
Goal	To guide and/or improve the learning process	To make a final judgment
Consequence	Decision about future didactic approach, to stimulate the learning process	Decision about the student

assessment activity has a formative function. There are three main distinctions between formative and summative assessment concerning the timing of the assessment, the goal of the assessment, and the consequences of the assessment (see **Table 1**). The consequences of some types of summative assessments can be far-reaching. For instance, failing a national language exam may mean that a candidate does not get a visa to stay in the respective country.

These high-stakes consequences for summative assessments lead to high demands on the quality of the speaking test (see the sidebar titled Quality Requirements in Language Assessment). One of the most important quality requirements is validity. Throughout all factors in the assessment process involved, there must be evidence for validity. Borsboom et al. (2004) argue that a test is valid if variations in the ability cause variations in the test outcome. However, because inferences and consequences follow from a score, it has been argued that the use of the score must be valid (Bachman & Palmer 2010, Kane 2013, Messick 1989). According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME 2014, p. 14), validity is thus defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test scores." In **Figure 2** (based on McNamara 1996), these two divergent views are depicted as two colors: If all factors and processes up to inferences (blue components in **Figure 2**) can be shown to reflect the proposed construct, the validity of the speaking assessment "stripped of all excess baggage" (Borsboom et al. 2004, p. 1070) is satisfactory; i.e., the score reflects the intended (speaking) construct. However, for a validity argument according to Kane (2013), it must be shown that the use of a test, including the inferences and consequences (orange components in **Figure 2**), leads to the anticipated beneficial effects and no (unanticipated) detrimental effects.

To reach validity, the relevant variation between performances by test takers should be variation in their abilities on the chosen construct of speaking proficiency only. This chosen construct leads to designing specific tasks, choosing specific ways to deliver the task (e.g., with or without an interlocutor present), and creating specific rubrics or criteria that are to be used by human raters or algorithms that calculate scores automatically. As can be understood from this simplified picture of the process of assessment and the factors involved (a more elaborate model can be found in Fulcher 2003), construct-irrelevant variance can occur at many different stages in this process. What if the task benefits some test takers more than others, for example, because of differences in topic knowledge? What if characteristics of the interlocutor (e.g., among many things, their gender) benefit some test takers more than others? Did the tasks lead to speaking performances that reflect the speaking ability of the candidate best? Do the criteria reflect the construct? Was the rater or the automatic calculation by an algorithm able to evaluate the criteria as they were intended? Did the rater use other criteria that are not relevant to the chosen construct of speaking? In other words, does the final score reflect the claimed ability of the candidate?

3.1. Speaking Assessment Tasks

Assessing speaking proficiency using tasks that sample all possible communicative functions in which speaking is involved is not feasible for the practice of language assessment, and fortunately,

QUALITY REQUIREMENTS IN LANGUAGE ASSESSMENT

Valid

A test is valid when it measures what it intends to measure, i.e., when variations in scores reflect variations in the intended construct (Borsboom et al. 2004). Test use is valid when evidence and theory support the interpretations of the test (e.g., Kane 2013).

Reliable

A test is reliable when it measures the construct it measures precisely, that is, with the smallest possible margin of error. Reliable assessment can be achieved only through repeated evaluations (i.e., multiple tasks or assignments for each participant). In addition, interlocutor and rater effects should be minimized.

Transparent

A speaking assessment is transparent when the teacher or testing agency is clear to the various stakeholders (at least the candidates) about what is expected during the assessment, i.e., what the requirements, criteria, and method of assessment will be.

Authentic

In addition to being valid, reliable, and transparent, a communicative speaking assessment activity should also strive to be authentic: The activity must create a situation that resembles the intended communicative situation in real life. This ensures positive washback.

this is often not needed. For assessment of speaking proficiency, the choice of tasks is dependent on the chosen subconstruct within speaking proficiency and on the type of language in the specific context being assessed. A speaking task used in (research on) a prevocational program in which students are prepared for working in hospitality, for example, could be "to explain to a customer how to open the door using a hotel key card," in which the test taker roleplays the hotel clerk and the interlocutor roleplays the customer (Van Batenburg et al. 2018). An opposite distribution of roles can be foreseen in a speaking task when it is likely that the test taker in an authentic situation would need to know how to ask for help from a hotel clerk when faced with a convoluted key-card procedure. The chosen tasks thus need to simulate the target language-use domain (Bachman & Palmer 2010).

Progress in speaking proficiency by a learner entails not only increasing linguistic and strategic competencies but also the ability to operate these competencies in more diverse contexts and on more diverse topics. A beginning language user will be able to talk about some concrete aspects of their personal life. A highly proficient speaker, however, should be able to handle many different situations and be able to talk about many topics. So, if gauging overall speaking proficiency is the goal of the assessment, it becomes increasingly difficult (if not impossible) to sample tasks soundly, especially at the higher levels of proficiency. Hulstijn (2011) therefore argues that core language skills such as linguistic knowledge (vocabulary, grammar, pronunciation) and automaticity (speed of processing) can be reliably and efficiently measured using discrete-point methods, assessing one element of language at a time. This is taken up by Van Moere (2012) who argued for using elicited imitation, a task in which speakers quickly and accurately repeat sentences, to assess automaticity in speaking.

