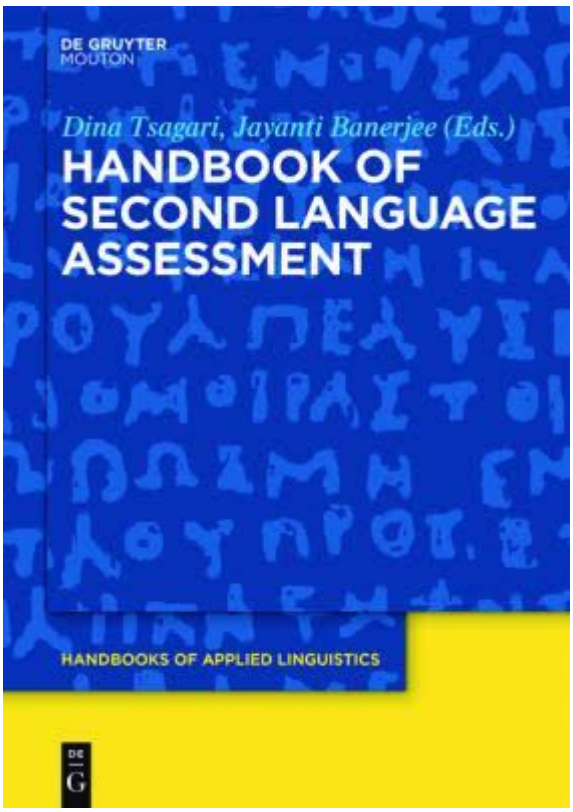
Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). Berlin: DeGruyter Mouton.



Assessing Speaking

Talia Isaacs

1. Introduction

Economic globalization and technological advances are bringing together people from different linguistic and cultural backgrounds who were once oceans apart (Gatbonton and Trofimovich 2008). On a global scale, speaking ability is increasingly, albeit sometimes tacitly acknowledged as a highly coveted skill and a source of cultural capital in many educational and workplace settings today. In the case of learning a second language (L2) in a host country, achieving effective oral communication in the target language is often emphasized as essential for achieving successful integration, removing barriers to academic performance, adequately performing on the job, accessing vital social services, and, on a more macro level, mitigating social isolation and reducing language barriers in linguistically heterogeneous societies (e.g. Derwing and Munro 2009; Isaacs 2013). In addition to L2 interactions occurring with native or native-like members of the target language community, oral communication is also increasingly common in learners from different linguistic communities who use the target language as the lingua franca to

carry out business transactions or to promote cultural exchange, particularly in prestigious or widely spoken languages with international currency (Jenkins 2000). Thus, it would seem that, to buttress and compliment nonverbal communication strategies, the ability to respond to an interlocutor in an appropriate and effective way during the time pressures of real-world face-to-face communication, be it in a dominant language or an L2, are of great human importance.

In light of the current global context, the notion that speaking ability has always been a source of instructional focus in the development of L2 proficiency and in classroom-based and high-stakes assessment instruments might seem intuitive. In fact, this has not always been the case, with speaking ability proving susceptible to paradigmatic swings in approaches to L2 teaching and, in turn, assessment over time (e.g. Weir, Vidaković and Galaczi 2013). While some standardized tests have included spoken production as a mandatory component of L2 proficiency testing for decades (e.g. the *Certificate of Proficiency in English* in 1913), others have incorporated compulsory speaking tasks only relatively recently (e.g. the *Test of English as a Foreign Language* in 2005). Existing commercial tests with an oral proficiency component underscore vastly different approaches to providing large-scale speaking tests that are practical to administer, that simulate test-takers' real world oral communication demands, and that yield reliable (sometimes defined as objective) assessments, and there are inevitable trade-offs (Bachman and Palmer 1996).

Situated in the context of a global L2 testing culture where technological innovations have resulted in novel approaches and possibilities for capturing speech, which is an essentially ephemeral and intangible medium, and where incorporating a speaking component in L2 proficiency testing is increasingly prevalent, the goal of this chapter is to elucidate contemporary challenges in the assessment of L2 speaking through a description of emergent notions and practices. The focus is broadly on the ways in which the construct of L2 speaking ability has been defined and operationalized in assessment instruments, on sources of variability in speaking performances, and on the validity of human judgments and technology-mediated assessments. A historical overview is used as a launching pad for highlighting matters that have currency in teaching, researching, and assessing speaking and as a springboard for discussing future trends.

2. Historical perspective and link to current conceptualizations and practice

In a 1929 *Modern Language Journal* article, Lundeberg (1929: 193) contends that "oral-aural skills are today recognized as desirable and attainable objectives. A great many teachers and administrators… rank the attainment of ear and tongue skills very high among their objectives. The layman, especially the parent, would often have this practical phase of language study placed first in the list".

