Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us

Catherine Elder, Tim McNamara and Hyejeong Kim, The University of Melbourne
John Pill, American University of Beirut
Takanori Sato, Sophia University

Highlights

- We consider the scope of current conceptualizations of communicative competence in tests of spoken language.
- We present studies illustrating aspects of performance underrepresented in traditional criteria for speaking tests.
- We propose the greater use of non-language specialists' views to determine assessment criteria.

Keywords

Communicative competence, Construct definition, Indigenous assessment, Language testing, Lay perspective, Testing Language for Specific Purposes

Abstract

Models of communicative competence in a second language invoked in defining the construct of widely used tests of communicative language ability have drawn largely on the work of language specialists. The risk of exclusive reliance on language expertise to conceptualize, design and administer language tests is that test scores may carry meanings that are misaligned with the values of non-language specialists, that is, those without language expertise but perhaps with expert knowledge in the domain of concern. Neglect of the perspective of lay (i.e. non-linguistic) judges on language and communication is a serious validity concern, since they are the ultimate arbiters of what matters for effective communication in the relevant context of language use.

The paper reports on three research studies exploring the validity of rating scales used to assess speaking performance on a number of high-stakes English-language tests developed for professional or general proficiency assessment purposes in Korea, Australia, China and the UK. Drawing on Jacoby and McNamara's (1999) notion of "indigenous assessment", each project attempted to identify the values underlying non-language specialists' judgements of spoken communication as they rated test performance or participated in focus-group workshops where they viewed and commented on video- or audio-recorded samples of performance in the relevant real-world domain.

The findings of these studies raise the question of whether language can or should be assessed as object independently of the content which it conveys or without regard for the goal and context of the communication. The studies' findings also cast doubt on the notion that the native speaker should always serve as benchmark for judging communicative effectiveness, especially with tests of language for specific purposes, where native speakers and second-language learners alike may lack the requisite skills for the kind of effective interaction demanded by the context.

Introduction and literature review

For nearly 50 years, language specialists have conceptualized communicative ability for second-language (L2) communication, and have attempted to identify the components of knowledge and ability involved. Such attempts began in response to Chomsky's (1965) competence/performance distinction in which competence is narrowly restricted to grammatical knowledge. Hymes (1972) proposed an influential theory of communicative competence looking at competence from a sociolinguistic perspective and adding various elements to those discussed by Chomsky. Subsequently, models building on Hymes's work were developed by Canale and Swain (1980) and Bachman (1990) for L2 teaching and testing, a departure from the exclusive concern with traditional grammar which, as Joseph (this issue) reminds us, had dominated the foreign language curriculum and associated methods of assessing achievement for decades. These new models explicated the multiple components of language ability in detail and have served as a framework of reference for defining the construct of both specific- and general-purpose proficiency tests (Bachman & Palmer, 2010; Douglas, 2000).

In general, however, these models consist of detailed specification of language-related components (e.g., grammatical, discourse, and sociolinguistic knowledge) and have paid less attention, if any, to non-linguistic cognitive, affective, and volitional factors, seeing them as too complex to deal with, even though these factors were discussed extensively by Hymes as part of what he called *ability for use*. As a result, the construct of most L2 performance tests is typically defined purely in terms of cognitive linguistic ability, and assessment criteria used for performance tests normally include only language-related components. McNamara (1996) calls such performance tests *weak* performance tests, as opposed to *strong* performance tests, which assess performance based on real-world criteria or task fulfillment. He also claims that the majority of L2 performance tests are weak performance tests. This situation persists, although a more socially oriented model of interactional competence has been proposed (Kramsch, 1986) and elaborated by Jacoby and Ochs (1995), Young (2008) and work on distributed cognition (e.g., Hutchens, 1995) deploys models which transcend the boundaries of individual actors to encompass complex social practices. While these social views of performance have been acknowledged in the language testing field (McNamara & Roever, 2006), most performance tests, even those focusing on the co-constructed nature of performance (e.g., see Taylor & Wigglesworth, 2009), tend to place greater emphasis on the underlying linguistic qualities of performance than on criteria reflecting the complexity of communication in the target language use (TLU) domain. Harding (2014), in a recent overview of communicative language testing, highlights the need to move beyond these narrowly linguistic criteria and ensure that test constructs are rich enough to reflect current communicative needs.