Either way, the language test constructer needs to consider the specific construct and design tasks accordingly. For instance, when interaction is part of the intended construct, using only

monologues as tasks will not capture the full construct (Roever & Ikeda 2022). As another example, Van Batenburg et al. (2018) targeted the measurement of strategic competence in interaction and therefore used scripted role-plays specifically eliciting these strategies.

3.1.1. Constructive alignment. As just discussed, the most important consideration when designing tasks is the specific subconstruct that is to be assessed. When the assessment is classroom based and learning related, the construct to be measured is set by the learning goals and reflects the type of tasks that learners practiced, following the principle of constructive alignment (Biggs 1996). For instance, if the learning goal for a series of lessons were "learning to request information," the teacher would use teaching materials such that students practice requests, and the aligned assessment would predictably gauge the extent to which the students can request information.

3.1.2. Washback effects. If no conscious or purposeful alignment has been created by the teacher, an assessment will still influence the teaching that precedes it. This is because teachers always want to prepare their students as well as possible for the assessment, and students in turn want to learn what will be covered in the assessment. The phenomenon of the influence of an assessment on the teaching and learning that precede it is called washback. Washback can be positive, for example, when the assessment consists of authentic tasks reflecting students' potential future communication outside the classroom. In this case, both teacher and students will want to practice such authentic tasks. Washback can also be negative, and in the case of speaking proficiency, washback has indeed been negative simply because assessment of speaking did not occur. The negative washback then consists of little practice of speaking in the classroom. The lack of speaking assessment in the classroom may be caused by the fact that teachers may find fulfilling the needs outlined in handbooks for valid assessment intimidating, especially because such guidelines are usually geared toward large-scale settings, rather than classroom settings (Fulcher 2012, Vogt & Tsagari 2014).

Although the purposeful teaching principle of constructive alignment does not play a role in (large-scale) external examination, washback does still occur. For instance, implementing speaking in the TOEFL® exam in 2005 led to more focus on teaching speaking in preparation courses (e.g., Wall & Horák 2011). The type of speaking tasks used in external examinations also leads to either positive or negative washback. If only elicited imitation were to be used in a large-scale oral proficiency test, preparation courses would likely teach language learners to become experts in repeating sentences. This would be an example of negative washback.

The challenge for the large-scale external examinations is, on the one hand, to include authentic communicative tasks (from the targeted language-use domain), and on the other hand, to standardize delivery of the tasks across contexts. One such standardized, yet authentic, external examination is the Occupational English Test (OET), an English language screening test for clinicians designed to reflect the language demands of health professional–patient communication. Woodward-Kron & Elder (2016) showed how the OET indeed exhibits a reasonable degree of authenticity. Despite the similarities, their study also showed that the linguistic choices and interactional approaches by the role-played patients in this language assessment setting differed from those in more authentic settings. Similarly, Johnson (2001), among others, has shown how in a general speaking proficiency test like the Oral Proficiency Interview, the test-interview as a task does not elicit natural conversation but elicits so-called test talk (Fulcher 1996), which is a unique genre by itself (Chalhoub-Deville & Fulcher 2003, p. 503). This is unsurprising since the main underlying goal of natural conversation is different from the main goal of speaking in a testing situation.

3.2. Interlocutor

No matter what the task is, a candidate always directs their speaking to a recipient, because speaking is always directed at someone. It may be face-to-face (either online or physical), it may be an AI interlocutor, or it may be a simulated role-play directed at a computer. In the case of a real interlocutor, it may be an examiner asking interview questions or playing a role, for instance, a patient, as in the OET described above. Alternatively, the interlocutor may be another examinee or multiple examinees, as in group assessments. Similar to the choice of tasks, the choice of the interlocutor(s) and their interactional behavior should be driven by the targeted construct. To reach authenticity, interlocutors would need to be an actual patient for the OET, and an actual tourist for the hotel clerk task. However, in addition to being authentic and valid, a speaking assessment must also be precise or reliable. In the case of real interlocutors, it should not matter who is chosen as interlocutor, and even if the same interlocutor is involved for all examinees of a certain test, the interlocutor should give each examinee the same opportunities and/or behave similarly. Not surprisingly, research has shown that this is impossible. Because the theory on communication states that interaction is coconstructed by all interlocutors involved, this theory already predicts that an examiner cannot behave the same in each exam and in each interaction. Additionally, the fact that in successful communication interlocutors align the way they speak (Pickering & Garrod 2004) predicts that the language use of one examinee will depend on the language use of the other in paired assessments, which may influence individual scores.

Indeed, both test takers and examiners introduce variability simply by bringing their sociocultural identity, personal characteristics, and speaking styles to the speaking test. The interlocutor effect (O'Sullivan 2002) is more pronounced in paired and group assessments compared to individual assessments. In these paired or group assessments, the acquaintanceship (O'Sullivan 2002), personality (Berry 2007, Ockey 2009), extroversion (Berry 2007), and language proficiency of the interlocutor (Davis 2009) have been found to have an influence. In addition, interlocutor effects are highly unpredictable (Brown & McNamara 2004), as they may interact in various ways (Berry 2007, Nakatsuhara 2011), including three-way interactions (O'Sullivan 2004). Galaczi & Taylor (2018) describe these and other studies in more detail when emphasizing the tension between authenticity/validity on the one hand and reliability on the other. Because of the nature of interaction, they conclude that the variability should be seen as part of the construct: "Engaging in interaction successfully with different interlocutors is now seen as a fundamental part of real-life interactional demands, and it can therefore be argued that the associated variability is part of the construct underlying communicative language tests" (Galaczi & Taylor 2018, p. 225).

