

10 The role of the L1 in testing L2 English

Fumiyo Nakatsuhara, Lynda Taylor, Suwimol Jaiyote
University of Bedfordshire

1 Introduction

A number of chapters in this volume have highlighted the role of testing and assessment in contributing to maintaining monolith conceptualisations of English, and conversely the hybrid nature of the language and lack of clear boundaries, especially in L2 versions. Accordingly, this chapter discusses the role of the L1 in assessing L2 English proficiency, focusing particularly on tests of L2 English speaking ability. It firstly considers Weir's (2005) socio-cognitive framework for developing and validating speaking tests (further elaborated in Taylor, 2011), in order to locate the issue of L1 influence comprehensively within an overall test validation framework. It then describes how the different test purposes for which tests are designed can determine the role of the L1, according to the specific construct of the test. To exemplify differing roles that can be played by the L1 in speaking tests, this chapter then presents two studies on L2 spoken English tests which sought to address the issue of test-takers' L1 in contrastive ways.

The first study explored the impact of test-takers' L1 backgrounds in the paired speaking task of a standardised test of general English provided by an international examination board (Nakatsuhara and Jaiyote, 2015). The key question in the research was how we can ensure fairness for test-takers who perform paired tests in shared and non-shared L1 pairs. The second piece of research is a test validation study conducted as a part of the development of a new English for Academic Purposes (EAP) test administered by a national examination board, targeting a monolingual population of learners who share a single L1 (Nakatsuhara, 2014). Of particular interest is the way in which its pronunciation rating scale was developed and validated for the single L1 context.

In light of these examples of research into international and locally-developed tests, this chapter aims to demonstrate the importance of the construct of a test and its score usage when considering what Englishes (rather than 'standard' English) should be elicited and assessed, and when/how we can reconcile notions of 'standard' English with local language norms and features without undermining the validity of a test or risking unfairness for test-takers. In so doing, this chapter reiterates the point made by Harsch (this volume) concerning the significance of establishing a transparent test construct with due considerations to the specific context of each test, leading to the best possible way to benefit learners by appropriately selecting the variety(ies) of English against which their 'l-language' development (Hall, this volume) should be assessed.

2 Test validity framework

Since Messick's (1989) seminal paper on a unitary conceptualisation of test validity, there has been a general consensus among language test researchers and practitioners that test validity concerns "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of inferences

and *actions* based on test scores or other models of assessment” (Messick, 1989, p. 13, italics in original). According to this understanding, validity is not a property of the test itself, but rather a concept pertaining to the meaning, interpretations, and inferences made on the basis of test scores. Messick (1989, p. 20) explains the centrality of construct validity and the significance of taking social dimensions into account within his unified theory of validity.

Building on Messick’s validity conceptualisation, Weir (2005) proposed a socio-cognitive framework for test development and validation, which is now widely recognised as a sound and comprehensive framework on the basis of which validity judgements can be made more confidently (e.g. Geranpayeh and Taylor, 2013; Khalifa and Weir, 2009; O’Sullivan and Weir, 2011; Shaw and Weir, 2007; Taylor, 2011). The framework pays special attention to the cognitive processing theory that underpins equivalent operations in real-life language use, and it views the use of language in performance tasks as a social rather than a purely linguistic phenomenon. In the socio-cognitive framework for validating speaking tests illustrated in Figure 10.1, Weir (2005; further elaborated in Taylor, 2011) proposed six distinguishable elements, for which we need to generate evidence to support a meaningful validity argument. These are *test-taker characteristics*, *context validity*, *cognitive validity*, *scoring validity*, *consequential validity* and *criterion-related validity*. They are briefly explained as follows (cf. Weir, 2005, pp. 48-49):

- *Test-taker characteristics* concern how the physical/physiological, psychological, and experiential characteristics of candidates are catered for by a test, and whether the test is likely to be appropriate for the target candidates;
- *Context validity* concerns to what degree the test tasks can be judged as being capable of eliciting language under appropriate linguistic and task-based performance conditions that are relevant to and representative of the real-life construct that the test is intended to measure;
- *Cognitive validity* concerns the extent to which the cognitive processes required to complete the tasks are shown to be processes that correspond to the intended underlying theoretical construct of language ability, as well as the extent to which candidates are likely to use the same cognitive processes as they would if performing the same task in a ‘real world’ context;
- *Scoring validity* concerns to what extent we can depend on the scores from the test being consistent, reliable, and generalisable in a non-test target language use context, what the scores or grades mean, and whether the relationship between task performance and the rating criteria used to assess the performance is appropriate;
- *Consequential validity* concerns the degree to which test scores are interpreted and acted upon in the intended way, and the extent to which the test has intended or unintended consequences associated with the washback effect on teaching and learning and the impact on society;
- *Criterion-related validity* concerns the extent to which the empirical relationship with external sources (i.e. different versions of the same test, other tests with an identical

construct, or the future performance of candidates) supports the way the score is meant to be interpreted.

Echoing Messick (1996, p. 253), Weir (2005) notes that these different components should be seen as complementary forms of validity evidence. The validity of a test relates to all these aspects, and the interpretation of evidence relying on just one aspect in isolation fails to treat validity as a whole.