A further limitation of L2 theories of communicative competence is that theory construction to date has not invited the perspectives of non-language specialists. (This lack of attention to lay views of language stems perhaps from linguists' dismissal of such views as unscientific (see Rajopolan. this issue).) As a result, the theories do not necessarily explain which features or behaviours of speakers are likely to be perceived as constituting competence in communication by those actually engaged in the communicative event. This could potentially undermine the validity of the theories and resulting test scores, since individuals with no specialized linguistic knowledge are in fact the ultimate arbiters of L2 speakers' oral performance in real-world language use domains; that is, L2 speakers are more likely to communicate with non-language specialists than with applied linguists, and to be judged based on their perspectives (Barnwell, 1986; Brindley, 1991; Chalhoub-Deville, 1996).

We thus have a double narrowing of the criteria by which performance is to be judged: linguistic features of performance are privileged; and the criteria by which those actually involved in the communication judge its success have not been attended to. At what cost has this narrowing of the construct of communicative language ability by applied linguists and language testers, partly in the interests of test manageability, been achieved? A number of studies have investigated (a) the dissonance between language specialist and linguistic lay perspectives on communicative competence and (b) the assessment criteria underlying the judgements of domain experts (i.e., non-language specialists) in specialized TLU domains.

Empirical research comparing non-language specialist and language specialist perspectives has shown that the former group tends to judge the communicative competence of L2 speakers differently from the latter (Brown, 1995; Elder, 1993; Galloway, 1980; Hadden, 1991). The two groups have been found to attend to different speech features and to show different levels of sensitivity to language form. Language specialists are generally more sensitive to linguistic form and more severe on linguistic errors. Furthermore, while studies analyzing patterns in data from judgements by language teachers (Brown, Iwashita, & McNamara, 2005; McNamara, 1990; Zhang & Elder, 2011) have typically[1] indicated that they are basing their overall judgment of

[1] Though not invariably: cf Hinofotis, Bailey, & Stern, 1981; Sato, 2012

communicative competence on language proficiency or grammatical accuracy, often to an extent of which they are unaware (Eckes, 2009), these features seem to play a less salient role in linguistic lay-people's evaluative judgments. Instead, non-language specialists are concerned more with successful communication and performance features influencing communicative success more directly.

Other studies have addressed the criteria used by domain experts in judging communication, including medical doctors, non-linguistics-related subject teachers, and professionals in various academic fields (Abdul Raof, 2011; Douglas & Myers, 2000; Jacoby, 1998). For example, Jacoby (1998) explored the criteria used by a group of physicists in providing feedback on practice oral presentations of post-doctoral researchers and PhD candidates. Using a Conversation Analytic methodology she analyzed the physicists' discussion of the presenters' rehearsals for conference presentations and uncovered the implicit criteria indigenous to that communicative context. She found that the group appeared to orient exclusively to non-linguistic criteria, paying little attention to linguistic errors made by presenters who were non-native English speakers. Although indigenous criteria derived from this and other studies vary significantly, they have all shown that assessment criteria used by domain experts in judging actual communication are considerably different from the conventional linguistically-oriented criteria used in L2 oral proficiency tests. Ubiquitous linguistic-related features developed based on theories of communicative competence—involving grammar, vocabulary, pronunciation, and fluency—play a less prominent role in indigenous assessments. These studies on domain experts' perspectives have contributed to defining the specific-purpose communication ability required in particular domains. Jacoby and McNamara (1999) argue that "studies of naturally occurring 'indigenous' socialization and assessment practices in professional settings, can provide more direct access to what counts as communicative competence in particular contexts" (p. 214).

Incorporating ultimate arbiters' perspectives into test development enhances the validity of language-for-specific-purposes tests and general-purpose proficiency tests. As Bachman and Palmer (2010) claim, developers need to ensure that “the criteria and procedures for recording the responses to the assessment tasks correspond closely to those that are typically used by language users in assessing performance in TLU tasks” (p. 236). But to do so is to challenge the way in which L2 communicative performance is typically judged, as we will show in our account of three recent studies addressing this issue.

The three studies

The paper draws on three independent PhD studies canvassing the views of non-linguistically expert judges about what they valued in the quality of spoken communication. All three studies examined the validity of language tests used to assess English proficiency with a particular focus on the criteria used to assess speaking performance. All took as their point of departure the notion that language experts and those from other fields may differ in their views of language and what it means to communicate effectively, and (as argued above) that theories of communicative competence and any attempt to measure such competence should take account of these different perspectives.

