

○ LANGUAGE TESTING AND ENGLISH AS AN INTERNATIONAL LANGUAGE

CONSTRAINTS AND CONTRIBUTIONS

Catherine Elder, University of Melbourne

Correspondence to Catherine Elder: caelder@unimelb.edu.au

Luke Harding, University of Melbourne

Correspondence to Luke Harding: lharding@unimelb.edu.au

INTRODUCTION

Clyne and Sharifian ("C & S") make special reference to English language testers in the conclusion of their paper, describing them (along with English language teachers) as "gatekeepers" of the English language. The authors suggest that these groups "need to explore and implement, in fundamental ways, the implications of the current and future situations of the complexity of English as an international language." More specifically, they argue that,

The expanding pluricentricity of English needs to be reflected in ELT materials, which currently lean towards either American or British English (and to a much lesser extent Australian English), under-representing other varieties. Language testing needs also to be informed by the unprecedented growth of variation in the norms of international communication (e.g. Canagarajah, 2006; Elder and Davies, 2006). Furthermore, the contents of language tests ought to correspond with the functions for which the testees will employ the language (see Zafar Khan, 2009). In many contexts, people who take language tests such as IELTS and TOEFL use English for intercultural communication, often in the absence of "native" speakers. In such cases, we believe the test should try to evaluate intercultural communicative skills instead of obsessively testing the "inner circle" Englishes.

From this passage, we can see that C & S critique the current practice of English language testing (in particular) on two main grounds:¹ firstly, that the growth of English varieties and the corresponding diversification of norms has not been adequately integrated into the design of English language tests; and secondly, that test developers should shift

their focus from measurement of proficiency in a particular (standard) variety to the assessment of a candidate's ability for "intercultural communication".

Each point has support in a tradition of papers which have commented on this apparent conservatism among test designers. For example, the first issue of "representation of diversity" has been the subject of a sustained critique in a series of papers by Lowenberg (e.g. 2002), who argues that in outer- and expanding-circle contexts, proficiency tests based on inner-circle norms may lack validity as they ignore the sociolinguistic reality of the candidates' language use (see also Brown and Lumley, 1998). The second point – that language tests are concentrating on the wrong skills – has found its most forceful expression in Canagarajah's (2006) paper (also cited by C & S), which argues for a fundamental shift in the object of language testing from one of "proficiency" to one of "negotiation" in recognition of the fact that the ability to communicate in EIL contexts may have more to do with using appropriate interpersonal strategies to accommodate differences across participants than with an individual's level of mastery of the English code. These are worthy points of contention, and form a serious critique of many current testing practices.

However the situation for language testers in grappling with these issues is more complex than C & S acknowledge. This paper will argue, firstly, that while C & S are right in pointing to language testers' reluctance to embrace EIL, they misrepresent the underlying reasons for language testers' resistance. The motivations of language testers in resisting a move to EIL testing are seldom as simple as "gatekeeping" for its own sake; rather they are driven by the testers' ethical responsibility to construct fair and useful tests, which reflect as faithfully as possible the norms operating in the relevant target language use situations, and which impose necessary constraints on test design. Secondly, this paper will argue that in exploring the limits of such constraints, language testers have, in fact, been responsible for a considerable amount of research which has pushed forward our understanding of EIL communication more broadly. It will also be shown that in many current testing situations, intercultural communicative skills are already, to a large extent, assessed. The paper concludes that, in spite of these promising developments, attempts by language testers to address the challenge of EIL must proceed in an evidence-based and consultative manner.

CONSTRAINTS

We shall begin by outlining some of the complex issues in language testing and how these can constrain language testers in what they measure and how they do so. As outlined

in Elder and Davies (2006), language testers need to adhere to certain principles or requirements in designing their instruments, if these are to be considered fair, meaningful and useful. Among these requirements are (a) construct validity or, put more simply, the fitness of the test for its purpose and context; (b) fairness, particularly with respect to absence of bias towards individuals or groups; and (c) accountability to a range of stakeholders, not least of which is a consideration of the views of test-takers themselves. These requirements bring with them certain constraints in test design, which have resulted in what might be seen as conservatism within the testing community in dealing with new varieties or uses of English. Each is discussed in greater detail below.