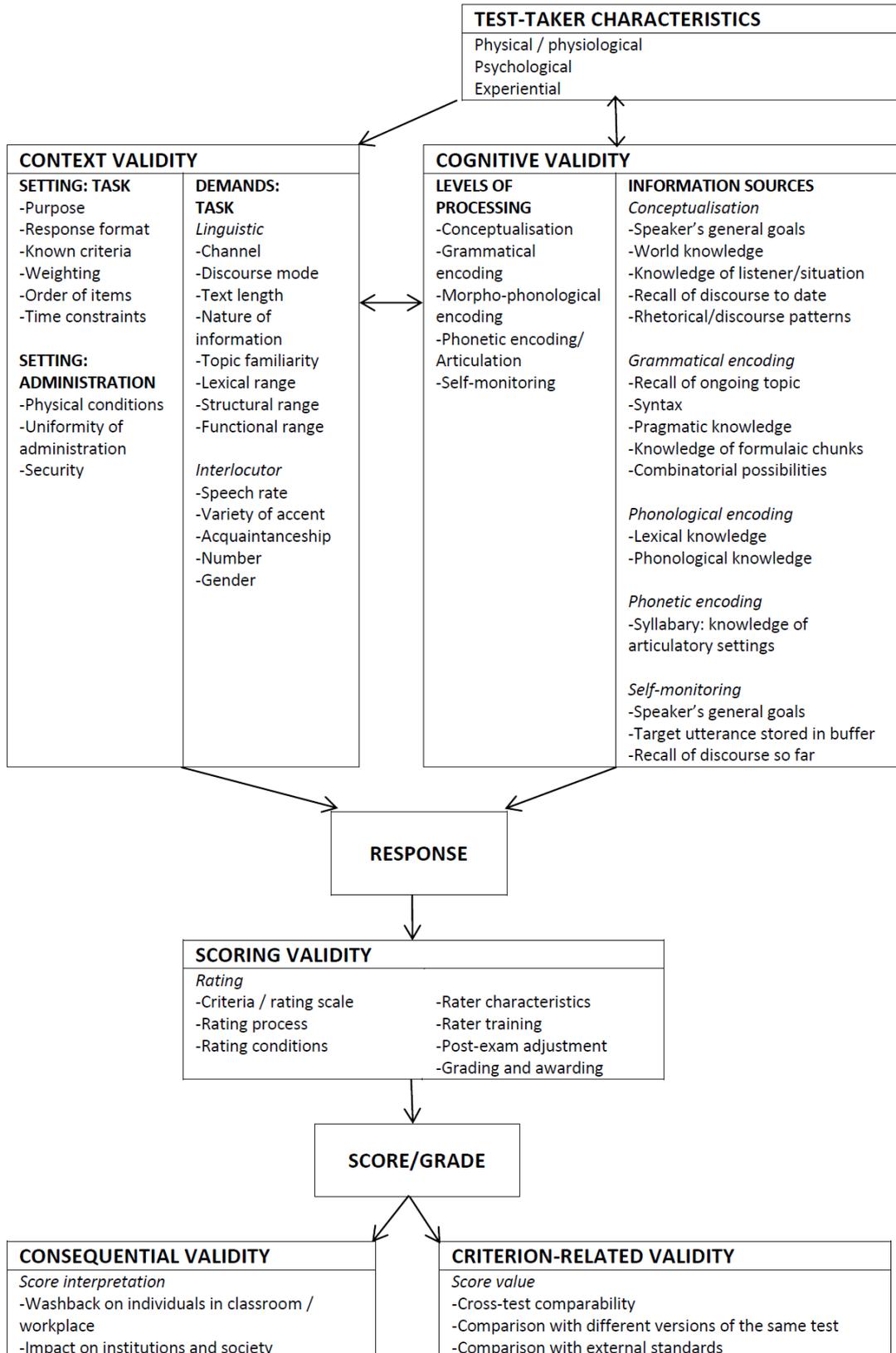


Figure 10.1: Weir's socio-cognitive framework for validating speaking tests (Taylor, 2011, p. 28)

When evaluating the role of the L1 in an L2 speaking test, we need to take all these validity components into consideration, since the L1 issue could potentially influence all parts of the framework. We will now briefly discuss the relationship between the L1 and each component of the socio-cognitive framework, while making some links to Hall's (this volume) ontological framework of English whenever possible.

Test-takers' own L1 background belongs to the top box of the framework, *test-taker characteristics*, which will affect the *cognitive* processes that the test-takers will engage while performing test tasks. For example, grammatical, phonological, and phonetic encoding stages could be to some degree influenced by test-takers' L1 transfer. These two aspects of the socio-cognitive framework could relate to what Hall calls 'l-language resources', since both *test-taker characteristics* and *cognitive* parameters are about features pertaining to individual learners and their internal cognitive capacity.

If the test format involves interaction between an examiner and a test-taker or between peer test-takers, then the 'interlocutor' aspect of *context validity* requires careful attention in terms of the L1 background of interlocutors. The selection of such contextual parameters can be considered as specifying the types of 'linguaging' we would like to assess in a test.

Test-takers' spoken output, which is in the expressive domain of one's language, is usually assessed by either an examiner who also acts as an interlocutor or a rater who listens to or watches recorded performances of test-takers; as a result, how familiar the examiners and/or raters are with the test-takers' variety of English and how they are trained to assess the performances under *scoring validity* become critical.

Furthermore, consideration for *consequential validity* should also be made, since the resulting scores then have to be interpreted to inform the users of the test score in a way that is appropriate to specific test purposes. Whether and the extent to which the test facilitates positive washback to teachers and learners and provides positive impact on educational systems and society are also key considerations. Finally, since a test is often used to predict learners' future performance in a real-life context, the predictive power of a test and its relation to the role of the L1 needs to be considered under *criterion-related validity*.

The centrality of the construct of a test and the use of test scores (Messick, 1989) means that all parameters relating to the role of the L1 in an L2 speaking test must accord with what the test is designed to measure and what score interpretations are anticipated. Therefore, while advances in research on World Englishes and English as a Lingua Franca are welcomed by language testing researchers and test providers, the actual decision on the treatment of L1-related issues and a variety of Englishes in a language test will need to be made on a case by case basis, depending on the construct of the specific test in question and its intended score interpretations. In doing so, all the validity parameters described above have to be given due consideration.