Notably, Lundeberg's reference to "ear and tongue skills" groups aural/oral skills together. Although the relationship between human speech perception and production has been the object of considerable scientific debate and inquiry (Baker and Trofimovich, 2005), with some L2 communication models particular to speech perception (e.g. Flege 1995), and others to L2 speech production (e.g. de Bot 1992), Lundeberg's reference to aural/oral skills in tandem reflects the notion that "listening and speaking are theoretically and practically very difficult to separate" (Douglas 1997: 25). From a macro-level perspective of real-world interactional demands (as opposed to a technical perspective involving input processing and/or output mechanisms), verbal communication between human interlocutors is often a two-way process, involving, to use the

transmission (broadcasting) metaphor, both a sender and a receiver of a speech signal (i.e. sound wave) that needs to be encoded and, in turn, decoded by the relevant parties, to achieve successful communication in real-time (De Bot, Lowie and Verspoor 2007). Thus, in settings involving the exchange of verbal information, conversational partners often switch seamlessly between the role of speaker and listener to keep the flow of the conversation going under time pressure, sometimes modifying or adapting their speech to promote informational exchange (Jenkins 2000). In light of such conversational contexts, it is sometimes conceptually difficult or artificial to separate speaking and listening. This may be an important consideration in more contemporary notions of construct definition within the field of language testing, in view of conceptualizing and concretizing the real-world context(s) that test scores need to extrapolate to, which could have bearing on task design in a language test (Bachman and Palmer 1996).

Grouping "ear and tongue skills" together is also at the exclusion of the written medium, which raises crucial challenges inherent to speech perception and production. In contrast to written language, speech is both ephemeral and intangible, with technology needed to both render it visually, and to generate and digitally store a record of a performance (Hewlett and Beck 2006). Audio recording devices afford speech some permanency so that an utterance can be captured in time then played, replayed, sped up, slowed down, edited to optimize the recording quality, or digitally altered to experimentally manipulate a particular linguistic feature that might be the subject of an empirical investigation (e.g. establishing even versus uneven syllable length conditions to examine temporal duration effects on test performance). Presuming that a reasonable sound quality is obtained in the original recording, capturing a speaking performance using a digital recorder offers possibilities for scoring or transcribing a spoken L2 performance after its live rendering, or for administering standardized prerecorded speech elicitation prompts or audio-mediated instructions for completing a task. Technology has also allowed invisible sound waves, which humans perceive auditorily, to be represented visually (e.g. via waveforms or spectrographs). The ability to visually inspect speech, coupled with the use of recording devices to render sounds permanent, has made the study of speech more conducive to precise scientific measurement. Thus, advancements in technology have made it possible to turn the auditory, dynamic, and ephemeral phenomenon of speech into a visual, static medium that can be stored indefinitely or transcribed at leisure. That is, technology has transformed speech, which "consists of inherently difficult patterns for humans to attend to" (Port 2007: 362) into a tangible entity that can be quantified instrumentally (e.g. using speech editing software such as *PRAAT*), and that also opens up possibilities of obtaining ratings after the fact from any number of human raters who may not have been present at the L2 performance.

The possibilities afforded by the use of technology give rise to crucial distinctions between different modes of speaking assessment. A *direct* speaking test denotes assessing speaking through face-to-face oral communication with an interlocutor, interviewer, or examiner. This contrasts with a *semi-direct* speaking test, which denotes a machine-mediated assessment involving the test-taker uttering responses into a recording device without a human interlocutor being present to elicit the output (Qian 2009). Finally, an *indirect* test involves assessing speaking without having the test-taker actually produce spoken language (e.g. using written multiple choice item as indicators or predictors of speaking ability; O'Loughlin 2001). Examples of speaking tests that incorporate these modes and a discussion of the use of technology on scoring are expanded upon in subsequent sections of this chapter.

Reverting to the substance of Lundeberg's (1929) quote cited above, it is evident that he champions the importance of aural/oral communication in classroom settings and emphasizes its importance to stakeholders. Echoing this, Lado's seminal book, *Language Testing*, which is widely regarded as constituting the birthplace of language assessment as a discipline in its own right (Spolsky 1995), heralds speaking as "the most highly prized language skill" (1961: 239). Lado buttresses this claim with the logic that being able to speak an L2 implies an ability to understand it. He further argues that in the case of most world languages, L2 speaking ability facilitates and accelerates learning how to write it.

Despite these historical views endorsing an emphasis on L2 speaking ability, which seem consonant with speaking needs in the 21st century, a focus on speaking ability in instruction and, by implication, assessment in modern history has not been a given, with speaking ability and the linguistic skills focused on within that broad construct proving unrobust to instructional pendulum swings over the course of modern pedagogical history. Lundeberg (1929: 193) decries the lack of instructional emphasis on aural/oral skills as "one of the heaviest broadsides delivered against our teaching". The Grammar Translation Method, which was introduced in the late 18th century and which dominated L2 instruction until the 1940s (Richards and Rodgers 2001) centered on "grammatical exactness and translating ability" as a means of intellectual development, effectively sidelining speaking (Kelly 1969: 382). With respect to assessment, the *Certificate of Proficiency in English* (*CPE*), introduced in 1913 and produced within the British assessment tradition (Spolsky 1995), is an early and pioneering exception. Theoretically inspired by the Reform Movement's emphasis on speaking ability, which centred on the instruction of phonetics (e.g. Sweet 1899), the mandatory speaking component of the *CPE* included a conversation task— spearheading a tradition of oral interaction in Cambridge exams— to complement more traditional read-aloud and oral dictation tasks (Weir, Vidaković and Galaczi 2013). The Cambridge English Centenary Symposium on Speaking Assessment, held in 2013 in Cambridge, UK, celebrated this major development in the evolution of the speaking construct and inclusion of speaking as an integral component in L2 proficiency testing.