The study by Kim (2012) was designed to interrogate the construct of radiotelephony communication as operationalized in the International Civil Aviation Organization (ICAO) guidelines, and in the English Proficiency Test for Aviation (EPTA) required for accreditation of non-native-English-speaking aviation professionals in Korea. The study by Pill (2013) aimed to revisit the criteria used on the speaking component of the Occupational English Test (OET; McNamara, 1996) designed to assess the communication skills of health professionals as part of the professional licensure process for those seeking to practise their profession in Australia, New Zealand or Singapore. The investigation by Sato (2014) sought to identify what determined lay-persons’ evaluations of speaking ability as represented in tests of general English proficiency: namely, the College English Test–Spoken English Test (CET–SET) designed to measure the oral English proficiency of graduating students in China (Zheng & Cheng, 2008) and three Cambridge English examinations for speakers of English as a foreign language.

Research questions

Although the research questions were formulated somewhat differently for each study, they can be summarized as follows:

1. What do non-linguistically-expert judges value about spoken communication in English for general or profession-specific purposes?
2. How might these values inform test constructs, rating scales and score interpretations in particular contexts and the way we conceive of spoken communication in general?

Methods

The non-linguistically-trained judges were different for each study as were the methods used to elicit their insights. Broadly speaking, the approach used in each case could be described as grounded ethnography, "an approach to describing and understanding a target language use situation from the perspective of language users in that situation" (Douglas, 2000, p. 93). The participants and elicitation techniques for each study are described briefly below.

Kim's study elicited feedback from aviation personnel via three primary methods, namely (a) a large scale survey of 300 pilots and 100 air-traffic controllers (b) follow-up structured interviews with a subset of 22 informants and (c) individual and focus-group commentaries from a sample of three experienced pilots and five experienced air-traffic controllers while listening to six audio-recorded episodes of radiotelephony discourse gathered in what the ICAO had classified as "non-routine", "abnormal", "emergency" or "distress" situations. The survey was designed to capture informants' views regarding the relevance of the ICAO proficiency guidelines and associated test in Korea to the requirements of radiotelephony communication, and the follow-up interviews served to illuminate the survey responses. The more detailed commentary on the radiotelephony discourse samples aimed both to explicate the specialist language of each episode for the benefit of the researcher and to uncover what communication practices the “insider” informants considered important for effective functioning in the aviation airspace.

Pill’s study drew its data from two major sources: (a) two workshops, convened expressly for the research, conducted with a purposive sample of 13 qualified health professionals in Melbourne with experience of supervising and giving feedback on performance to medical students and junior doctors and (b) 46 pre-existing written reports from medical educators drawn from a database used to track the progress of family medicine trainees taking a three-year clinically-based vocational training program to prepare them for the Fellowship examination of the Royal Australian College of General Practitioners. The workshops centred around training videos involving International Medical Graduates (IMGs) from non-English-speaking backgrounds practising consultation scenarios with simulated patients. These videos were used as stimuli to elicit medical educators’ views of effective communication. Participants were asked by the facilitating researcher to comment, one at a time, on the stronger and weaker aspects of each IMG performance in a manner resembling how they might give feedback in an actual training situation. The written reports, by contrast, consisted of actual feedback given by educators after observing a trainee engaging in a series of clinical consultations with their regular patients.

Sato‘s study canvassed the views of 23 graduate students from disciplines other than applied linguistics or Teaching English to Speakers of Other Languages Thus, the participants in his study had neither (a) specialized knowledge of applied linguistics, (b) experience of any training in language assessment and teaching, or (c) experience of rating and teaching L2 learners formally. Furthermore, the participants were drawn from all of Kachru’s (1988) concentric circles: the Expanding Circle (N=10), the Outer Circle (N=6), and the Inner Circle (N=7) in the interests of avoiding any bias in favour of native speaker norms. These lay judges viewed seven individual monologic presentations from the CET–SET and three paired interactions from the Cambridge English Certificate in Advanced English (CAE), First Certificate of English (FCE) and Preliminary English Test (PET) examinations. They then recorded their intuitive impressions of each test taker's performance on a scale from 1 (Poor) to 7 (Excellent) and provided concurrent verbal justifications for their ratings. Each speech sample was then reviewed and the judges were asked via a stimulated recall procedure to verbalize the features of the performance that influenced their judgments. A subsequent semi-structured interview was undertaken to elicit supplementary information.