3 Test validation studies addressing L1-related issues

To exemplify how differing L1 roles can be operationalised in actual language tests, we will now present selected parts from two validation studies on two speaking tests which have clearly different test constructs and different purposes to serve for different target test-taker groups, highlighting the differing roles played by the L1 in these two tests.

The first study described in Section 3.1. is taken from Nakatsuhara and Jaiyote (2015) on the *B2 First* examination (formerly known as First Certificate in English or FCE) developed and delivered by Cambridge Assessment English. *B2 First* has a very large international candidature cohort and is one of the most popular General English tests offered by Cambridge English which administers over 5 million English tests annually in more than 130 countries (Cambridge English, 2018a). For example, *B2 First* was administered in 94 countries in 2016 (Cambridge English, 2018b).

The second study illustrated in Section 3.2. is extracted from Nakatsuhara (2014) on the *Test of English for Academic Purposes (TEAP)*, which is a relatively new admissions test for Japanese colleges and universities. The test was taken by 24,434 test-takers in 2017, and it is expected that *TEAP* will be recognised by all national universities when the current National Center Test for university admissions (with 550,000 candidates in 2017; National Centre for University Entrance Examinations, 2017) is phased out from 2020 (Tanaka, 2018). Therefore, a significant increase of the test-taker numbers is expected in the next few years.

As such, *B2 First* is a General English test for a multilingual population of international learners, while *TEAP* is an English for Academic Purposes (EAP) test targeting a monolingual population of learners who share a single L1.

3.1 *B2 First*: English test for international test-takers

The focus of Nakatsuhara and Jaiyote's (2015; see also Jaiyote, 2016 for more details) study was the paired speaking task (Part 3: collaborative task) of *B2 First*. The research investigated the extent to which test-takers having a shared or non-shared L1 partner can affect their performance on the task. More specifically, this research addressed three different aspects of test validation in relation to the role of the L1 in a paired speaking format: fairness in test scores, listening demands imposed by an interlocutor, and communication patterns. Firstly, since *B2 First* is a large-scale, high-stakes test, "issues of quality and fairness must be paramount" (Taylor, 2006, p. 56). The test construct should be comparable in both shared and non-shared L1 pairings, and test-takers should not be advantaged or disadvantaged in terms of awarded scores due to their partners' L1 backgrounds. Secondly, it has been demonstrated that interactive speaking tests tap into both speaking and listening constructs to some degree, because test-takers' listening proficiency is inevitably required in order to respond to their partner appropriately (e.g. Nakatsuhara, 2012; Seedhouse and Egbert, 2006). It is necessary therefore to examine the extent to which the two types of pairing are comparable in terms of the listening demands imposed by the paired partner. Lastly, it is important to investigate how comparable communication patterns are between the two types of pairing. In Jenkins' (1997) study, in which she observed paired discourse in *C1 Advanced (CAE)*, it was found that test-takers with a shared L1 partner deliberately used more L1-influenced pronunciation features to make their utterances mutually intelligible. As her study was relatively small-scale, this

study also aimed to re-examine whether and to what extent speech accommodation and any other salient communication patterns can be identified in the two types of pairing and what roles the learners' listening proficiency may play in communicating with shared and non-shared L1 partners. In terms of locating the focus of this study within the socio-cognitive validation framework, it relates to *scoring validity* and the 'interlocutor' component of *context validity* in addition to the test-taker's own *test-taker characteristics*.

The research questions for the study were as follows:

RQ 1: Are there any differences in speaking test scores when test-takers are paired with a shared L1 speaker as compared with a non-shared L1 partner?

RQ 2: Are there any differences between shared and non-shared L1 pairs in terms of the impact of their listening proficiency on performance in paired oral tests?

RQ 3: What are the similarities and differences in communication patterns between shared and non-shared L1 pairs?

3.1.1 Methodology

Forty pre-sessional English programme students at a UK university participated in the study, of which 20 were Thai L1 speakers and 20 were Urdu L1 speakers. To minimise the potential effects of other test-taker characteristics, the participants' test-taker characteristics other than L1 background were controlled as much as possible. 10 males and 10 females from each L1 background were recruited, and only single-sex pairs were formed. Most of the participants were in their 20s (Mean=27.20, SD=2.84), and their speaking and listening proficiency levels were comparable according to their recent IELTS Band scores (Speaking: Mean=5.61, SD=0.35, Listening: Mean=5.28, SD=0.39). A demographic questionnaire also confirmed the test-takers' perceived familiarity with the English spoken by shared and non-shared L1 speakers with Likert-scale questions. Unsurprisingly, they were significantly more familiar with the English spoken by the same L1 speakers than the English spoken by the other L1 speakers (i.e. Thai or Urdu).

All 40 learners were asked to take four tests: a listening test, two paired speaking tests, and a monologic speaking test. The listening test was to assess the learners' general listening ability, using 30 items taken from *B1 Preliminary (PET)* and *B2 First* practice materials (Cambridge English, 2008, 2009). The reliability of the listening test was acceptable, showing a Cronbach alpha value of 0.91. All participants took two paired speaking tests, one with a shared L1 partner and one with a non-shared L1 partner. Two paired speaking tasks were selected from *B2 First* practice materials (Cambridge English, 2009), and the two tasks were used in a counter-balanced order between the shared and non-shared L1 conditions. The 40 learners also took a monologic speaking test. This was to obtain their baseline performance data without any influence of a paired partner. Both the paired and monologic speaking tests were video-recorded and double-marked by two trained examiners. The paired performance was assessed using the four rating criteria of *B2 First* (i.e. *Grammar and vocabulary*, *Discourse management*, *Pronunciation*, and *Interactive communication*), and the monologic performance was assessed using the same rating scales except for *Interactive*

communication. Inter-rater reliability between the two raters was relatively high, with Pearson correlations ranging from 0.89 to 0.93. The absolute agreement rate of the two raters was 55.5% (243 of the total 440 score points in total), and the remaining 44.5% (197 scores) fell within one-point difference.