However, for an assessment to be reliable, their conclusion implies that the only way to include variability as part of the assessment itself is to have test takers converse with many different interlocutors, thereby sampling most if not all possible interlocutor effects and interactions between interlocutor effects. If a test taker is paired with only one interlocutor, the score on a speaking assessment for a candidate will be dependent on either fortunate or unfortunate pairing. Galaczi & Taylor (2018) therefore propose to combine different types of interactional formats, including monologue, alongside paired/group discussion. This solution would indeed diminish the variability induced by random pairing with other examinees but not get rid of it entirely. Similarly, awarding shared scores to examinees in paired or group assessments, as suggested by May (2009) and Taylor & Wigglesworth (2009), while truly admitting the shared responsibilities in the coconstructed interaction of all interlocutors involved, would make the score unreliable for each individual in the assessment. Another approach is to use a more scripted dialogue with a real interlocutor or to use Spoken Dialog Systems (SDSs). Ockey & Chukharev-Hudilainen (2021), for

instance, investigated how scores in human–human interaction may differ from scores awarded in the SDS interaction and found raters to award similar scores for several criteria, such as grammar and vocabulary, but not for interactional competence.

3.3. Criteria

According to Weigle (2002), the scoring criteria of a test represent the theoretical basis on which it is founded and are thereby the embodiment of a test's construct. The speaking assessment criteria thus need to consider the many aspects that L2 learners need to master to become proficient (see **Figure 1**). Indeed, judges usually score multiple components to evaluate speaking proficiency, following Bachman & Palmer's (1996, 2010) construct of language ability. One of the most influential descriptions of scales and criteria is the Common European Framework of Reference (CEFR) (Council of Europe 2001). The general framework is language and context independent and based on models of communicative competence (Bachman & Palmer 1996, Canale & Swain 1980). The descriptors in the scales are based on 40 existing proficiency scales (North 2002), which means that they are based on "professional consensus rather than empirical evidence" (Harsch 2014, p. 156). One of the illustrative scales involving oral production is the scale depicted in table 3 of the CEFR (Council of Europe 2001, pp. 28–29) describing "Qualitative aspects of spoken language use." This scale has the following linguistic criteria: range, accuracy, fluency, interaction, and coherence. Because the scale criteria were based on existing scales and expert opinions, and because the CEFR has been highly influential in language assessment since 2001, these criteria can be seen as a summary of all (qualitative) criteria in existing general speaking proficiency scales.

Differences between assessment scales, however, do exist. For instance, De Jong (2018) describes the differences between the International English Language Testing System (IELTS), Pearson Test of English (PTE) Academic, and TOEFL® internet-based test (iBT) for the construct of fluency: Fluency is apparent as a separate criterion in the PTE Academic (Pearson 2022) but part of the criterion "Fluency and coherence" in the IELTS scale (IELTS 2022). In addition to criteria on the quality of language, most scales also include an aspect of content. In a brief comparison of these and additional scales on this matter, De Jong (2021) likewise described differences between scales. Whereas the PTE Academic (Pearson 2022) and TOEFL® iBT (ETS 2022) have separate criteria for content (labeled "Content" and "Topic development," respectively), in the IELTS descriptors (IELTS 2022), content is mentioned under "Fluency and coherence," and only at the two highest levels (band 8: "develops topics coherently and appropriately" and band 9: "develops topics fully and appropriately"). Like the descriptions for fluency (De Jong 2018), the descriptions for content are sometimes vague and leave room for subjectivity.

The criteria in language assessment largely overlap with the aspects that are used in L2 acquisition research, when researchers measure aspects of speech objectively. The linguistic aspects of complexity, accuracy, and fluency (CAF) in L2 acquisition research have been researched extensively. Recently, the notion of functional adequacy (i.e., content or task fulfillment) has been added (e.g., De Jong et al. 2012a, Kuiken & Vedder 2022, Pallotti 2009). As Pallotti (2020) mentions, the measures of CAF may be more appropriate for formative rather than summative assessment (p. 207). The precise analytic measures or scores in a rubric help L2 learners to identify which aspects of L2 speaking they have mastered and on which aspects they can still improve (Bachman & Damböck 2018, p. 154). The information provided by a filled-out analytic rating scale also helps teachers to be aptly adaptive to the individual needs of learners (Hughes 2003, p. 105). In this way, analytic rating scales are useful in learning and teaching processes (assessment for learning, see Black & Wiliam 2009).