Although a range of different methods was used to elicit and interpret feedback from the informants in the three studies, all yielded self-report data and each study used spoken stimuli (whether these were samples of test discourse, simulated interactions or actual workplace encounters), along with other methods, to elicit views of what constituted effective communication. Informants' commentary, whether spoken or written, was coded thematically by each researcher using an inductive, bottom-up approach (Lincoln & Guba, 1985) with rigorous documentation of the process and double-coding to ensure replicability. By scrutinizing the themes emerging from these different studies we were able to draw links between the findings as reported below.

Findings

The findings of each study are summarized briefly in turn below after which general trends linking the three studies are noted.

Survey and interview responses from the Korean aviation informants in Kim's (2012) study revealed strong resistance to current ICAO language proficiency requirements and to the test in Korea used to implement these requirements. Both were seen as placing undue emphasis on a decontextualized native-speaker standard of English proficiency at the expense of what were seen as the more critical issues of professional experience and expertise, and preparedness to cooperate in the English as a lingua franca communication that is characteristic of aviation interaction.

Some of the six linguistic criteria—Comprehension, Fluency, Interactions, Pronunciation, Structure, Vocabulary—specified in the ICAO guidelines and applied to the assessment performance on the associated EPTA (developed in

Korea) were seen as largely irrelevant to the professional situation, as indicated in the following comment from an experienced aviation professional.

> *Say someone speaks proficient English at the highest level of the ICAO rating scale. That is, a speaker provides all the details in a situation that he wants to say and all the details a hearer may want to know. How can the hearer process all the provided information doing other multiple tasks at the same time? And, radiotelephony communication is mostly comprised of instructions and requests and there's no need for description. Why does structure or grammar matter? It's the same for fluency. When we encounter an abnormal situation or emergency, because it's an unexpected situation, we have to think before making a judgement and a request. Of course this has an effect, so our fluency decreases. What's the point of being "fluent" in that situation?*
> KT, captain, 13 years of experience (cited in Kim & Elder, 2015, p. 143)

Analysis of the expert feedback on the recorded samples of radiotelephony discourse confirmed that a lack of professional knowledge by either pilot or air-traffic controller was deemed responsible for unnecessarily extended and potentially ambiguous communication as both parties attempted, and in some cases failed, to reach mutual understanding. Lack of adherence to standard phraseology conventions also impeded communication and the tendency of some pilots to give detailed information was not necessarily helpful given the pressures of the communication context. While it was conceded that reduced intelligibility (see Kim & Billington, 2016) and limited vocabulary knowledge could sometimes be an obstacle to understanding, Kim's analysis of informants' feedback also suggested that responsibility for misunderstanding is shared between interlocutors and that appropriate use of accommodation strategies by both native and non-native English-speaking aviation personnel is critical to achieving the plain English qualities of precise and efficient communication in aviation. She proposed that any test of communicative competence in aviation English should be required for both native and non-native English-speaking personnel, and should take into account these issues of co-construction in context (Jacoby & Ochs, 1995) as well the distribution of cognition beyond the individual to encompass the broader social and physical environment in which interaction takes place (e.g., Hutchins and Klausen, 2000 and see Joseph, this issue).

Pill’s (2013) study of the indigenous criteria oriented to by medical educators giving feedback on trainee doctors’ interactions with patients was used to build a model of what is valued by doctors in the doctor–patient consultation. The model included three overlapping and interdependent skill sets, Communication Skills, Clinical Skills and Practitioner Skills, all drawing on a shared repertoire of Interactional Tools used for performance of the consultation. (For a detailed exposition of this model see Pill, 2013, pp.189-200.)

While not all of these skill sets were amenable to inclusion in a language test like the OET, Pill concluded that language was essential to an effective consultation, because the interactional tools used by doctors with their patients accomplished clinical work by linguistic means. He found sufficient evidence in the analysis of the dataset for keeping the existing four analytic criteria—Intelligibility, Fluency, Appropriateness of Language, and Resources of Grammar and Expression—used to assess speaking performance on the OET speaking sub-test. However, he also proposed that two additional criteria Clinician Engagement and Management of Interaction be added to the existing set. These concern the ability of the health professional, respectively, to demonstrate his/her awareness of the patient (patient-centredness) and to gather and give information efficiently. These elements were highly valued by the participants in the study and seen as being realized through appropriate language behaviours, as illustrated in the following instances of health educator feedback. The first is directed to the trainee and proposes alternative and more sensitive wording for a question.