After each paired speaking test, a stimulated recall interview (Gass and Mackey, 2000) was carried out with each of the paired test-takers, to gain insights into the interaction in both shared and non-shared L1 pairings. Using a video-recording of test performance as a stimulus, they were asked to explain and elaborate on their communicative behaviours to help the researchers to interpret the salient communicative patterns of each pairing.

3.1.2 Results and analysis

The listening test scores, and paired and monologic speaking test scores were statistically analysed to address RQ1 and RQ2. First, non-parametric Wilcoxon Signed Rank tests were used to examine differences in paired speaking scores between shared and non-shared L1 pairs (RQ1). As presented in Table 10.1, the results show that there was no statistically significant difference for any of the analytic categories in the two types of pairing. This suggests that the type of pairing does not affect the test-takers' paired test scores.

Rating category (1-5 points)	Pair type	Median	Mean	SD	Wilcoxon
<i>Grammar and vocabulary</i>	shared L1	3.50	3.36	1.12	Z=-.12, p=.90
	non-shared L1	3.50	3.36	1.21	
<i>Discourse management</i>	shared L1	3.50	3.13	1.16	Z=-1.90, p=.06
	non-shared L1	3.50	3.45	1.12	
<i>Pronunciation</i>	shared L1	3.00	3.19	1.16	Z=-.23, p=.82
	non-shared L1	3.00	3.19	1.10	
<i>Interactive communication</i>	shared L1	3.50	3.23	1.20	Z=-.81, p=.86
	non-shared L1	3.25	3.16	1.28	

Table 10.1: Comparison of paired speaking-test scores between the shared and non-shared L1 pairs (N=40)

To address RQ2, Spearman correlations were performed to examine the strength of the correlations between the listening test scores and the analytical scores of monologic and paired speaking tests in the whole group, and in shared and non-shared L1 pairs separately. Table 10. 2 shows that while none of the correlations between listening and monologic test scores was statistically significant, positive significant correlations were found between listening and paired speaking scores for *Grammar and vocabulary* ($\rho=0.32$, $p=0.04$) and for *Discourse management* ($\rho=0.35$, $p=0.03$). Although the strength of these correlations was only medium (Cohen, 1998), there seems to be a positive relationship between the learners' listening proficiency and their paired speaking performance displayed for *Grammar and vocabulary* and *Discourse management*.

	Grammar and vocabulary	Discourse management	Pronunciation	Interactive communication
--	------------------------	----------------------	---------------	---------------------------

Speaking test mode	Mono	Paired	Mono	Paired	Mono	Paired	Mono	Paired
Spearman's rho	.19	.32	.13	.35	.19	.25	-	.08
Sig.	.25	.04	.44	.03	.24	.13	-	.63

Table 10. 2: Correlations between listening and speaking scores (N=40)

The same correlational analyses were then carried out for shared and non-shared L1 pairs in the paired test. Interestingly, the identical results were found only for non-shared L1 pairs, but not for shared L1 pairs. As reported in Table 10.3, no significant relationship was found between test-takers' listening and speaking proficiency displayed in paired tests when they were paired in shared L1 pairs. However, their listening proficiency seemed to matter in non-shared L1 pairs for their *Grammar and vocabulary* ($\rho=0.37$, $p=0.02$) and *Discourse management* ($\rho=0.38$, $p=0.02$) scores. That is, although the strength of correlations was only medium, the higher the learners' listening scores were, the higher *Grammar and vocabulary* and *Discourse management* scores they received when they were paired with a different L1 partner.

Pair type	Grammar and vocabulary		Discourse management		Pronunciation		Interactive communication	
	shared L1	non-shared L1	shared L1	non-shared L1	shared L1	non-shared L1	shared L1	non-shared L1
Spearman's rho	.26	.37	.26	.38	.22	.22	-.01	.14
Sig.	.10	.02	.11	.02	.17	.18	.97	.37

Table 10. 3: Correlations between listening and paired speaking scores for shared and non-shared L1 pairs (N=40)

To address RQ3, all video recordings of paired speaking performance were transcribed following Conversation Analysis (CA) conventions (Atkinson and Heritage, 1984). CA is a type of discourse analytic approach which Harsch (this volume; cf. also Eskildsen, this volume) also suggests as useful to gain insights into actual 'languaging' instances, and it was carried out to explore communication patterns in the paired tests which were related to test-takers' listening abilities and their L1s. Data from retrospective interviews with test-takers and raters was used to support and elaborate on the researchers' interpretation of the CA analysis. After salient communication patterns were identified in both shared and non-shared L1 patterns, the transcripts were coded by two researchers independently, to count the number of occurrences of each feature and to examine them statistically between the two pair groups. Due to space limitations, only one main difference between shared and non-shared L1 pairs is presented below.

The number of communication breakdowns, the test-takers' attitude towards repairing communication breakdowns, and their success rates seemed different between the two types of pair. Among the entire transcripts of 20 shared and 20 non-shared L1 interactions, 12 communication breakdowns were observed in shared L1 pairs and an attempt to solve

the breakdown was made for all 12 cases, all of which were successful. On the other hand, non-shared L1 pairs encountered as many as 25 communication breakdowns, and an attempt to solve the breakdown was however observed only for 19 cases (76%), of which 15 cases (60%) were successful. That is, non-shared L1 pairs had more than twice as many communication breakdowns as shared L1 pairs, and while 100% of the communication breakdowns were repaired in shared L1 pairs, only 60% were repaired in non-shared L1 pairs. An example of a non-attempt to solve a communication problem in a non-shared L1 pair is exemplified in Excerpt 1 (U08 is an L1 Urdu speaker and T08 is an L1 Thai speaker; see the line with an arrow).