A fundamental requirement of language testing is that the construct underpinning a test, or the language knowledge and skills that a test is aiming to measure, be carefully defined. Once the test construct has been determined, testers need to demonstrate that their tests are capturing that construct through a process of argument- and/or evidence-based validation. If they cannot produce these justifications, doubts may be raised about the meaning of test scores, and in turn, about the validity of decisions based on these scores. Since it is widely acknowledged (e.g. by Seidlhofer, 2001 and Jenkins, 2007) that the very nature of EIL communication is that norms are fluid, it is not surprising that language testers have been reluctant to explicitly invoke international English as the basis for test design. To do so would arguably be unfair to test takers who would not know what kinds of language use were acceptable, what sources to draw upon in preparing themselves to pass a test, or indeed what standards were being applied by those judging their performance.

This is one reason why the default model for language testing is still often Standard English (SE) as used in “inner circle” countries and as codified in English grammars, dictionaries and the like. It allows for greater certainty about what is being assessed. Nevertheless, while SE remains as a point of reference, it is not clear that it dominates the design of language tests and disenfranchises users of other Englishes in quite the way that C & S and others such as Jenkins (2006) imply. The emphasis in the last three to four decades on communicative competence in English teaching curricula around the world – which has in turn influenced language test design – means that there is far less focus on the formal code than was formerly the case. For example, formal accuracy in Standard English grammar is only a small part of what is assessed on the IELTS and TOEFL tests that are named as offenders in C & S’s paper, and this element is never assessed independently of other aspects of language such as appropriateness. In many testing contexts, communicative effectiveness has replaced rigid adherence to SE norms

in the criteria used to assess productive language tasks (see rebuttal of Jenkins' position by Taylor, 2006).

But what do we mean by communicative effectiveness? How to define this construct? It is not particularly helpful to invoke EIL (or for that matter SE) when deciding what to include on a test or where to place the threshold between mastery and non mastery of English. Both SE and EIL are too general and cover too broad a range of users and uses. What goes into a test and how standards of communicative effectiveness are defined depends on the test's purpose. If the test is designed to measure medical, aviation or academic English, then it should simulate the language demands of that domain as closely as possible drawing on expert descriptions where these exist. The criterion for communicative effectiveness in the particular domain will be determined through consultation with the relevant stakeholders. What test designers will ask these stakeholders to do once the test has been trialed is to review samples of test performance at various levels and decide how much, and what nature, of English proficiency is sufficient to ensure optimum health outcomes, air safety or academic learning as the case may be. The language construct, in other words, needs to be defined very carefully with respect to the particular situation or purpose for which the test is intended. The validity of the test thus depends on adequate descriptions of the domain, adequate sampling in the form of representative test tasks or items, context-sensitive criteria, well-informed decision-making about thresholds between satisfactory and unsatisfactory performance, and appropriate evidence and argument supporting any inferences drawn from test scores.

Rather than invoking EIL or SE as constructs we are arguing for a targeted and contextualised description of what we are attempting to measure. This has long been accepted as good practice in language testing (Douglas, 2000). The TOEFL iBT, for example, which is radically different from earlier versions of this test, has moved in this direction by clearly defining itself as a measure of *academic* language ability and aligning itself more closely to "the variety of language use tasks that examinees are expected to encounter in everyday academic life" (Sawaki et al., 2009, p. 5). Extensive arguments and evidence have been adduced in support of the academic language construct which underpins the test (Chapelle et al., 2008). It is not therefore appropriate to criticize TOEFL for not measuring EIL or intercultural competence (as C & S have done) since that is not the purpose for which this particular test was intended.

What tests like the TOEFL can more reasonably be criticised for is any under-representation of the academic domain which they have explicitly claimed to sample from. As a case in point, the listening components of TOEFL (and, for that matter, of IELTS and the new Pearson test of English academic [PTE]), are restricted to inputs from educated

speakers of American, British or New Zealand English rather than from the full range of non-native English accents likely to be present amongst staff and students in any English-medium university. Why? We suggest that this is not driven by blindness on the part of language testers to the presence of such varieties, but rather by the second test requirement of fairness. The problem – as seen by test developers – of including the accents of non-native speakers on a test is that those test takers who share a speaker's accent, or who are more familiar with a particular accent, may find the speaker more intelligible than do other test takers.