You had asked the patient “Do you want to harm yourself?” [whereas] it would be more appropriate to ask “Sometimes when people feel down, they feel like escaping/hurting themselves[.] Have you ever thought like this?” [R23-1-14]
(Pill, 2013, p. 206)

The second is more general in nature and reflects on how IMGs perform poorly under time pressure on simulated roleplay assessments used in the context of medical training to the detriment of efficient diagnosis.

A lot of them are scared of open questions because they think they'll lose time um although it always works the other way if they've got to keep thinking of a question every five seconds it takes them way more time than just saying "Tell me about your symptom" [wk2-330]
(Pill, 2013, p. 193)

Refocusing of the OET speaking assessment scheme to incorporate such professionally relevant considerations, Pill proposed, would extend the test's construct beyond a somewhat restrictive view of language as a decontextualized set of elements displayed in the performance of individual test takers to include aspects of their interactional competence (Kramsch, 1986) in a workplace setting.

Sato's (2014) study differed from the other two in that it canvassed views of test performance and the tests in question were not designed for any specific communicative purpose. Nor were the informants chosen for their expertise in a particular professional field. They were selected simply as representative of a general lay population with no specialist training in language matters. As noted under Method above, participants rated test performances both quantitatively (assigning scores) as well as qualitatively.

The quantitative results showed first of all that the informants' judgments did not always accord with the proficiency assessments of the same performances made by language-trained raters, and this was particularly true for the Cambridge English exams involving paired interactions, perhaps because the lay judges were more concerned with the flow of communication across the pair or their confluence (McCarthy, 2010) than with individual contributions to the exchange.

Thematic analysis of verbal protocol data identified a number of different elements—Demeanour, Non-verbal Behaviour, Pronunciation, Linguistic Resources, Fluency, Content, Interaction, and Overall Impression—as being influential in the lay informants' judgements. The frequency with which these elements were mentioned differed somewhat for monologic and dialogic speech samples as indicated in the table below.

Table 1. Elements influencing lay informants' judgments of speaking test performance (ranked by frequency of mention).

CET–SET (monologue)	Cambridge English (dialogue)
Overall Impression (21.1%)	Overall Impression (19.4%)
Content (15.1%)	Content (13.7%)
Fluency (13.5%)	Linguistic Resources (12.7%)
Other (12.4%)	Interaction (12.2%)
Pronunciation (11.5%)	Pronunciation (10.2%)
Linguistic Resources (10.4%)	Non-verbal Behaviour (9.6%)
Non-verbal Behaviour (9.4%)	Other (9.2%)
Demeanour (5.7%)	Fluency (6.6%)
Interaction (0.7%)	Demeanour (6.3%)

In spite of these frequency differences, it is interesting that Overall Impression and Content come so high on both lists, accounting for 36% and 33% of all comments across the two tests. The main components of these two largest categories—in particular, message conveyance, comprehensibility, and ideas—were considered by the informants to be closely related to the outcome of communicative performance. Although linguistic features such as grammar, vocabulary, pronunciation, and fluency were also recognized as factors impacting the outcome of communication, the participants neither considered them crucial nor penalized errors harshly unless comprehensibility was seriously impeded. Their impressions were also influenced by the test-takers' non-verbal behaviour and by non-language-exclusive cognitive and affective factors, such as perceived level of confidence, anxiety and/or willingness to communicate. In addition, in the paired interactions, the participants frequently noted interactional features such as engagement and the size of contribution.

Discussion and conclusion

Returning to our original question about what non-linguistically-expert judges value in spoken communication, it would seem that although the three studies involved diverse informants, methods and contexts and yielded somewhat different findings, all revealed reduced or little attention by non-language specialists to some of the linguistic categories such as accuracy which feature strongly in traditional language test rating scales, and greater emphasis on those aspects of message conveyance and interactional competence perceived as relevant to the goals of communication in each context. Informants' judgments were also coloured by non-linguistic features such as content quality and/or professional competence, which, as noted earlier, tend to present problems for existing models of communicative competence.