3.4. Scoring

Until recently, most scores on speaking assessments were given by raters, using scoring criteria as described in Section 3.3. Human raters bring another challenge for reliability. The score should be the same, irrespective of who rated the speaking performance. In addition to the interlocutor effect, however, rater effects have been identified in language assessment contexts (Eckes 2009, McNamara 1996). Raters may be systematically lenient or severe, or otherwise systematically use the scale divergently, for instance, by showing a central tendency and avoiding both ends of the scale or by showing a systematic bias. From research on rater cognition (Han 2016), it is also well-known that judges have trouble keeping aspects apart (Engelhard 1994). This is the common understanding of a halo effect, where raters "fail to distinguish between conceptually distinct features of examinee performance, but rather provide highly similar ratings across those features" (Eckes 2009, p. 5). For instance, when judging fluency, an already formed opinion on accuracy may influence the rating. Cai (2015) showed how raters confuse criteria while assigning scores, due to high cognitive load. Using specific assessment criteria for specific parts within a test (Khabbazbashi & Galaczi 2020, Taylor & Galaczi 2011) may be one option to reduce such cognitive load.

Some of the bias effects can be directly related to rater characteristics that have been found to influence test scores, such as the raters' accent familiarity (Winke et al. 2013), the rater's native speaker status, and social bias (Kang et al. 2019). Kang et al. (2019) also showed how a short rater training mitigated these effects, at least in the short run. Additionally, to maximally reduce the risk of rater effects, multiple raters should be used: Xi & Mollaun (2006) examined the reliability of the TOEFL® Academic Speaking Test and found that averages across six tasks and two ratings would be acceptable for sufficient reliability when using analytic scores. This reflects the finding from a review study by In'nami & Koizumi (2016) that task effects are larger than rater effects.

In addition to being difficult, rating is also time-consuming and resource intensive. In large-scale testing, this is avoided by introducing automatically obtained measures and, as proposed by Isaacs (2018), the combination of human ratings on certain aspects with automatic measures on other aspects has been investigated (Davis & Papageorgiou 2021). This leads to more efficient scoring procedures for large-scale tests. For classroom assessment, however, automatically measuring speech is currently not feasible. Gu et al. (2021) showed that automated analytic feedback as calculated by the SpeechRater™ engine of the TOEFL® iBT speaking test can be helpful for learners preparing for such a test.

3.5. Inferences and Consequences

When a score has been awarded to the test taker's performance by an algorithm or a rater, or by a combination of the two, the score still needs to be assigned a meaning. An inference is made, based on the score and based on the requirements of the task. For instance, if the task is a monologue, little if any inference on interactional competence can be made. Or if the task was to give answers in an interview, little if any inference can be made on competencies for orally summarizing a newspaper article. The score on the assessment can be given to the examinee, but the inference that follows from the score also needs to be made clear to the examinee. Especially in formative assessment, what the score means, in combination with feedback to the learner, helps the learning process, as briefly described in Section 3.4. Which components of speaking proficiency has the candidate already mastered, and which components can still develop? Next, for both formative and summative assessments, consequences follow (see **Table 1**).

5. CONCLUSION AND FUTURE CHALLENGES

This section describes some future challenges for the assessment of speaking. The existing challenges have become clear in the previous sections: Assessing speaking is a difficult task. Before the assessment can take place, the test constructor needs to define the construct and accompanying language use domain. The factors needed to assign a score, such as tasks, interlocutors, rating scales, and raters or algorithms, should introduce little if any construct-irrelevant variance. For most acts of speaking, the ability to successfully communicate, i.e., the construct-relevant variance, does not necessarily entail using complex language without any disfluencies (see also De Jong 2018); rather, aligned and predictable speech drives efficient communication.

5.1. Plurilingual Competencies

In describing speaking assessment thus far, I have not taken into account recent developments in language learning that acknowledge the fact that learners may successfully use more than one language in communication [see, e.g., the new companion volume to the CEFR (Council of Europe 2018)]. When involved in a communicative task, speakers may need to draw upon their plurilingual competencies. For instance, a communicative task may involve describing a graph at a finance meeting in English, followed by a discussion in English plus another language. In addition to using sociocultural competencies to gauge which register is needed in a specific context, a successful speaker must be able to switch between dialects and languages appropriately. The assessment of such plurilingual competencies has not made its way into actual speaking assessments yet.

5.2. Classroom Assessment Development

A practical challenge in the research and development of speaking assessment is to make the innovative and sometimes technological advancements that are researched and introduced available for classroom-based assessment. Assessing speaking remains a difficult task, but perhaps language teachers would be helped if practical versions of some of the technical tools that have been developed became available. For instance, using an SDS (Ockey & Chukharev-Hudilainen 2021) could be helpful for both learning and assessment. Likewise, automatic scoring systems, including meaningful feedback to learners (Gu et al. 2021), could make classroom assessments less resource- and time-consuming. The recent need to administer language assessments online, due to the coronavirus disease 2019 (COVID-19) pandemic, could potentially further drive advancements that could leak into classroom (online) assessment. Isbell & Kremmel (2020) note that online assessments "potentially allow for more interactivity, multimodality, and more authentic representations of today's communication and target language use domains than has been possible in traditional paper-based tests" (p. 616).