In order to overcome this problem, test designers would need to sample very broadly from the range of possible accents encountered in academic contexts to make a test which is equally fair to test-takers from all L1 backgrounds. Yet limited sampling might also raise charges of bias against one or other language group. The fallback position, clearly safer and more practical, has been to stay with the standard varieties. The appeal of SE lies in its neutrality, in the sense that it is the variety most likely to be equally familiar to all test taker groups. Decisions such as these may be unduly conservative (see discussion of Harding, 2008 in the next section), but they have generally (if not historically then at least currently) been driven by a concern for fairness rather than a view that local or non-standard varieties of English have no claim to legitimacy.

There is a further and equally powerful reason for test designers' reluctance to embrace some of the challenges posed by the presence of an ever greater variety of uses and users of English and that has to do with the attitudes and needs of the English users themselves, to whom they are accountable. As Elder and Davies (2006) point out, local non-standard varieties of English have strong appeal in identity terms, but are often stigmatised by the users themselves who prefer to learn and be tested in high prestige varieties. As a case in point, an early attempt by Brown and Lumley (1998) to develop a test which is sensitive to local norms and uses of English has never been used. This test was commissioned by SEAMEO to determine the proficiency of teachers of English in Indonesia and was designed to reflect local classroom realities by selecting texts from local English-medium sources on topics of local relevance, and by using non-native (Indonesian) raters as assessors (Hill, 1996). It seems that other large-scale tests, not designed for the measurement of teacher proficiency, continue to have greater currency in Indonesia. Similar problems have been faced in Hong Kong, where the GSLPA, a context-sensitive exit test which mirrors the demands for English in the Hong Kong employment context (Lumley and Qian, 2003) has long struggled for due recognition because tests from the inner circle are often viewed by local stakeholders as “the gold standard”.

It is perhaps for these same reasons that, while there are increasing numbers of locally developed and administered tests of English like the College English Test (CET) in China (Zheng and Cheng, 2008), all of them, to our knowledge, draw on SE as their reference point, even though the test takers share the same L1 and could feasibly be assessed according to local English norms (for a recent description of the particular features of China English see He and Li, 2009). It seems however that the test developers in countries such as China are as concerned as the test takers to align themselves with prestige varieties of English and perhaps view such norms as inferior or at least less useful or portable than those of SE. So, while there are strong arguments for developing tests which are sensitive to local contexts, test users in those contexts are often the first to reject them for a range of reasons, including the desire to move beyond the context they are studying or working in. Language testers have a responsibility to counsel stakeholders about inappropriate test use, but we are not always in a position to dictate policy. We must also be wary, as Taylor (2006) has cautioned, of dismissing stakeholders' views as uniformly unenlightened.

CONTRIBUTIONS

Thus far we have outlined some of the issues that may have constrained language testers in embracing the challenges of EIL assessment and, in so doing, question the basis for some of the criticisms made by C and S. In this section we will point to some areas of language testing research which not only demonstrate language testers' awareness of sociolinguistic variation and its implications for the design of assessment tasks, but also, we would argue, contribute to a better understanding of EIL more generally.

Firstly, with regard to the use of L2 accents in listening test input, a program of research has emerged over the past ten years which was initiated with funding from the Educational Testing Service (ETS) – which produces the TOEFL test – and gave rise to two studies which investigated the effect of accents on listening test scores (Major et al., 2002; Major et al., 2005). Llurda (2004) has noted that the existence of such a research program signals a shift in language testing towards EIL concerns, acknowledging both “the existence of the huge number of non-native English speakers and the need to incorporate their voices into mainstream English language teaching and language testing” (Llurda, 2004, p. 315). While Major et al.'s (2002; 2005) studies were not conclusive, more recent research by Harding (2008) has found that a shared-L1 or familiarity effect is not pervasive on a test when highly-intelligible (though characteristically accented) L2 speakers are used, but that these effects may emerge when certain conditions are present

on a listening test relating to the types of listening skills that items are targeting (i.e., bottom-up or top-down), particular characteristics of a speakers' pronunciation and the linguistic demands of the text. These findings provide the basis for several proposed approaches to using speakers with L2 accents in listening assessment without compromising test designers' fundamental concerns for construct validity and fairness.