However, it was not until the Second World War that a concerted focus on the assessment of L2 speaking emerged (Fulcher 2003), with the American role in international affairs catalyzing the "demand for ability to actually speak the foreign tongue" (Kaulfers 1944: 136). In a *Modern Language Journal* article on wartime developments in American testing, Kaulfers emphasizes the need of ordinary citizens travelling abroad, as military personnel, to "communicate information in an intelligible way" in their spoken interactions. For assessments for military purposes in particular, he emphasizes the high-stakes consequences of miscommunications, commenting that the role of the task from a test development perspective is to generate "recognizable evidence of the examinee's readiness to perform in a life-situation where lack of ability to understand and speak the language extemporaneously might be a serious handicap to safety and comfort, or to the effective execution of military responsibilities" (1944: 137).

Kaulfers' "oral fluency" test, which assesses L2 readiness for military deployment, requires test-takers to translate key phrases from the first language (L1) into the L2 under the broad categories of asking for or responding to information requests and giving directions or polite commands. Speaking performances are then scored by human raters on a rating instrument comprised of two 4-level subscales, with a procedure for deriving median scores across language functions. The first subscale, which assesses "quality of language," describes the functions that the

test-taker demonstrates being able to execute, ranging from "can make known only a few essential wants in set phrases or sentences" to "can converse extemporaneously on any topic within the range of his knowledge or experience" (1944: 144). This scale is arguably a precursor to performance descriptors or "can do statements" in contemporary rating scales that outline the functions that learners are able to accomplish when carrying out selected communicative tasks, such as the *Canadian Language Benchmarks* (Pawlikowska-Smith 2002) or the *English Profile* linked to the *Common European Framework of Reference* (Salamoura and Saville 2010). This represents a source of continuity between past and present for gauging raters' judgments of the extent to which communicative goals on envisioned tasks have successfully been achieved, with an emphasis on extrapolating score results to tasks that test-takers are likely to perform in real-world settings.

Kaulfers' second subscale, which assesses the "quality of oral performance," appears as a scale with the bolded headings "unintelligible or no response," "partially intelligible," "intelligible but labored," and "readily intelligible," underpinned by more elaborate descriptors designating the extent to which a "literate native" (possibly the rater him- or herself) is able to understand the speech at each level of the scale (1944: 144). At the lowest band, the test-taker's speech results in the "literate native" (listener) being "confused or mislead [sic]" whereas "poor pronunciation or usage" or the presence of pronunciation errors designates the mid-levels of the scale. Finally, the highest band descriptor refers to the listener's inability "to identify the speaker's particular foreign nationality," which implies the absence of a detectable, or at least readily identifiable, L2 accent.

Of great historical importance in Kaulfer's scale is the very early reference to the notion of intelligibility—a construct that has currency in the teaching and assessment of L2 speech. Although intelligibility has been defined and measured in different ways for research and assessment purposes, paralleling the multiple interpretations that abound for a term such as "fluency," for example (Koponen and Riggenbach 2000), intelligibility, in the broad sense of the word, denotes the ease or difficulty with which a listener understands L2 speech (Isaacs and Trofimovich 2012). Widely regarded as the goal of L2 pronunciation instruction, and by implication, assessment (Levis 2006), intelligibility is featured as an assessment criterion in the speaking scales of standardized tests, including the *Test of English as a Foreign Language* (*TOEFL*) and the *International English Language Testing System* (*IELTS*). However, there is a need to pinpoint, with greater precision and empirical substantiation, the linguistic factors most important for achieving intelligible speech at different scale bands to guide raters' judgments and to crystalize the linguistic factors associated with this construct that are most important for making level distinctions (Isaacs and Trofimovich 2012). Although intelligibility has traditionally been associated with pronunciation, recent research has revealed that a listener's ability to understand L2 speech is not confined to pronunciation, but also extends to other linguistic domains, including discourse measures, lexical richness, lexicogrammatical, and fluency (temporal) variables (Trofimovich and Isaacs 2012). That is, many linguistic features in addition to pronunciation can have bearing on a listener's understanding of L2 speech. By contrast, L2 accent appears to be a much narrower construct that is most strongly related to the linguistic factors commonly referred to under the umbrella term "pronunciation," including word stress, rhythm, and segmental (vowel and consonant) production accuracy.

Although the notion of the native speaker is not without controversy in language testing circles and among applied linguists more generally (Davies 2011; Kubota 2009), in operational assessment and research settings, native or native-like speakers still routinely conduct ratings of L2 speech (Levis 2006). Reference to the absence of a recognizable nonnative accent in Kaulfers's

(1944) scale is reminiscent of L2 speaking scales in use today that explicitly refer to achieving the native speaker standard (i.e. L2 accent-free speech) at the highest level of the scale (e.g. *Cambridge ESOL Common Scale for Speaking*; see Taylor 2011). However, it is possible to be highly intelligible but to still have a lingering L2 accent. The consensus view among applied linguists today is that, although accents are perceptually salient, accent reduction is not an appropriate instructional goal (Derwing and Munro 2009). This is in part because L2 learners are able to integrate into a new society and to perform well in workplace or academic settings without needing to sound like native speakers. This argument can be extrapolated to assessment contexts (Isaacs 2014). Notwithstanding the extremely rare situations in which the L2 learner needs to sound like a native speaker to conceal his/her identity in the real-world domain to which speaking scores are being generalized (e.g. employment as a secret service agent), it is only when the accent interferes with a learner's ability to be understood by the listener that it should be alluded to in speaking scale descriptors (Alderson 1991). In sum, referring to intelligible speech at the high end of the scale without resorting to a native-like accent is an assessment challenge now as then.