The three studies point to the practical consequences for assessment of the narrower model of communicative competence which currently informs many language assessments, in both specific-purpose and general-purpose contexts. By not acknowledging all that is relevant for successful communication in real-world situations, decisions are made to exclude individuals from participation in professional settings who may in fact be competent to practise and to allow others access to professional practice whose actual competence may cause problems of communication, with potentially serious, even fatal consequences. Thus in the Korean context, older pilots and air-traffic controllers were at risk of losing their right to practise as professionals under the terms of the new ICAO policy of requiring increased levels of English proficiency, even though their experience and knowledge of the communicative demands of workplace settings equipped them to compensate for any limitations in their English knowledge. In contrast, pilots with relatively high levels of proficiency, including native speakers, are often exempted from testing requirements and allowed to fly even when their communicative behaviour (ignoring the conventions of aviation communication designed to make it safe) and relative lack of experience mean that their communication ability is compromised. In fact, the Korean aviation authorities have subverted the impact of the test by disclosing all the items on the Internet prior to the test administration, so that candidates can memorize their answers and be certain of achieving the minimum required level. In this way the authorities have avoided the necessity of sacking their older and more experienced staff, who have demonstrated the safety of their practice over many years (Kim, 2012).

In the study of medical communication, the impact of adopting the new criteria emerging from the investigation of health professionals' indigenous assessment practices described above, has resulted in somewhat different decisions about who should be admitted to practice and who excluded (Pill & McNamara, 2016). Arguably, these are more valid decisions as they reflect more what is actually considered important and relevant in communication in healthcare settings by those most familiar with the context. Again, the views of applied linguists without experience of the health context, focusing exclusively on linguistic features of communication, resulted in what can be seen as faulty decisions about admission to practice, with possible implications for patient safety. And more generally, Sato's study of the views of lay-persons about the communicative effectiveness of performance on a range of spoken language tests suggests the necessity for a revision of the criteria we should be using in such tests if they are to validly represent what matters to those likely be making judgements about the quality of communication in non-test situations.

The adoption of a new orientation to judging communicative effectiveness, based on the views of those actually involved in the communication would have far-reaching implications for both general-purpose and specific-purpose tests. It suggests that we need to revise our understanding of the nature of communicative competence, embracing and elaborating on the rich model proposed nearly 50 years ago by Dell Hymes, but never, to our knowledge, fully implemented in L2 assessment. Communicative success is dependent not only on language skills but on the abilities, cognitive and non-cognitive, of the whole person as deployed in the particular context of use. This also means that the already tenuous distinction between first- and second-language speakers (as discussed by Davies 2003, 2004 and see Joseph, this issue) will be further reduced in scope, as so much of what contributes to successful communication in occupational or academic settings will be the same whether an individual is speaking their first or an additional language. Reducing the significance of this distinction will reflect the reality of most workplaces and academic settings in contemporary urban societies, where the demands of the setting are felt by native and non-native speakers alike and where participants routinely collaborate in meeting those demands, without specific attention to native-speaker status. Accordingly the relevance of the native-speaker norm, so central to applied linguistics for many years (Davies, 2003, 2013) may once again need to be reconsidered, and the justification for specific tests for L2 speakers, when assessing readiness to manage the complex communicative demands of real-world encounters, called again into question.

Would Alan Davies be sceptical our proposal to redefine language proficiency in light of the views of lay persons' orientations and reset test norms accordingly? Yes indeed, as seen in his unnervingly trenchant response to a preview of the abstract for this paper:

> How do we know that the laity's belief that they are right is indeed right? Isn't it possible that they have a folk linguistic view of language which does not stand up in the criterion situation? Asking them what they think is all very well but are they thinking straight?" (Davies, personal communication 17/06/2015)

Alan's point is well taken, but our response to his critique is that language experts do not have a monopoly on straight thinking about language in that we too have been socialized into accepting certain understandings as given. Indeed we have argued that this is the weakness of our current models, which emphasise aspects of language that seem salient to us but may not be the ones that serve people well in the complex acts of communication that they engage in. Our role as applied linguists is surely to examine to interrogate our own understandings and also to understand the norms by which others operate, using methodologies which are open to critical scrutiny, such as those applied in the studies reported here. While the new models and associated rating scales we devise to incorporate lay perspectives will be imperfect, based as they are on limited sampling and the inevitable biases of both informants and researchers, they will hopefully come closer to capturing what matters to language users in the contexts of concern than those currently available. We believe that our position is in keeping with Alan’s broader intellectual stance as cited in Rajopolan (this issue):