Another area attracting increasing attention in the field of language testing is whether native speakers of English, who have been traditionally responsible for scoring the performance of L2 test-takers (Lazaraton, 2005; Lowenberg, 2002; Seidlhofer, 2001), should continue to be the exclusive "norm makers" given that they are far outnumbered by non-native users of English (Graddol, 1997) for whom SE norms may have limited relevance. A number of studies have been conducted comparing rater behaviour across native- and non-native speaker groups with a view to discovering whether they rate equally consistently and whether they apply different standards in judging performance. For example, a recent study by Kim (2009) comparing evaluations of oral test performance by native English-speaking (NES) teachers from Canada and non-native English speaking (NNES) teachers from Korea, shows high levels of agreement between the two rater groups in terms of scores, but marked differences in the criteria applied in judging performance. These findings are echoed in a subsequent study involving NES and NNES teachers of English in China (Zhang and Elder in press) judging the speaking proficiency of Chinese learners on the CET-SET, a group oral test administered to around 100,000 test-takers annually (Zhang & Elder 2009). While such studies leave unanswered the practical question of who is better equipped to act as rater in any given situation, they provide empirical evidence to support what EIL/ELF researchers have long asserted, namely that there are no a priori grounds for privileging native speakers of English as judges. Moreover these qualitative investigations of the criteria invoked by non-native raters from a range of L1 backgrounds when judging L2 learner performance have the potential to illuminate the alternative (non SE) norms which may underlie communication in a range of EIL contexts. Of course, it cannot be assumed that alternative norms will emerge. In the Zhang & Elder study, the NNES raters proved to be more prescriptive and more focused on formal (SE) accuracy than the NES.

Finally, C & S's call for assessment of intercultural communication skills has, we would argue, already been addressed with the increasing use of pairwork (i.e. peer-to-peer interaction between learners from different L1 backgrounds) in the assessment of second-language speaking. This format is now being used for a number of high stakes tests including the Certificate of Proficiency in English (Weir and Milanovic, 2003; Macqueen and Harding, 2009). A forthcoming special issue of the journal *Language*

Testing brings together the most recent research in this area. Much of this research suggests that this format can overcome many of the limitations of the traditional one-on-one interview between the L2 learner and a native speaker examiner by removing the power imbalance and encouraging more complex interaction of the kind likely to occur in many EIL contexts (Brooks, 2009). The traditional view of paired assessment has seen variability of interlocutors (on factors such as gender, personality and proficiency) as a potential threat to fairness, with the risk that one interlocutor may dominate the communicative exchange at the expense of the other (e.g. see Berry, 2007; Iwashita, 1998; O'Sullivan, 2000). However, a study by Davis (2009) has shown that differences in proficiency do not necessarily prevent candidates from engaging in collaborative interaction. A further paper by May (2009), reports on the reactions of raters to instances of such negotiated interaction, concluding that key features of the communication are regarded as jointly constructed rather than individual achievements. In other words, the variability formerly seen as a threat to fair measurement is viewed by these raters as an integral part of the test construct. Successful handling of this kind of variability, as Taylor and Wigglesworth (2009) point out, requires the same kind of flexibility and sensitivity that characterise effective communication in real world encounters between non-native users of English. A further rater study by Ducasse and Brown (2009), investigates how these features of collaborative interaction can be incorporated into a rating scale used for assessing pairwork in classroom contexts. It is clear then that this kind of assessment has potential for use in both classroom and more formal assessment contexts, where the purpose of the test is to measure intercultural communicative competence.

CONCLUSION

While this outline is by no means comprehensive, the three strands of research sketched above show us that language testers are grappling with issues that are similar to those which preoccupy EIL researchers, and are making contributions which are relevant not only for assessment but also for the broader field of EIL research. Having said that, the field of language testing is steeped in the tradition of psychometrics, and, as language testing practitioners ourselves, we can attest to the fact that there will not be a revolution in language testing with respect to embracing EIL. Language testing is, after all, often concerned with making decisions that can affect people's lives. As with any other serious area of policy-making, changes to language testing policy must be evidence-based, and may evolve slowly in response to changes in social mores.

While we endorse the general thrust of C and S's article (as, we believe, would many in the language testing community who, after all, are mostly applied linguists themselves) we would also urge those working with EIL from outside the testing field both to engage more seriously with the concerns and ethical responsibilities of testers in order to better understand the constraints they work within, and also to acquaint themselves with current developments in language assessment which can offer an important source of empirical evidence and insight on some of the issues which most interest EIL researchers.

ENDNOTES

- ¹ The sentence – “Furthermore, the contents of language tests ought to correspond with the functions for which the testees will employ the language (see Zafar Khan 2009)” appears to be a comment on content validity in general, rather than a specific remark on the relationship between testing and English as an International Language.