In terms of the logistics of his proposed speaking assessment, Kaulfers (1944: 150) suggests that test-takers' responses be audio recorded, with an examiner administering the test and independent raters scoring the recorded responses. Although he concedes that this "pioneering" speaking test development effort is not without its limitations, he suggests that the proposed test may expose the shortcomings of "silent pen-and-ink exercices" (i.e. indirect speaking test items). After inviting classroom teachers to trial the proposed assessment, he emphasizes that examining the reliability, validity, and norming procedures in speaking assessments in general would comprise a good dissertation topic, since "the need is real."

3. Key issues and challenges in assessing L2 speech

Issues associated with assessing L2 speech articulated in historical texts decades ago shed light on the nature of the challenges inherent in the medium that still resonate today. Arguments that aural/oral skills "are less measurable because they are less tangible, more subject to variation, and probably will involve the cumbersome and time-consuming expedient of the individual oral examination" (Ludenberg 1929: 195) have served as a historical deterrent for developing and implementing direct or semi-direct speaking tests as part of L2 proficiency testing, particularly in the American testing tradition (Spolsky 1995). Wood's (1927) concern that some teachers base their scoring decisions on the content or substance of the test-taker's message while other teachers' decisions are informed by properties of the test-taker's oral production could at least partially be redressed through clearly defining the focal construct and explicitly addressing sources of construct-irrelevant variance in rater training (i.e. factors that should have no bearing on raters' scoring decisions; Bachman and Palmer 1996; Winke, Gass and Myford 2013). Alternatively, a more controlled task-type that constrains test-takers' expected output, such as a read-aloud task, would remove content as a source of variation across test-takers to enable a sole focus on the linguistic properties of their speech, if this was desired. However, as Lado (1961) clarifies, even in the case of decontextualized, discrete-point items measuring goodness of articulation (e.g. individual words or sounds), listeners may not achieve consensus in making simple binary judgments about whether a test-taker's oral production is "right" or "wrong" (e.g. an accurate or inaccurate production of a target sound), as this is, to an extent, contingent upon listeners' subjective perceptions. Indeed, recent research has shown that listeners' familiarity with the L2 accent of the test-taker may facilitate their understanding of L2 speech (Harding 2012) and bias

their assessments (Winke, Gass and Myford 2013). The implication is that ideally, this rater effect should be controlled for in high-stakes speaking assessments so as not to unduly influence raters' scoring decisions on a factor extraneous to the L2 speaking ability being measured.

Even if listeners do arrive at a numerically identical scores based on a given speaking performance, they may assign scores for qualitatively different reasons (Douglas 1994). To elaborate, when it comes to dichotomously scored items which are objectively verifiable and are not dependent on human judgment, the measurement model is relatively simple: the test-taker interacts with the test task to produce a score (Upshur and Turner 1999). When the assessment is listener-mediated and the resulting scores are not objectively verifiable, additional sources of systematic variance are introduced into the measurement model and reflected in the score that is generated. One such factor, listener or rater characteristics, has the potential to influence both the quantitative scores that raters assign, including the criteria that raters attend to when making their scoring decisions (Brown, Iwashita and McNamara, 2005), and the strategies they use to condense their possibly complex impressions of an L2 performance into a single numerical score (Lumley 2005). Further, the assessment framework, most often a rating scale, is yet another source of variance. The need to balance the practical consideration of providing raters with a user-friendly assessment instrument which features a manageable number of criteria in scale descriptors appears to be at odds with representing the construct comprehensively in rating scales, particularly when the construct of interest is global or multifaceted. Thus, rating descriptors necessarily oversimplify the complex and possibly nonlinear processes involved in L2 acquisition or task performance that they aim to represent (Isaacs and Thomson 2013). In addition, they are likely to fall short of reflecting the myriad factors that listeners attend to when arriving at their scoring decisions or the possible interactions among these factors, especially when the speaking task is complex (e.g. an extemporaneous speech sample or an interaction between a test-taker and an interlocutor, the latter of which would introduce yet another source of variance). In sum, the intangible nature of speech makes it arguably more challenging to assess than the written medium. This notwithstanding, eliciting raters' evaluative judgments of speech opens up a series of complexities and possible threats to the validity of the assessment that is also common to the scoring of L2 writing production.