Excerpt 1: Non-attempt to repair communication breakdown

- U08: er: to become ah:: successful artist it-it basically depend upon er:: (.)
personality ... for person who is too shy [to come and=
T08: [mm:: ((frowning))
U08: =perform in front of many people and especially in these days
→ T08: mm:: ((frowning)) ok my my turn (0.3) .hhh in this picture

Related to the increased level of understanding in shared L1 pairs, it seems that shared L1 pairs understood each other even when a partner's utterance was somewhat confusing - in other words, non-target. For instance, in a Thai L1 pair in Excerpt 2, T19 used the verb "make" instead of the verb "do". Nonetheless, T20, without even making a clarification request, understood what T19 intended to say, and T20 continued the conversation. In T20's stimulated recall interview, T20 reported, "*I knew what she meant. In my language, 'make' and 'do' have the same meaning.*", indicating that T20's familiarity with the specific variety of English helped her understand T19. Echoing May (2011), the two trained raters in this study also reported that test-takers from the same L1 background seemed to understand each other well, even when they had difficulty in comprehending the talk.

Excerpt 2: Understanding an utterance with a non-target element

- T19: what do you think about (0.7) er::: girlfriend and (.) boyfriend make
something together?
→ T20: i think for girls, they like maybe share feeling or talk something

It was also noted that misunderstandings occurred in non-shared L1 pairs because of the different cultural backgrounds of the test-takers. Excerpt 3 illustrates miscommunication between U10 (Urdu) and T10 (Thai). T10 suggested that U10 should organise a party to make new friends. U10 then imagined that alcohol would need to be offered at a party, which is not compatible with U10's religious beliefs. In U10's stimulated recall interview, he said, "*I am Muslim and our religion doesn't allow us to drink alcohol. I tried to tell my partner about it.*" However, U10 explained the cause of his unwillingness to organise a party only very implicitly, referring to his lack of skills in arranging parties. T10 did not understand his partner's hidden problem with a party, and proposed that they should move on to a new topic. This example highlights the importance of explicit explanation for successful communication in non-shared L1 pairs.

Excerpt 3: Misunderstanding due to different cultural backgrounds

- U10: i only have two or three friends
T10: you- you can do parties if you want to make a lot of friends
→ U10: actually problem's that i'm not good at party ha ha ha
. .
. ((awkward conversation for 16 lines))
. .
T10: so let's go to the next...

Interestingly, there was no instance of speech accommodation as observed in Jenkins's (1997) study. There were two occurrences where two Thai test-takers used L1-influenced back-channelling and a L1 word during the paired test interaction, but both instances were observed in non-shared L1 pairs, and they reported in retrospective interviews that they unconsciously inserted the L1 back-channelling and L1 word.

To summarise the results, there seemed some differences in the construct measured between shared and non-shared L1 pairs. While the construct under the non-shared L1 condition included listening, listening proficiency did not seem to matter under the shared L1 condition. More communication breakdowns and misunderstandings were observed by non-shared L1 pairs, and it seemed harder to repair them. Nevertheless, fairness to test-takers in both pairs was retained, as test scores were not different between the two conditions. These findings will be revisited and interpreted in the Discussion and Conclusions section below in relation to the role of L1 in international tests.

3.2 TEAP: English test for single L1 test-takers

As noted earlier, *TEAP* is a university admissions test in Japan, aiming to assess English skills required by Japanese students to study at the university level in Japan. The *TEAP* Speaking rating scales were therefore developed considering the construct and the use of test scores in this particular context. This section will focus only on its *Pronunciation* scale, and report on how the rating descriptors to be used to assess the single L1 group, Japanese, were developed and validated. Hence, the focus is the *scoring validity* of the *TEAP* Speaking test in relation to the role of L1 in this test.

As was decided from the outset of the project, the CEFR (Council of Europe, 2001) played a central role in designing the *TEAP* test, while making it relevant to the specific Japanese context. To this aim, the test design team built on the CEFR performance descriptors with more descriptors specific to target language use (TLU). In addition to the *Course of Study for High Schools* published by Japanese government (MEXT, 2008), various rating scales that were developed for Japanese learners of English such as the Standard Speaking Test rating scales (ALC, 2006) and Kanda English Proficiency Test rating scales (Bonk and Ockey, 2003) informed drafting of the *TEAP* rating scales. Once an early version of the rating scales was drafted, the scales were discussed iteratively between the three test development partners, and based on these discussions, draft rating scales to be used for a pilot study were prepared. The draft scales contained five analytical categories, *Grammatical range and accuracy*, *Lexical range and accuracy*, *Fluency*, *Pronunciation*, and *Interactional Effectiveness*,

each of which had four levels (0=below A2, 1=A2, 2=B1, 3=B2). Of the five scales, this section focuses only on *Pronunciation*. Since the *TEAP* analytic scales are not publicly available, pronunciation descriptors included in the *TEAP* holistic scale are extracted in Table 10.4.

B2	Speech easy to understand; accurate stress and intonation; some L1 influence on individual sounds.
B1	Speech intelligible; noticeable L1 influence on stress, intonation, and individual sounds.
A2	Speech mostly intelligible; heavy L1 influence on stress, intonation, and individual sounds; some mispronunciations impede communication.
Below A2	No response OR often unintelligible.