5.3. Individual Differences in L1 Speaking

It is well known that L2 learners exhibit large individual differences (IDs) in success at learning to speak an L2. Therefore, an important aim in L2 acquisition research has been to explain such IDs. Concerning oral abilities, research has established that L2 speakers differ in the complexity, accuracy, and fluency with which the message is delivered (Pallotti 2020) and that (skills in) these linguistic aspects of speech are related to L2 proficiency (De Jong et al. 2012b, Iwashita et al. 2008). Concerning speaking fluency, however, it has been found that not all aspects are related to underlying L2 knowledge and skills (De Jong et al. 2013, Kahng 2020) and that strong correlations between L1 and L2 fluency exist (Bradlow et al. 2017, De Jong et al. 2015, Kahng 2020). In other words, speakers have individual speaking styles, which carry over from their L1 behavior.

Whereas research into L2 learning and assessment has usually focused on IDs thus far, there has been little research into L1 IDs in oral abilities and aspects of speaking. There is some evidence that L1 speakers show IDs in their use of linguistic aspects, not just concerning fluency as mentioned above but also concerning (grammatical) accuracy and complexity (Dąbrowska 2012). These findings raise an important question. If aspects of L2 speaking are partly unrelated to L2 proficiency but are due to personal speaking style, we need to investigate how such IDs in speaking styles emerge in both L2 and L1. Not only do we need to better understand L1 and L2 speaking processes by incorporating individual variations in a meaningful way (see also Kidd et al. 2018), we also need to understand which aspects of L2 speaking are relevant for L2 speaking proficiency assessment and which reflect construct-irrelevant personal speaking styles.

5.4. Assessing Interactional Competence Through Controlled Interaction

In a way, this plea for research on L1 IDs can be seen as a research agenda for psycholinguistics raised by L2 researchers (Hulstijn 2019). Conversely, L2 research and assessment must learn from psycholinguistic studies into dialogues. As Pickering & Garrod (2004) discuss, psycholinguistic research used to shy away from investigating speaking at all, because of uncontrolled spontaneous responses (Bock 1996). And when controlled speaking experiments were introduced, this new research shied away from investigating dialogue for the same fear of lack of control. Branigan et al. (2000), however, showed how the use of a confederate in controlled, dialogic experiments, has been successful in studying dialogue, where participants are unaware that the confederate is working (partly) from a script. Language assessment practice may learn from these innovations to use dialogue in a controlled manner (cf. Van Batenburg et al. 2018). Such formats may lead to more controlled and thus reliable inclusion of interaction in assessment while still upholding authenticity. Leaving out interaction from most targeted constructs in speaking assessment cannot be an option, interaction being the prime act in which all the components of speaking proficiency are elicited, including strategic competencies, as well as all three linguistic competencies: structural, predictive, and pragmatic.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

AERA, APA, NCME. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: Am. Educ. Res. Assoc.

Bachman L, Damböck B. 2018. *Language Assessment for Classroom Teachers*. Oxford, UK: Oxford Univ. Press

Bachman LF, Palmer AS. 1982. The construct validation of some components of communicative proficiency. *TESOL Q.* 16(4):449–65

Bachman LF, Palmer AS. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford, UK: Oxford Univ. Press

Bachman LF, Palmer AS. 2010. *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford, UK: Oxford Univ. Press

Bates E, D'Amico S, Jacobsen T, Székely A, Andonova E, et al. 2003. Timed picture naming in seven languages. *Psychon. Bull. Rev.* 10(2):344–80

Berry V. 2007. *Personality Differences and Oral Test Performance*. Berlin: Peter Lang

Biggs J. 1996. Enhancing teaching through constructive alignment. *High. Educ.* 32(3):347–64

Black P, Wiliam D. 2009. Developing the theory of formative assessment. *Educ. Assess. Eval. Account.* 21(1):5–31