Lundeberg's quote that aural-oral tests "probably will involve the cumbersome and time-consuming expedient of the individual oral examination" (1929: 195) underscores the practical problems associated with human-mediated assessments of speech. They are neither cost effective nor time efficient. For example, examiners' salaries for individually administering and scoring the test task(s), and, in the case of test performances that need to be recorded, mechanisms for securely storing test performance data (sound or video files) may make it too expensive or time consuming to engage in the direct testing of L2 speech en masse, and this consideration may ultimately outweigh the use of more authentic task types. Such practical considerations led Lado (1961) to advocate the use of indirect testing as an alternative to oral production tests, although his hypothesis that written responses and actual oral productions of the tested word would strongly correlate was not borne out in empirical studies which tested his hypothesis. In fact, indirect test items modelled on Lado's prototype indirect items yielded "catastrophically low reliabilities" and concomitant validity issues (Buck 1989: 54). Thus, despite the practical challenges involved in assessing speaking, direct or semi-direct tests are considered the only valid formats for assessing L2 speech today (O'Loughlin 2001).

The reliance on listeners' impressionistic judgments of speech is arguably at odds with the strongly psychometrically-influenced American assessment tradition. Some of the hallmarks of this tradition are the superordinate focus on the technical (statistical) reliability of test items, heralding of multiple choice as the gold standard of measurement, and tendency to opt for the most administratively feasible test formats and item types in the context of large-scale, high-stakes tests (Bachman et al. 1995; Spolsky 1995). The *TOEFL*, a product of this tradition, was launched in 1964 as an English proficiency test for academic purposes (ETS 2011). In the initial paper-based version (*pBT*) of the test, only the reading and listening sections were compulsory. In the subsequent computer-based version (*cBT*), separate speaking and writing tests were developed as optional supplementary tests, with the speaking test (*Test of Spoken English*) becoming operational in the early 1980s (Spolsky 1995). It was only with the introduction of the internet-based version (*iBT*) of the test in 2005—the most recent revision to date—that the speaking component became compulsory and was included as part of the mainstream *TOEFL*. Due to its widespread use as a language screening instrument at many English-medium universities internationally, students who took earlier versions of the test (*pBT* or *cBT*) were admitted to postsecondary institutions without any assessment of their speaking ability, with no separate speaking component generally mandated for university entrance purposes (Isaacs, 2008). In addition, some higher education institutions did not screen international teaching assistants for spoken proficiency, although speaking is clearly of importance in the academic domain, particularly in the case of individuals bearing instructional responsibilities (Isaacs, 2013; Saif 2002). To complement research on the validity of the *TOEFL iBT* (e.g., Farnsworth 2013), there is a growing body of work examining washback effects, particularly as English teachers have sought to adapt to the addition of the speaking section, which constituted a major change. In a multiphase washback study that included a baseline study to describe pedagogical practice prior to the introduction of the *TOEFL iBT*, Wall and Horák (2006; 2008; 2011) found overall positive teacher attitudes toward testing speaking, with the effect of more class time reportedly being allocated to speaking at the Central and Eastern European TOEFL test centers examined, suggesting a positive influence of the *TOEFL iBT* speaking test on instruction from the teachers' perspective.

The *TOEFL iBT* speaking tasks are semi-direct (computer-mediated) with standardized recorded prompts. The scoring is conducted by trained *TOEFL* raters using separate scales for the two different speaking task types—independent (involving speaking only) and integrated (involving speaking after listening to either a spoken prompt, or to both a spoken and written prompt, with the task directive of summarizing and synthesizing information from these sources in the spoken response; ETS 2008). These factors are major innovations of the *TOEFL iBT*. The first, related to constructing and using rating scales specific to the task type, arguably draws, at least implicitly, on the sociointeractional view that L2 ability and the assessment context are inseparable, since assessments are locally-situated, with the task embedded in the way that the construct is operationalized in the scale (Chalhoub-Deville 2003). This contrasts with the traditional psychometric view that the L2 ability being measured can be completely disentangled from the task or task type and, more broadly, from the context of the assessment (i.e. what is being assessed is the individual's cognitive ability, as opposed to his/her interaction with a particular task or task type to yield the performance; Bachman 2007). This locus of debate has implications for the specificity of rating scales and for the generalizability of the speaking performance beyond the local context of the assessment. The second innovation is challenging the Ladoesque (1961) notion of partitioning the assessment of language ability into separate component skills. Because speaking, listening, and writing skills often occur in tandem in the real-world tasks that test-takers

are likely to undertake, grouping such skills together arguably enhances the authenticity of the test task to a greater extent than does simply assessing isolated skills and abilities (Plakans 2013).

In contrast to the American assessment tradition, the British tradition has not been driven by forefronting technical reliability, nor by using practically expedient task types in terms of test administration and scoring, with a speaking assessment tradition involving the use of direct, face-to-face test tasks (Weir, Vidaković and Galaczi 2013). Less driven by psychometric concerns about the need to replicate exact testing conditions across test-takers, the British tradition has tended to emphasize interactional tasks involving human interlocutors, with little experimentation with semi-direct or indirect task types and with reliance on expert judgments in scoring as an integral part of the tradition (Taylor 2011). In comparison with semi-direct tasks, face-to-face interactions tend to be more appealing to test-takers and may result in more authentic assessments, which is a trade-off for a more efficient and possibly cheaper assessment involving the use of technology for test administration and scoring (Qian 2009).