Table 10. 4: Pronunciation descriptors in the *TEAP* holistic scale (Eiken Foundation of Japan, 2014)

The research questions relevant to the *Pronunciation* scale in this round of *a priori* validation study were as follows.

RQ1: Is there any evidence from test-takers’ output language that validates the descriptors used to define the levels on each rating scale?

RQ2: How well is the scoring validity of the *TEAP* Speaking tests supported by the three-facet Multi-faceted Rasch Model (MFRM) analysis (i.e. test-taker, rater, and rating category)?

3.2.1 Methodology

Twenty-three first-year Japanese university students and three trained raters participated in this study. The three raters were experienced English native speaker teachers at Japanese universities. They all attended a rater training session prior to the pilot test.

All 23 students took a trial version of the *TEAP* Speaking test, and all performances were video-recorded, which were then rated by the three trained raters using the draft rating scales. As a method to validate the descriptors used to define the levels on each rating scale, the test-takers’ actual speech samples were transcribed and analysed linguistically to address RQ1. Previous studies have employed this approach to validating rating scale descriptors, including Brown (2006), and Iwashita et al. (2008). A variety of linguistic measures were selected to reflect the features of performance relevant to the wording within the draft analytical rating scales, so as to investigate whether these measures differ in relation to the proficiency levels of the test-takers assessed using the rating scales. After that, test scores were analysed using MFRM analysis with the FACETS program. The FACETS analysis was carried out with three major facets for the score variance in this study: test-takers, raters, and rating categories. This analysis can identify inconsistencies in test-takers’ rating scores, rater severity, and differences in the difficulty levels across the five rating categories.

After the three raters had completed the rating of all recorded performances, they were invited to take part in a focus group session. During the focus group, the raters watched videos of three test-takers once again, and the videos were paused after each task to allow for discussion. The focus group was designed to elicit the raters' reasons for awarding the scores they had given, to provide insights into the rating process(es) and inform the way in which the results from the linguistic and statistical analyses are interpreted.

3.2.2 Results and analysis

Key assessment areas specified in the draft *Pronunciation* scale were intelligibility, prosodic features such as intonation, rhythm, word/sentence stress, assimilation/elision, and L1 influence. In practice, pronunciation was the hardest category to quantify. Iwashita et al. (2008) employed measures of phonology using specialists' judgements on different phonological features. Brown (2006) dropped phonology analysis because of the difficulty of measurement. Post (2011) used acoustic analysis software to analyse pronunciation features, which is the most accurate way of quantifying pronunciation features. However, it involves extremely labour-intensive work in segmenting and making judgements on each phoneme. Iwashita et al.'s method would have been less labour-intensive, but it was not feasible for this project either due to practical constraints. Instead, for the purpose of this study, it was decided to measure only the quantity of L1-influenced (Japanese *katakana*-like) words, by counting the number of obviously L1-influenced words as a percentage of total words produced. Words spoken with noticeable *katakana*-like pronunciation such as inserting extra vowels (e.g. [dogʊ] for [dog]), all syllables evenly stressed without using [ə] or L1 influenced consonants (e.g. [ɹ] for [l], [s] for [θ]) were coded on the transcripts. Examples include:

S1-1: and: what's (.) your ah problem [pʊɹɔbʊɹemʊ] in class:.

S2-4: (.) Ah, I: enjoyed (1.5) club [kɹabʊ].

This analysis does not cover all features included in the *Pronunciation* scale, and it does not take it into account either that "not all aspects of the English phonological system are equally important for international intelligibility" (Swell, 2013, p. 428; cf. Field, 2005; Jenkins, 2000; Tsuzuki and Nakamura, 2009). However, it was thought that this measure should tap into the 'intelligibility' and 'L1 influence' aspects of the given rating scale to some extent, which were both important in validating TLU-domain specific rating descriptors. Approximately 10% of the data were co-coded by another researcher for inter-coder reliability.

Table 10.5 presents the results, showing the extent to which the analysed feature differs between the adjacent levels of the rating scales. Since there was only one student who scored 1 in *Pronunciation*, the analysis combined Levels 1 and 2 students together as one category and compared the group with the Level 3 students. The Level 3 students showed less L1 influence (Mean=1.14%) than the Level 1 and 2 students (Mean=1.38%). The results therefore indicated that the L1 influence measure exhibited a change in the expected direction, providing evidence that the rating descriptors on L1 influence are functioning in a way congruent with the test designers' intention.

Focus	Measure	Level	N	Min	Max	Mean	SD
L1 influence	Percentage of words pronounced with L1 influence	Level 1 + 2 (A2 and B1)	1 + 15	0.00	8.00	1.38	2.00
		Level 3 (B2)	7	0.00	4.00	1.14	1.35

Table 10. 5: Pronunciation features in two proficiency levels (%)

Regarding the FACETS analysis, we present only the analysis related to five rating categories, in Table 10.6. Of relevance to the focus of this section is the second column, 'Fair average' (which indicates expected average raw score values transformed from the Rasch measure). The analysis showed that the five rating categories exhibited different degrees of difficulty and these differences were statistically significant ($X^2(4)=23.5$, $p<0.005$). In particular, *Pronunciation* and *Interactional Effectiveness* were found to be easier than the other scales.