Bock K. 1996. Language production: methods and methodologies. *Psychon. Bull. Rev.* 3(4):395–421
Borsboom D, Mellenbergh GJ, van Heerden J. 2004. The concept of validity. *Psychol. Rev.* 111(4):1061–71
Bradlow AR, Kim M, Blasingame M. 2017. Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *J. Acoustical Soc. Am.* 141(2):886–99
Branigan HP, Pickering MJ, Cleland AA. 2000. Syntactic co-ordination in dialogue. *Cognition* 75(2):B13–25
Brown A, McNamara T. 2004. "The devil is in the detail": researching gender issues in language assessment. *TESOL Q.* 38(3):524–38
Brown P, Levinson SC. 1987. *Politeness: Some Universals in Language Usage*. Cambridge, UK: Cambridge Univ. Press
Bybee JL. 2013. Usage-based theory and exemplar representations of constructions. In *The Oxford Handbook of Construction Grammar*, Vol. 1, ed. T Hoffmann, G Trousdale, pp. 1–24. Oxford, UK: Oxford Univ. Press
Bygate M. 1987. *Speaking*. Oxford, UK: Oxford Univ. Press
Cai H. 2015. Weight-based classification of raters and rater cognition in an EFL speaking test. *Lang. Assess. Q.* 12(3):262–82
Canale M, Swain M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Appl. Linguist.* 1(1):1–47
Celce-Murcia M. 2007. Rethinking the role of communicative competence in language teaching. In *Intercultural Language Use and Language Learning*, ed. EA Soler, MPS Jordà, pp. 41–57. Dordrecht, Neth.: Springer Netherlands
Chalhoub-Deville M, Deville C. 2005. A look back at and forward to what language testers measure. In *Handbook of Research in Second Language Teaching and Learning*, ed. E Hinkel, pp. 815–32. Abingdon, UK: Routledge. 1st ed.
Chalhoub-Deville M, Fulcher G. 2003. The oral proficiency interview: a research agenda. *Foreign Lang. Ann.* 36(4):498–506
Clark HH. 2002. Speaking in time. *Speech Commun.* 36(1–2):5–13
Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg, Fr.: Council of Europe
Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg, Fr.: Council of Europe
Dąbrowska E. 2012. Different speakers, different grammars: individual differences in native language attainment. *Linguist. Approaches Bilingualism* 2(3):219–53
Davis L. 2009. The influence of interlocutor proficiency in a paired oral assessment. *Lang. Test.* 26(3):367–96
Davis L, Papageorgiou S. 2021. Complementary strengths? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assess. Educ.: Principles Policy Pract.* 28(4):437–55
De Bot K. 1992. A bilingual production model: Levelt's 'speaking' model adapted. *Appl. Linguist.* 13(1):1–24
De Jong N, Steinel MP, Arjen F, Florijn AF, Schoonen R, Hulstijn JH. 2012a. The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In *Dimensions of L2 Performance and Proficiency*, ed. A Housen, F Kuiken, I Vedder, pp. 121–42. Amsterdam: John Benjamins Publ. Co.
De Jong NH. 2018. Fluency in second language testing: insights from different disciplines. *Lang. Assess. Q.* 15(3):237–54
De Jong NH. 2021. *Assessing language when content matters: language assessment viewpoint*. Paper presented at EALTA Speaking SIG: Assessing Content When Language Matters, online, Nov. 19
De Jong NH, Groenhout R, Schoonen R, Hulstijn JH. 2015. Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Appl. Psycholinguist.* 36(2):223–43
De Jong NH, Steinel MP, Florijn AF, Schoonen R, Hulstijn JH. 2012b. Facets of speaking proficiency. *Stud. Second Lang. Acquis.* 34(1):5–34

De Jong NH, Steinel MP, Florijn A, Schoonen R, Hulstijn JH. 2013. Linguistic skills and speaking fluency in a second language. *Appl. Psycholinguist.* 34(5):893–916

Dell GS, Schwartz MF, Martin N, Saffran EM, Gagnon DA. 1997. Lexical access in aphasic and nonaphasic speakers. *Psychol. Rev.* 104(4):801–38

Dingemanse M, Roberts SG, Baranova J, Blythe J, Drew P, et al. 2015. Universal principles in the repair of communication problems. *PLOS ONE* 10(9):e0136100

Dörnyei Z, Kormos J. 1998. Problem-solving mechanisms in L2 communication: a psycholinguistic perspective. *Stud. Second Lang. Acquis.* 20(3):349–85

Dörnyei Z, Scott ML. 1997. Communication strategies in a second language: definitions and taxonomies. *Lang. Learn.* 47(1):173–210

Drew P. 2012. Turn design. In *The Handbook of Conversation Analysis*, ed. J Sidnell, T Stivers, pp. 131–49. Hoboken, NJ: Wiley. 1st ed.

Eckes T. 2009. Many-facet Rasch measurement. In *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, ed. S Takala, Section H. Strasbourg, Fr.: Council of Europe

Ellis NC. 2001. Memory for language. In *Cognition and Second Language Instruction*, ed. P Robinson, pp. 33–68. Cambridge, UK: Cambridge Univ. Press. 1st ed.

Engelhard G. 1994. Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *J. Educ. Meas.* 31(2):93–112

ETS (Educ. Test. Serv.). 2022. *TOEFL iBT® Independent Speaking Rubrics*. Rubric, Educ. Test. Serv., Princeton, NJ. **https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf**

Fan J, Yan X. 2020. Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Front. Psychol.* 11:330

Fulcher G. 1996. Testing tasks: issues in task design and the group oral. *Lang. Test.* 13(1):23–51

Fulcher G. 2003. *Testing Second Language Speaking*. Harlow, UK: Pearson Longman

Fulcher G. 2012. Assessment literacy for the language classroom. *Lang. Assess. Q.* 9(2):113–32

Fulcher G. 2015. Assessing second language speaking. *Lang. Teach.* 48(2):198–216

Galaczi E, Taylor L. 2018. Interactional competence: conceptualisations, operationalisations, and outstanding questions. *Lang. Assess. Q.* 15(3):219–36

Goh CCM, Burns A. 2012. *Teaching Speaking: A Holistic Approach*. New York: Cambridge Univ. Press

Goldberg A. 2005. *Constructions at Work*. Oxford, UK: Oxford Univ. Press

Grice HP. 1975. Logic and conversation. In *Speech Acts*, ed. P Cole, JL Morgan, pp. 41–58. Leiden, Neth.: Brill

Griffin ZM, Bock K. 2000. What the eyes say about speaking. *Psychol. Sci.* 11(4):274–79

Gu L, Davis L, Tao J, Zechner K. 2021. Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: a user perception study. *Assess. Educ.: Principles Policy Pract.* 28(1):58–76

Han Q. 2016. Rater cognition in L2 speaking assessment: a review of the literature. *Working Papers TESOL Appl. Linguist.* 16(1):1–24

Harsch C. 2014. General language proficiency revisited: current and future issues. *Lang. Assess. Q.* 11(2):152–69

Hickok G. 2012. Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* 13:135–45

Huettig F, Audring J, Jackendoff R. 2022. A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition* 224:105050

Hughes A. 2003. *Testing for Language Teachers*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.