Drawing on the oral proficiency interview tradition, which can be traced back to the American Council on the Teaching of Foreign Languages in the 1950s (Fulcher 2003), the *IELTS* is a face-to-face oral interview with a trained *IELTS* examiner. The interviewer variable opens up an additional source of systematic variance in the measurement model, with individual differences in the interviewer's behavior while administering the *IELTS* having been shown to exert an influence on both test-takers' performance, and on raters' perceptions of the test-taker's speaking ability (Brown 2005). The scripting of test prompts is the primary mechanism that the exam provider has implemented to minimize interviewer variation and ensure a degree of standardization in test administration procedures (Taylor 2011).

Although oral interviews have been the preferred and most commonly used method of speaking assessment since the Communicative era (Luoma 2004), there has been a growing trend toward the assessment of peer performance on interactional tasks. This has been buttressed by findings from Second Language Acquisition (SLA) research on the facilitative effects of peer interactions on L2 learning (e.g. Mackey and Goo 2007) and on the value placed on pair and group work in the L2 classroom—which is generally regarded as good pedagogical practice—in promoting positive washback. In addition, peer interactions have the advantage of bypassing the power imbalance that is inherent in oral proficiency interviews (Ducasse and Brown 2009).

The Cambridge main suite of exams first adopted the paired speaking test format in 1996. Following a brief monologue task and response to interviewer questions, the test culminates in a peer collaborative task, thereby eliciting different interactional patterns within the scope of the speaking test and sampling from a broader range of dimensions within the L2 oral proficiency construct than would be possible on a single monologic task (Saville and Hargreaves 1999). However, the caveat is that interlocutor variables (i.e. test-taker characteristics) on the paired interaction task could affect the quality of the interaction that is generated and, consequently, test-takers' individual and collective performance and scoring outcomes. Such peer interlocutor variables (e.g. L2 proficiency, L1 background, gender, personality characteristics, attitudinal variables, etc.) are extraneous to the construct being measured and could threaten the validity of the assessment (Isaacs 2013). The different interactional patterns elicited in the Cambridge main suite of exams arguably serve as a buffer against this concern, since test scores are based on examiner cumulative judgments of test-taker performance on all three tasks (monologue and

interactions with both the examiner, and the test-taker partner; Saville and Hargreaves 1999). An issue of current debate associated with paired speaking test tasks is whether they should be individually scored or subject to joint scoring to reflect the co-constructed nature of the resulting discourse, with implications for rating scale design (May 2009). This topical debate often revolves around issues regarding the appropriate definition and operationalization of the speaking construct (which could include interactional competence) and on concerns about fairness to test-takers. Notably, a difficulty with research in this area is in isolating the effects of interlocutor variables on test-taker performance, since they may interact in complex ways and are difficult to control for, sometimes resulting in contradictory findings about interlocutor effects across studies (Davis 2009).

Taken together, the British assessment tradition offers a striking contrast to the American tradition, with advantages and disadvantages to both.

4. From current trends to future directions

A central theme that has permeated this discussion has been issues and challenges associated with the assessment of L2 speech, with reference to the role of technology in rendering this fleeting and intangible medium more amenable to measurement. In addition to enabling time-delayed human subjective judgments of L2 performances, technological innovations have made it possible to instrumentally analyze speech using measures such as grammatical accuracy, lexical frequency, acoustic variables (e.g. pitch range), and temporal variables (e.g. mean length of run; Isaacs and Trofimovich 2012; Zechner et al. 2009). Such machine-derived measures can then be weighted and combined to derive machine-scored assessments of L2 speaking proficiency. In this way, it is possible to address the age-old concern about the reliability of subjectively scored oral test data without resorting to the indirect testing of objectively scored items (e.g. written multiple choice questions).

The debate about the removal of the human element in language testing has been catapulted by the launch of fully automated (i.e. machine-mediated) L2 speaking tests, including the *Versant* test (formerly *Phonepass*), available in various languages and used for high-stakes purposes (Downey et al. 2008) and *SpeechRater*, which is, as yet, intended for low-stakes *TOEFL iBT* test preparation purposes (Zechner et al. 2009). To train the speech recognizer or machine scoring system, L2 speech samples are scored by a large cross-section of human raters. Averaging scores across the resulting large volume of ratings effectively neutralizes individual rater idiosyncrasies, thereby removing concerns about rater effects that are present in nonautomatically-scored tests that normally involve the scalar judgments of just two or three trained raters (Van Moere & Downey, this volume).

Validation studies on the *Versant* have demonstrated strong statistical associations between the automated scores generated by the speech recognition algorithm and human ratings. In addition, strong correlations with scores on traditional L2 speaking subtests (e.g. *TOEFL iBT*; *IELTS*) suggest that the *Versant* is measuring a related speaking construct to these more traditional tests (Downey et al. 2008), although this is only part of the validity argument. This notwithstanding, *qualitatively* speaking, automated measures of speech do not replicate the factors that human listeners attend to or that feed into their scoring decisions (Isaacs 2014). Furthermore, because automated scoring involves pattern matching, controlled tasks that generate highly predictable L2

productions (e.g. constructed response or utterance repetition tasks) are much more amenable to automated scoring than communicatively-oriented extemporaneous speech tasks that elicit relatively unpredictable test-taker output (Zechner et al. 2009). It follows that the use of "constrained" tasks on the *Versant* has led to concerns within the language testing community about the narrowing of the speaking construct, especially in comparison with speaking tests that elicit different interactional patterns and are deemed more authentic (Chun 2006). However, as the *Versant* test developers have argued, there is a strong psycholinguistic basis for the use of such tasks, although this is not likely to be embraced by language testers coming from a more sociocultural tradition (Downey et al. 2008). Such debates about the value of automated speaking assessments are ongoing and serve to counterbalance and occasionally interface with contemporary debates on the use of paired or group oral tests.