Rating category	Fair average	Measure	Real S.E.	Infit MnSq
Pronunciation	2.07	-.74	.31	1.19
Interactional Effectiveness	2.07	-.74	.30	1.06
Fluency	1.97	.08	.29	.86
Lexical Range and Accuracy	1.91	.57	.29	1.01
Grammatical Range and Accuracy	1.87	.83	.30	.75

Table 10. 6: Rating category measurement report

In the post-rating focus group discussion with the three raters, one of the topics was whether the difficulty level of *Pronunciation* and *Interactional Effectiveness* should be increased to be more aligned to the other scales. However, unlike the *Interactional Effectiveness* descriptors for which ambiguous wording was identified for revising, the raters agreed that changes to wording in the *Pronunciation* scale was not necessary. 'Intelligibility' is central to the construct of pronunciation defined in this rating scale. As Issacs (2008) noted, intelligibility is an 'evasive' concept which is hard to pin down, and a lot of the discussion with raters concerned what 'intelligible' means and how to interpret 'impeding communication' (cf. Wicaksono, this volume). Raters felt that they were able to understand the test-takers because of their familiarity with Japanese speakers' pronunciation of English. They were concerned about whether this would also apply to 'unsympathetic' or 'naïve' listeners. However, the TLU domain for *TEAP* is the EFL context in Japanese universities. Students in the context will be interacting with tutors and peer students who are obviously familiar with pronunciation features that are typical of Japanese EFL learners. Given the target context, it was felt that it was appropriate to judge students' pronunciation on the basis of raters' own ease of understanding that pronunciation. That is, while descriptors on L1 influence seemed to function well, being lenient about the impact of L1 influence on intelligibility and communication effectiveness was perfectly justified. As such, the *TEAP* Speaking test embraces the English variety spoken by Japanese speakers, reflecting the construct of the test and the usage of the test scores.

4 Discussion and conclusions

Following the description of the socio-cognitive framework for developing and validating speaking tests (Weir, 2005; Taylor, 2011), this chapter has exemplified two studies that highlight the different roles of the L1 in speaking tests. It is believed that the *TEAP* Speaking rating scale offered an example of reconciling notions of ‘standard’ English with local English norms and features. In contrast, the *B2 First* example showed the complexity in treating L1-related issues in a valid and fair manner in international examinations. Given the increased number of communication breakdowns in non-shared L1 pairings in the paired discussion part of *B2 First*, it would be indeed interesting if non-shared L1 pairs were always formed for such General English tests “to examine [test-takers’] ability to negotiate their own dialectal differences in conversation” (Canagarajah, 2006, p. 239). However, its practicality is highly questionable, since such international tests are administered all over the world, and most of the learners who take a test in their home country are likely to share the same L1. In addition, as Field (2018, p. 60) notes, “the ability to decode strongly accented speech is not a matter of adjusting immediately and flexibly to unfamiliar features, but the result of a period of exposure that is often a matter of chance”. Therefore, having only non-shared L1 pairings does not guarantee that the test can measure a more consistent construct. Some learners may happen to be familiar with the particular variety of English spoken by their paired partners, while others may not. And this is essentially impossible to control in large-scale international examination contexts. Taylor (2006, p. 58) pointed out over a decade ago that testing is “the art of the possible”. As discussed above in conjunction with the socio-cognitive framework for test development and validation, what is important is that every component of ‘the possible’ is selected in a principled and justifiable way. This is in line with Harsch’s argument (this volume) for the necessity of ‘pragmatic’ decisions that we have to make.

However, this does not mean that international language examination boards have neglected advances in research on World Englishes and English as a Lingua Franca. Most language tests and language benchmark standards no longer make reference to Native Speaker competence (Taylor, 2006: 52; see also Harsch, this volume, about recent changes made in the 2018 CEFR Companion Volume). Furthermore, many interactive speaking tests in interview, paired, and group speaking formats now have a scale to measure *interactional competence* (e.g. Young, 2011), which is “the ability to co-construct interaction in a purposeful and meaningful way, taking into account socio-cultural and pragmatic dimensions of the speech situation and event” (Galaczi and Taylor, 2018, p. 226). Example descriptors which have operationalised the construct of interactional competence include:

- Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome. (Interactive communication scale, *B2 First*)
- Fulfils the task very well; Initiates and responds with effective turn-taking; Effectively maintains and develops the interaction; Solves communication problems naturally, if any. (Communicative effectiveness scale, Trinity’s *Integrated Skills in English II*)

It is believed that these descriptors are in line with and benefited from a body of research in World Englishes and English as Lingua Franca, valuing what learners can do with their individual communicative resources.

As discussed thus far, the negotiation of the local and global dimension of the English language is a challenging task to address in international examinations (cf. Sewell, 2013). However, a new initiative taken by the British Council's Aptis test seems to suggest a possible way for international examination boards to attempt to reconcile the notions of 'standard' English with local language norms and features and to reconceptualise the role of the L1 in their tests. O'Sullivan (2011) argues for the need for 'localisation' of assessment systems when the conditions of a test are found to be appropriate. He states that localisation is appropriate especially when a test is used to make specific claims about a particular population in a particular domain or context. Namely, when test scores are not meant to be generalised beyond a specific context, such as within a company or a university, then the test that is designed to reflect features of that context (e.g. in the use of visuals, specific language, or cultural references) is far more likely to work, and Aptis was developed to take localisation into account when the conditions are met. O'Sullivan and Dunlea (2015, p. 8) describe different degrees of localisation, and the localisation scheme seems to indicate that Aptis can be localised to reflect local language norms and features. The test also uses the socio-cognitive framework as its validation framework, and special care is taken for test localisation not to undermine the validity of a test or put at risk fairness for test-takers. While we should bear in mind that the test scores generated from localised tests are meaningful only within the specific context for which the tests were localised, this approach taken by the British Council appears to take us a step forward in reconceptualising the role of the L1 in English language testing.