Hulstijn JH. 2011. Language proficiency in native and nonnative speakers: an agenda for research and suggestions for second-language assessment. *Lang. Assess. Q.* 8(3):229–49

Hulstijn JH. 2015. *Language Proficiency in Native and Non-Native Speakers: Theory and Research*. Amsterdam: John Benjamins

Hulstijn JH. 2019. An individual-differences framework for comparing nonnative with native speakers: perspectives from BLC theory. *Lang. Learn.* 69:157–83

Hymes D. 1972. On communicative competence. In *Sociolinguistics*, ed. J Pride, J Holmes, pp. 263–93. Harmondsworth, UK: Penguin

IELTS. 2022. *Speaking: Band Descriptors*. Rubric, IELTS, London. **https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx**

In'nami Y, Koizumi R. 2016. Task and rater effects in L2 speaking and writing: a synthesis of generalizability studies. *Lang. Test.* 33(3):341–66

Isaacs T. 2016. Assessing speaking. In *Handbook of Second Language Assessment*, ed. D Tsagari, J Banerjee, pp. 131–46. Berlin: DeGruyter Mouton

Isaacs T. 2018. Shifting sands in second language pronunciation teaching and assessment research and practice. *Lang. Assess. Q.* 15(3):273–93

Isbell DR, Kremmel B. 2020. Test review: current options in at-home language proficiency tests for making high-stakes decisions. *Lang. Test.* 37(4):600–19

Iwashita N, Brown A, McNamara T, O'Hagan S. 2008. Assessed levels of second language speaking proficiency: how distinct? *Appl. Linguist.* 29(1):24–49

Jeon EH, In'nami Y, Koizumi R. 2022. L2 speaking and its external correlates: a meta-analysis. In *Understanding L2 Proficiency*, ed. EH Jeon, Y In'nami, pp. 339–67. Amsterdam: John Benjamins

Johnson M. 2001. *The Art of Non-Conversation: A Reexamination of the Validity of the Oral Proficiency Interview*. New Haven, CT: Yale Univ. Press

Kahng J. 2020. Explaining second language utterance fluency: contribution of cognitive fluency and first language utterance fluency. *Appl. Psycholinguist.* 41(2):457–80

Kane MT. 2013. Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50(1):1–73

Kang O, Rubin D, Kermad A. 2019. The effect of training and rater differences on oral proficiency assessment. *Lang. Test.* 36(4):481–504

Kasper G, Ross SJ. 2013. Assessing second language pragmatics: an overview and introductions. In *Assessing Second Language Pragmatics*, ed. SJ Ross, G Kasper, pp. 1–40. London: Palgrave Macmillan UK

Kempen G, Hoenkamp E. 1987. An incremental procedural grammar for sentence formulation. *Cogn. Sci.* 11(2):201–58

Khabbazbashi N, Galaczi ED. 2020. A comparison of holistic, analytic, and part marking models in speaking assessment. *Lang. Test.* 37(3):333–60

Kidd E, Donnelly S, Christiansen MH. 2018. Individual differences in language acquisition and processing. *Trends Cogn. Sci.* 22(2):154–69

Kormos J. 2006. *Speech Production and Second Language Acquisition*. New York: Routledge

Kramsch C. 1986. From language proficiency to interactional competence. *Mod. Lang. J.* 70(4):366–72

Kuiken F, Vedder I. 2022. Measurement of functional adequacy in different learning contexts: rationale, key issues, and future perspectives. *TASK* 2(1):8–32

Levelt WJM, Roelofs A, Meyer AS. 1999. A theory of lexical access in speech production. *Behav. Brain Sci.* 22(1):1–38

Levinson SC. 2016. Turn-taking in human communication – origins and implications for language processing. *Trends Cogn. Sci.* 20(1):6–14

May L. 2009. Co-constructed interaction in a paired speaking test: the rater's perspective. *Lang. Test.* 26(3):397–421

McNamara TF. 1996. *Measuring Second Language Performance*. London: Longman

Messick S. 1989. Validity. In *Educational Measurement*, ed. RL Linn, pp. 13–103. Washington, DC: Am. Counc. Educ. 3rd ed.