> One of the tasks of Applied Linguistics is to investigate which social model a speech community in practice selects as its language standard or model, to attempt an explanation of that choice, however hegemonic it may be, and to explore the concomitant institutional implications. (Davies, 1997: 5)

References

Abdul Raof, A. H. (2011). An alternative approach to rating scale development. In B. O’Sullivan (Ed.), *Language testing: Theories and practices* (pp. 151-163). Basingstoke: Palgrave Macmillan.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Barnwell, D. (1986). *Who is to judge how well others speak? An experiment with the ACTFL/ETS oral proficiency scale*. Proceedings Eastern States Conference on Linguistics, 3, 37-45.

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: SEAMEO Regional Language Centre.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, A., Iwashita, N., & McNamara, T.F. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. TOEFL Monograph Series. Princeton, NJ: Educational Testing Service.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 55-73). Cambridge: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Davies, A. (1997). Real language norms: description, prescription and their critics. A case for Applied Linguistics.' *View[z] Vienna English Working Papers. Vol 6,* 2, 4-18.

Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, Avon: Multilingual Matters.

Davies, A. (2004). The native speaker in applied linguistics, In Davies, A. & Elder, C. (eds), *Handbook of Applied Linguistics*. (pp.431-450). London: Blackwell

Davies, A. (2013). *Native speakers and native users*. Cambridge, Cambridge University Press.

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

Douglas, D., & Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 60-81). Cambridge: University of Cambridge Local Examinations Syndicate.

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment* (pp. 43-74). Frankfurt am Main: Peter Lang.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*(3), 235-254.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *The Modern Language Journal, 64*(4), 428-433.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, *41*(1), 1-24.

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

Hinofotis, F. B., Bailey, K. M., & Stern, S. L. (1981). Assessing the oral proficiency of prospective foreign teaching assistants: Instrument development. In A. S. Palmer, P. J. M. Groot & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence: Including proceedings of a colloquium at TESOL '79, Boston, February 27-28, 1979* (pp. 106-126). Washington, DC: Teachers of English to Speakers of Other Languages.

Hutchins, E. (1995). *Cognition in the wild.* MIT Press, Cambridge, Mass.

Hutchins, E. and Klausen, T. (1996). Distributed cognition in an airline cockpit. In D. Middleton & Y. Engeström(E ds.), Communication and cognition at work (pp. 15-34). Cambridge, UK Cambridge University Press

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J.

Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Jacoby, S. W. (1998). Science as performance: Socializing scientific discourse through the conference talk rehearsal. Unpublished PhD thesis, University of California, Los Angeles.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes*, *18*(3), 213-241.

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

Kachru, B. B. (1988). The sacred cows of English. *English Today, 16*(1), 3-8.

Kim, H. (2012). Exploring the construct of aviation communication: A critique of the ICAO language proficiency policy. Unpublished doctoral thesis, University of Melbourne, Australia.

Kim, H., & Billington, R. (2016). Pronunciation and comprehension in English as a lingua franca communication: Effect of L1 influence in international aviation communication. *Applied Linguistics*. doi: 10.1093/applin/amv075

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing*, *32*(2), 129-149.

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

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, *1*(1), 1-15.

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-75.

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

McNamara, T. F. & Roever, C. (2006). *Language testing: the social dimension*. Blackwell.

Pill, T. J. H. (2013). What doctors value in consultations and the implications for specific-purpose language testing. Unpublished doctoral thesis, University of Melbourne, Australia.

Pill, J. & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing 33,*2:217-334..

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223-241.

Sato, T. (2014). Linguistic laypersons' perspective on second language oral communication ability. Unpublished doctoral thesis, University of Melbourne, Australia.

Spolsky, B. (1989). Communicative competence, language proficiency, and beyond. *Applied Linguistics*, *10*(2), 138-156.

Taylor, L., & Wiggleworth, G. Are two heads better than one? Pairwork in L2 assessment contexts. *Language Testing, 26*(3), 325-339.

Young, R. F. (2008). *Language and interaction: An advanced resource book*. Oxford: Routledge.

Zhang, Y., & Elder, C. (2011). Judgements of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50*.*

Zheng, Y., & Cheng, L. (2008). College English Test (CET) in China. *Language Testing, 25*(3), 408-418.