With continued improvements in technology, speech recognition will become increasingly prominent in operational speaking tests and, thus, a growing subject of research and debate in the field. However, in the same way that the world is not, as yet, run by robots, machine scoring systems are not likely to completely supplant assessments of L2 speech. In our increasingly globalized world, it is ultimately human interlocutors who, in informal evaluative contexts, are the ultimate arbitrators of the extent to which the intended message has been transmitted and whether or not the communicative transaction has successfully been achieved. Thus, human judgments are likely to remain the standard against which automated speaking assessment systems will continue to need to be trained and, ultimately, benchmarked. Finally, as technology (e.g. social media) continues to revolutionize the nature of human communication and to open up new interactional possibilities on a global scale (Kramsch 2012), the need to perform and assess complex speaking tasks in reliable and valid ways will continue to persist.

5. References

Alderson, J. Charles. 1991. Bands and scores. In J. Charles Alderson & Brian North (eds.), *Language testing in the 1990s: The communicative legacy*, 71-86. London: Macmillan.

Bachman, Lyle F. 2007. What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In Janna Fox, Mari Wesche, Doreen Bayliss, Liying Cheng, Carolyn E. Turner & Christine Doe (eds.), *Language testing reconsidered*, 41–70. Ottawa, ON: University of Ottawa Press.

Bachman, Lyle F. & Adrian S. Palmer. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, Lyle F., Fred Davidson, Katherine Ryan & Inn-Chull Choi. 1995. *An investigation into the comparability of two tests of English as a foreign language*. Cambridge: Cambridge University Press.

Baker, Wendy & Pavel Trofimovich. 2005. Perceptual paths to accurate production of L2 vowels: The role of individual differences. *International Review of Applied Linguistics in Language Teaching* 44(3). 231-250.

Brown, Annie. 2005. *Interviewer variability in oral proficiency interviews*. Frankfurt: Peter Lang.

Brown, Annie, Noriko Iwashita & Tim F. McNamara. 2005. *An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks*. Monograph Series MS-29. Princeton, NJ: Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-05-05.pdf (accessed 12 April 2014)

Buck, Gary. 1989. Written tests of pronunciation: Do they work? *ELT Journal* 43(1). 50-56.

Chalhoub-Deville, Micheline. 2003. Second language interaction: Current perspectives and future trends. *Language Testing* 20(4). 369-383.

Chun, Christian W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly* 3(3). 295-306.

Davies, Alan. 2011. Does language testing need the native speaker? *Language Assessment Quarterly* 8(3). 291-308.

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

De Bot, Kees. 1992. A bilingual production model: Levelt's "speaking" model adapted. *Applied Linguistics* 13(1). 1-24.

De Bot, Kees, Wander Lowie & Marjolijn Verspoor. 2007. A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition* 10(1). 7-21.

Derwing, Tracey M. & Murray J. Munro. 2009. Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review* 66(2). 181-202.

Douglas, Dan. 1994. Quantity and quality in speaking test performance. *Language Testing* 11(2). 125-144.

Douglas, Dan. 1997. *Testing speaking ability in academic contexts: Theoretical considerations*. TOEFL-MS-08. Princeton, NJ: Educational Testing Service.

Downey, Ryan, Hossein Farhady, Rebecca Present-Thomas, Masahiro Suzuki & Alistair Van Moere. 2008. Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly* 5(2). 160-167.

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

ETS. 2008. *TOEFL® iBT Tips: How to prepare for the TOEFL iBT*. Princeton, NJ: Educational Testing Service.

ETS. 2011. TOEFL® program history. *TOEFL iBT® Research, 1*. Princeton, NJ: Educational Testing Service.

Farnsworth, Timothy L. 2013. An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly* 10(3). 274-291.

Flege, James E. 1995. Second language speech learning: Theory, findings, and problems. In Winifred Strange (ed.), *Speech perception and linguistic experience: Issues in cross-language research*, 233-277. Timonium, MD: York Press.

Fulcher, Glenn. 2003. *Testing second language speaking*. London: Pearson.

Gatbonton, Elizabeth & Pavel Trofimovich. 2008. The ethnic group affiliation and L2 proficiency link: Empirical evidence. *Language Awareness* 17(3). 229-248.

Harding, Luke. 2012. Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing* 29(2). 163-180.

Hewlett, Nigel & Janet Mackenzie Beck. 2006. *An introduction to the science of phonetics*. Mahwah, NJ: Lawrence Erlbaum.

Isaacs, Talia. 2008. Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review* 64(4). 555–580.

Isaacs, Talia. 2013. International engineering graduate students' interactional patterns on a paired speaking test: Interlocutors' perspectives. In Kim McDonough & Alison Mackey (eds.), *Second language interaction in diverse educational settings*, 227-246. Amsterdam: John Benjamins.

Isaacs, Talia. 2014. Assessing pronunciation. In Antony J. Kunnan (ed.), *The companion to language assessment*, 140-155. Hoboken, NJ: Wiley-Blackwell.