Morsella E, Miozzo M. 2002. Evidence for a cascade model of lexical access in speech production. *J. Exp. Psychol. Learn. Mem. Cogn.* 28(3):555–63

Nakatani Y. 2010. Identifying strategies that facilitate EFL learners' oral communication: a classroom study using multiple data collection procedures. *Mod. Lang. J.* 94(1):116–36

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

North B. 2002. Developing descriptor scales of language proficiency for the CEF common reference levels. In *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies*, ed. JC Alderson, pp. 87–105. Strasbourg, Fr.: Council of Europe

Ockey GJ. 2009. The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Lang. Test.* 26(2):161–86

Ockey GJ, Chukharev-Hudilainen E. 2021. Human versus computer partner in the paired oral discussion test. *Appl. Linguist.* 42(5):924–44

O'Sullivan B. 2002. Learner acquaintanceship and oral proficiency test pair-task performance. *Lang. Test.* 19(3):277–95

O'Sullivan B. 2004. Modelling factors affecting oral language test performance: a large-scale empirical study. In *European Language Testing in a Global Context. Studies in Language Testing*, Vol. 18, ed. M Milanovic, CJ Weir, pp. 129–42. Cambridge, UK: Cambridge Univ. Press/Cambridge ESOL

Pallotti G. 2009. CAF: defining, refining and differentiating constructs. *Appl. Linguist.* 30(4):590–601

Pallotti G. 2020. Measuring complexity, accuracy, and fluency (CAF). In *The Routledge Handbook of Second Language Acquisition and Language Testing*, pp. 201–10. Abingdon, UK: Routledge

Pawley A, Syder FH. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In *Language and Communication*, ed. JC Richards, RW Schmidt, pp. 191–226. Abingdon, UK: Routledge

Pearson. 2022. *PTE Academic Institutions Score Guide*. Rep., Pearson, London. **https://assets.ctfassets.net/yqwtwibiobs4/4GzZV6iHiWMfLX1y2CK29l/ef5f0aa73267f157fde173aa499c23d9/PTE_Academic_Score_Guide_for_Institutions_-_June_2022.pdf**

Pekarek Doehler S, Pochon-Berger E. 2015. The development of L2 interactional competence: evidence from turn-taking organization, sequence organization, repair organization and preference organization. In *Usage-Based Perspectives on Second Language Learning*, ed. T Cadierno, SW Eskildsen, pp. 233–68. Berlin: De Gruyter Mouton

Pickering MJ, Garrod S. 2004. Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27(2):169–90

Pickering MJ, Garrod S. 2013. An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36(4):329–47

Plough I, Banerjee J, Iwashita N. 2018. Interactional competence: genie out of the bottle. *Lang. Test.* 35(3):427–45

Roever C, Ikeda N. 2022. What scores from monologic speaking tests can(not) tell us about interactional competence. *Lang. Test.* 39(1):7–29

Roever C, Kasper G. 2018. Speaking in turns and sequences: interactional competence as a target construct in testing speaking. *Lang. Test.* 35(3):331–55

Sacks H, Schegloff EA, Jefferson G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4):696–735

Searle JR. 1976. A classification of illocutionary acts. *Lang. Soc.* 5(1):1–23

Segalowitz N. 2010. *Cognitive Bases of Second Language Fluency*. Abingdon, UK: Routledge

Skehan P. 1998. *A Cognitive Approach to Language Learning*. Oxford, UK: Oxford Univ. Press

Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, et al. 2009. Universals and cultural variation in turn-taking in conversation. *PNAS* 106(26):10587–92

Taylor L, Galaczi ED. 2011. Scoring validity. In *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Vol. 30, pp. 171–233. Cambridge, UK: Cambridge Univ. Press

Taylor L, Wigglesworth G. 2009. Are two heads better than one? Pair work in L2 assessment contexts. *Lang. Test.* 26(3):325–39

Timpe-Laughlin V, Youn SJ. 2020. Measuring L2 pragmatics. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, ed. P Winke, T Brunfaut, pp. 254–64. Abingdon, UK: Routledge

Van Batenburg ESL, Oostdam RJ, Van Gelderen AJS, De Jong NH. 2018. Measuring L2 speakers' interactional ability using interactive speech tasks. *Lang. Test.* 35(1):75–100

Van Moere A. 2012. A psycholinguistic approach to oral language assessment. *Lang. Test.* 29(3):325–44

Vogt K, Tsagari D. 2014. Assessment literacy of foreign language teachers: findings of a European study. *Lang. Assess. Q.* 11(4):374–402

Wall D, Horák T. 2011. The impact of changes in the TOEFL® exam on teaching in a sample of countries in Europe: Phase 3, the role of the coursebook. Phase 4, describing change. *ETS Res. Rep. Ser.* **https://doi.org/10.1002/j.2333-8504.2011.tb02277.x**

Weigle SC. 2002. *Assessing Writing*. Cambridge, UK: Cambridge Univ. Press

Winke P, Gass S, Myford C. 2013. Raters' L2 background as a potential source of bias in rating oral performance. *Lang. Test.* 30(2):231–52

Woodward-Kron R, Elder C. 2016. A comparative discourse study of simulated clinical roleplays in two assessment contexts: validating a specific-purpose language test. *Lang. Test.* 33(2):251–70

Xi X, Mollaun P. 2006. Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Res. Rep. Ser.* **https://doi.org/10.1002/j.2333-8504.2006.tb02013.x**

Young RF. 2011. Interactional competence in language learning, teaching, and testing. In *Handbook of Research in Second Language Teaching and Learning*, Vol. 2, ed. E Hinkel, pp. 426–43. Abingdon, UK: Routledge

Annual Review of Linguistics
Volume 9, 2023

Contents