References

- ALC (2006). *Standard speaking tests*. Available online at: www.alc.co.jp/edusys/sst/english.html.
- Atkinson, J.M., and Heritage, J. (1984). *Structures of social action*. Cambridge: Cambridge University Press.
- Bonk, W.J., and Ockey, G.J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89–110.
- Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. *IELTS Research Report 6*, 71-89.
- Cambridge English (2008). *Cambridge Preliminary English Test: Official Examination papers from University of Cambridge ESOL Examinations (PET Practice Tests)*. Cambridge: Cambridge University Press.
- Cambridge English (2009). *Cambridge First Certificate in English: Official Examination papers from University of Cambridge ESOL Examinations (FCE Practice Tests)*. Cambridge: Cambridge University Press.
- Cambridge English (2018a). *Cambridge English First (FCE)*. Available online at: www.cambridgeenglish.org/exams-and-tests/first/
- Cambridge English (2018b). *Grade statistics*. Available online at: <http://gradestatistics.cambridgeenglish.org/2016/fce.html>
- Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, 3(3), 229-242.

- Cohen, J. (1998). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Erlbaum.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Council of Europe (2018). CEFR Companion Volume. Available online at: <https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989>.
- Eiken Foundation of Japan (2014) *TEAP Speaking rating scales*. Available online at: www.eiken.or.jp/teap/construct/sp_rating_crit.html
- Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39, 399–423.
- Field, J. (2018). *Rethinking the second language listening test: from theory to practice*. Sheffield: Equinox.
- Galaczi, E.D. and Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219-236.
- Gass, S.M., and Mackey, A (2000). *Stimulated recall methodology in second language research*. NJ: Lawrence Erlbaum.
- Geranpayeh, A., and Taylor, L. (eds.) (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge: UCLES/Cambridge University Press.
- Issacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *The Canadian Modern Language Review*, 64, 555-580.
- Iwashita, N., Brown, A., McNamara, T., and O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24-29.
- Jaiyote, S. (2016). *The relationship between test-takers' first language, listening proficiency and their performance on paired speaking tests*. Unpublished PhD thesis, University of Bedfordshire, UK.
- Jenkins, J. (1997). Testing pronunciation in communicative exams. In M. Vaughan-Rees, ed., *A special issue of Speak Out! Bringing together the interests of the IATEFL pronunciation and testing sigs*, pp. 7-11.
- Jenkins, J. (2000). *The phonology of English as an International Language*. Oxford: Oxford University Press.
- Jenkins, J. (2006). The spread of EIL: a testing time for testers. *ELT Journal* 60(1), 42–50.
- Khalifa, H and Weir, C.J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: UCLES/Cambridge University Press.
- May, L. (2011). *Interaction in a paired speaking test*. Frankfurt am Main: Peter Lang.
- Messick, S. (1989). Validity. In R. L. Linn, ed., *Educational measurement (3rd edition)*. London, NY: McMillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13, 241-56.
- MEXT (2008) *The course of study for upper secondary school*. Available online at: www.mext.go.jp/a_menu/shotou/new-cs/index.htm.
- Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking Test. In L. Taylor, and C.J. Weir (eds.), *IELTS Collected Papers 2: Research in reading and listening assessment* (pp.519-573). Cambridge: UCLES/Cambridge University Press.
- Nakatsuhara, F. (2014). *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1*

- and Study 2. Available online at:
www.eiken.or.jp/teap/group/pdf/teap_speaking_report1.pdf
- Nakatsuhara, F., and Jaiyote, S. (2015). *Exploring the impact of test-takers' L1 backgrounds on paired speaking test performance: how do they perform in shared and non-shared L1 pairs?* Paper presented at the BAAL/CUP Applied Linguistics Seminar. York St John University, UK, 24-26/06/2015.
- National Centre for University Entrance Examinations (2017). *Transition in the Number of Applicants*. Available online at: www.dnc.ac.jp/data/suii/suii.html
- O'Sullivan, B. (2011). Introduction – Professionalisation, Localisation and Fragmentation in Language Testing (pp.1-12). In B. O'Sullivan, ed., *Language Testing: Theory and Practice*. Oxford: Palgrave.
- O'Sullivan, B., and Weir, C.J. (2011). Test development and validation. In B. O'Sullivan, ed., *Language testing: Theories and practices* (pp. 13-32). Basingstoke: Palgrave Macmillan.
- Post, B. (2011). *Using acoustic analysis software to analyse L2 pronunciation features*. Paper presented at the 3rd BAALTEASIG conference. University of Warwick, UK. 18/11/2011.
- Seedhouse, P., and Egbert, M. (2006) The interactional organisation of the IELTS Speaking Test. *IELTS Research Report 6*, 161-205.
- Sewell, A. (2013) Language Testing and International Intelligibility: A Hong Kong Case Study. *Language Assessment Quarterly*, 10(4), 423-443.
- Shaw, S.D., and Weir, C.J. (2007). *Examining Writing: Research and practice in assessing second language writing*. Cambridge: UCLES/Cambridge University Press.
- Tanaka, C. (2018, June 13). Private-sector test results to account for 20% of national university English entrance exam scores. *The Japan Times*. Retrieved from <https://www.japantimes.co.jp/news/2018/06/13/national/private-sector-test-results-account-20-national-university-english-entrance-exam-scores/#.W5aI8uhKhAk>
- Taylor, L. (2006). The changing landscape of English: implications for English language assessment. *ELT Journal*, 60(1), 51-60.
- Taylor, L. (ed.) (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: UCLES/Cambridge University Press.
- Tsuzuki, M., and Nakamura, S. (2009). Intelligibility assessment of Japanese accents. In T. Hoffmann and L. Siebers (eds.), *World Englishes: Problems, properties and prospects*. Amsterdam: John Benjamins, pp. 239–261.
- Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Young, R. (2011). Interactional Competence in Language Learning, Teaching, and Testing. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning Volume II* (pp. 426–443). NY: Routledge.