Isaacs, Talia & Ron I. Thomson. 2013. Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10(2). 135-159.

Isaacs, Talia & Pavel Trofimovich. 2012. "Deconstructing" comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34(3). 475-505.

Jenkins, Jennifer. 2000. *The phonology of English as an international language*. Oxford: Oxford University Press.

Kaulfers, Walter V. 1944. Wartime development in modern-language achievement testing. *Modern Language Journal* 28(2). 136-150.

Kelly, Louis G. 1969. *25 centuries of language teaching: An inquiry into the science, art, and development of language teaching methodology, 500 B.C.-1969*. Rowley, MA: Newbury House.

Koponen, Matti & Heidi Riggenbach. 2000. Overview: Varying perspectives on fluency. In Heidi Riggenbach (ed.), *Perspectives on fluency*, 5-24. Ann Arbor, MI: University of Michigan Press.

Kramsch, Claire. 2012. Why foreign language teachers need to have a multilingual outlook and what that means for their teaching practice. *Muitas Vozes* 1(1-2). 181-188.

Kubota, Ryuko. 2009. Rethinking the superiority of the native speaker: Toward a relational understanding of power. In Neriko Musha Doerr (ed.), *The native speaker concept: Ethnographic investigations of native speaker*, 233-247. Berlin: Mouton de Gruyter.

Lado, Robert. 1961. *Language testing: The construction and use of foreign language tests*. London: Longman.

Levis, John. 2006. Pronunciation and the assessment of spoken language. In Rebecca Hughes (ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice*, 245–270. New York: Palgrave Macmillan.

Lumley, Tom. 2005. *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Luoma, Sari. 2004. *Assessing speaking*. Cambridge: Cambridge University Press.

Lundeberg, Olav K. 1929. Recent developments in audition-speech tests. *The Modern Language Journal* 14(3). 193-202.

Mackey, Alison & Jaemyung Goo. 2007. Interaction research in SLA: A meta-analysis and research synthesis. In Alison Mackey (ed.), *Conversational interaction in second language acquisition: A collection of empirical studies*, 407-453. Oxford: Oxford University Press.

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

O'Loughlin, Kieran J. 2001. *The equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.

Pawlikowska-Smith, Grazyna. 2002. *Canadian Language Benchmarks 2000: Additional sample task ideas*. Ottawa, ON: Centre for Canadian Language Benchmarks.

Plakans, Lia. 2013. Assessment of integrated skills. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics* (Vol. 1). 204-212. Hoboken, NJ: Wiley-Blackwell.

Port, Robert F. 2007. The graphical basis of phones and phonemes. In Ocke-Schwen Bohn & Murray Munro (eds.), *Language experience in second language speech learning: In honor of James Emil Flege*, 349-365. Amsterdam: John Benjamins.

Qian, David D. 2009. Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly* 6(2). 113-125.

Richards, Jack C. & Theodore S. Rodgers. 2001. *Approaches and methods in language teaching* (2nd ed.). Cambridge: Cambridge University Press.

Saif, Shahrzad. 2002. A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics, 5(1-2),* 145–167.

Salamoura, Angeliki & Nick Saville. 2010. Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. In Inge Bartning, Maisa Martin & Ineke Vedder (eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 101-132. http://eurosla.org/monographs/EM01/101-132Salamoura_Saville.pdf (accessed 12 April 2014).

Saville, Nick & Peter Hargreaves. 1999. Assessing speaking in the revised FCE. *ELT Journal* 53(1). 42-51.

Spolsky, Bernard. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Sweet, Henry 1899. *The practical study of languages: A guide for teachers and learners*. London: Dent.

Taylor, Lynda. 2011. (ed.). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Trofimovich, Pavel & Talia Isaacs. 2012. Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition* 15(4). 905-916.

Upshur, John A. & Carolyn E. Turner. 1999. Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing* 16(1). 82-111.

Van Moere, Alistair & Ryan Downey. This volume. Technology and artificial intelligence in language assessment. In Dina Tsagari & Jayanti Banerjee (eds.), *Handbook of second language assessment*. Berlin: DeGruyter Mouton.

Wall, Dianne & Tania Horák. 2006. *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study*. TOEFL Monograph MS-34. Princeton, NJ: Educational Testing Service.

Wall, Dianne & Tania Horák. 2008. *The Impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change*. Research Report RR-08-37. Princeton, NJ: Educational Testing Service.

Wall, Dianne & Tania Horák. 2011. *The impact of changes in the TOEFL examination on teaching in a sample of countries in Europe: Phase 3, the role of the coursebook and phase 4, describing change.* TOEFL iBT Research Report TOEFL iBT-17. Princeton: NJ, Educational Testing Service.

Weir, Cyril J., Ivana Vidaković & Evelina Galaczi. 2013. *Measured constructs: A history of Cambridge English language examinations 1913-2012*. Cambridge: Cambridge University Press.

Winke, Paula, Susan Gass & Carol Myford. 2013. Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30(2). 231-252.

Wood, Ben D. 1927. *New York experiments with new-type modern language tests*. New York: MacMillan.

Zechner, Klaus, Derrick Higgins, Xiaoming Xi & David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51(10). 883-895.