REFERENCES

- Berry, V. (2007). Ethical considerations when assessing oral proficiency in pairs. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds), *Current developments and alternatives in language assessment – Proceedings of LTRC 96* (pp. 107–123). Jyväskylä: University of Jyväskylä and University of Tampere.
- Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, 26 (3), 341–366.
- Brown, A.; Lumley, T. (1998). Linguistic and cultural norms in language testing: A case study. *Melbourne Papers in Language Testing*, 7 (1), 80–96.
- Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, 3, 229–242.
- Chapelle, C. A.; Enright, M. K.; Jamieson, J. M. (Eds) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.
- Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26 (3), 367–396.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- Ducasse, A.; Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26 (3), 423–444.
- Elder, C.; Davies, A. (2006). Assessing English as a lingua franca. *Annual Review of Applied Linguistics*, 26, 282–304.
- Graddol, D. (1997). *The future of English? A guide to forecasting the popularity of the English language in the 21st century*. London: British Council.

- Harding, L. (2008). *The use of speakers with L2 accents in academic English listening assessment: A validation study*. Unpublished PhD dissertation, The University of Melbourne.
- He, D.; Li, D. (2009). Language attitudes and linguistic features in the China English debate. *World Englishes*, 28 (1), 70–89.
- Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing*, 5 (2), 29–50.
- Iwashita, N. (1998). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 5 (2), 51–65.
- Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal*, 60 (1), 42–50.
- Jenkins, J. (2007). *English as a lingua franca: attitude and identity*. Oxford: Oxford University Press.
- Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26 (2), 187–217.
- Llurda, E. (2004). Non-native-speaker teachers and English as an International Language. *International Journal of Applied Linguistics*, 14 (3), 314–323.
- Lazaraton, A. (2005). *Non-native speakers as language assessors: Recent research and implications for assessment practice*. Paper presented at the BAAL, Bristol.
- Lowenberg, P. (2002). Assessing English proficiency in the Expanding Circle. *World Englishes*, 21 (3), 431–435.
- Lumley, T.; Qian, D. (2003). Assessing English for employment in Hong Kong. In C. A. Coombe and N. Hubley (Eds), *Assessment practices: Case studies in TESOL. Practice Series* (pp. 135–147). Alexandria, VA: TESOL.
- Macqueen, S.; Harding, L. (2009). Review of the Certificate of Proficiency in English (CPE) Speaking test. *Language Testing*, 26 (3), 467–476.
- Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36 (2), 173–190.
- Major, R. C.; Fitzmaurice, S. F.; Bunta, F.; Balasubramanian, C. (2005). Testing the effects of regional, ethnic and international dialects of English on listening comprehension. *Language Learning*, 55 (1), 37–69.
- May, L. (2009). Co-constructed interaction in a paired speaking test. The rater's perspective. *Language Testing*, 26 (3), 397–422.
- O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, 28, 373–386.
- Sawaki, Y.; Stricker, L. J.; Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, 26 (1), 5–30.
- Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics*, 11 (2), 133–58.
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT Journal*, 60 (1), 51–60.

- Taylor, L. & Wigglesworth, G. (2009). Are two heads better than one? Pairwork in L2 assessment contexts. *Language Testing*, 26 (3), 325–340.
- Weir, C. J.; Milanovic, M. (Eds) (2003). *Continuity and innovation: The history of the CPE 1913-2002*. Studies in Language Testing 15, Cambridge: Cambridge University Press/UCLLES.
- Zafar Khan, S. (2009). Imperialism of international tests: An EIL perspective. In F. Sharifian (Ed.), *English as an international language: Perspectives and pedagogical issues* (pp. 190–205). Bristol: Multilingual Matters.
- Zhang, Y.; Elder, C. (2009). *Native and non-native teachers' evaluation of oral English proficiency in the CET-SET of China*. Paper presented at the annual meeting of American Association of Applied Linguistics, Denver, CO.
- Zhang, Y.; Elder, C. (2009, in press). Measuring the speaking proficiency of advanced EFL learners in China: the CET-SET solution. *Language Assessment Quarterly*.
- Zheng, Y. & Cheng, L. (2008) Test review: College English Test (CET) in China. *Language Testing*, 25 (3), 408–417.

Cite this article as: Elder, Catherine; Harding, Luke. 2008. 'Language testing and English and an international language: Constraints and contributions. *Australian Review of Applied Linguistics* special forum issue edited by Sharifian, Farzad; Clyne, Michael. 31 (3): pp. 34.1–34.11. DOI: 10.2104/aral0834.