OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

Language Education & Assessment, 2 (3), 135-154 (2019)
https://doi.org/10.29140/lea.v2n3.148

Building a Validity Argument for the Use of Academic Language Tests for Immigration Purposes: Evidence from Immigration-Seeking Test-Takers

NGOC THI HUYEN HOANG [a]

[a] *University of Queensland, Australia*
Email: huyenngochoang@gmail.com

Abstract

As validity pertains to test use rather than the test itself, using a test for unintended purposes requires a new validation program using additional evidence from relevant sources. This small-scale study contributes to the validation of the use of originally academic language tests—the International English Language Testing System and the Test of English as a Foreign Language—for assessing skilled immigration eligibility. Data were collected from 39 immigration-seeking test-takers, who are arguably under-represented in validation research. Analysis was informed by contemporary validity theory, which treats validity as a unitary concept incorporating score reliability, score interpretation, score-based decisions and their consequences. Results showed that the test-takers' perceptions varied widely. The evidence supporting this use included generally positive perceptions of the scores' reliability, washback effect, and fairness of score-based decisions. The refuting evidence concerned factors perceived to interfere with test-takers' performance and the complex consequences for the test-takers in aspects other than washback. However, overwhelmingly, as test-takers found the score-based decisions as fair, the validity judgement appeared tilted towards the positive side from the perspectives of these key stakeholders. Although the ultimate validity judgement requires the examination of evidence from other significant stakeholders as well, the present study has contributed valuable and unique evidence and bears important implications for research, practice, and policy particularly in high-stakes contexts such as immigration.

Keywords: contemporary validity theory, immigration, test-taker voices, test-taker inclusive validation, test use for unintended purposes, language assessment.

Introduction

Recent years' phenomenal growth in international migration has seen an ever high demand for assessment tools measuring aspired immigrants' proficiency in the destination polity's official

Data Availability Statement: *All relevant data are within this paper.*

language (Extra, Spotti, & Van Avermaet, 2009; Shohamy & McNamara, 2009). Yet, standardised language tests designed specifically for immigration remain scarce. Therefore, most immigration countries rely on existing language tests, which are intended for other purposes, even against test developers' guidelines for appropriate use. For example, the most widely used tests for processing migration visas—the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL)—are originally designed for academic purpose (i.e., assessing prospective students' readiness for English-medium secondary education).

The use of a test for purposes which it is not intended for raises critical questions about validity, a fundamental concern in testing and assessment (AERA, APA, & NCME, 2014). A test for academic purposes is designed so as to yield data to support inferences about the test-taker' ability to use language in universities and colleges. Using this test to screen immigrants involves making inferences about test-takers' capacity to use English as immigrants in the destination country based on inferences about their ability to use English in academic settings. While there is arguably some overlap between these two domains, a certain degree of misfit is inevitable since no test fits multiple purposes perfectly (AERA et al., 2014; Koch, 2013). Fulcher and Davidson (2009) emphasise that in such cases, a new, separate validity argument needs to be constructed to avoid test misuse and abuse. It is thus crucial to validate the use of standardised academic language tests for assessing skilled migration eligibility, which to date has been under-researched. Such validity inquiry contributes to the making of well-informed visa grant decisions which are fair to immigration-seeking test-takers and ultimately benefit immigration countries (Shohamy & McNamara, 2009).

Contemporary validity theory (AERA et al., 2014) holds that validation must involve examination of both the technical (i.e., test reliability and suitability for the purposes in question) and the social aspects (i.e., reasonableness of test use and its social impact). In addition, making a sound, unbiased validity judgement requires a compelling, comprehensive body of validity evidence, and the adequate representation of multiple stakeholders (AERA et al., 2014; Kane, 2006; Messick, 1989). Of all the key stakeholders, test-takers are the only ones to experience the test first-hand (Nevo, 1995) and are profoundly affected by the test score use (Kane, 2002), thus are considered the most important stakeholders in language testing (Rea-Dickins, 1997). Nonetheless, they have not been adequately represented in validation research(Cheng & DeLuca, 2011).

The current study seeks to fill the existing research gaps by investigating both technical and social dimensions of the use of academic language tests for immigration purposes from the perspective of immigration-seeking test-takers. It aims to answer the overarching research question of "How valid is the use of academic language tests for immigration purpose, through the lens of immigration-seeking test-takers?"

To situate this study within the broader literature, the next section reviews the expanding validation studies on language-in-immigration. Then a brief description of the methodology is provided, followed by the discussion of the results and some insight into the unified validity judgement. The concluding remarks close the paper with a set of recommendations for key stakeholders and suggestions for further research.

Validity of the Use of Language Testing in Immigration

Previous studies on language testing in the domain of immigration can be categorised into three broad groups according to which element of this phenomenon they focus on. The first group deals with technical features of tests and the test administration process; the second investigates rationales for setting language requirements for immigrants; while the third scrutinises its consequences.

One of the earliest studies in the first group is Merrylees (2003), which examined the suitability of the IELTS test for immigration purposes from the perspectives of two test-taker groups, one taking the test for immigration and the other for education in the UK. In particular, it explored these two groups' general attitudes and perceptions of the test as a whole as well as its four components in terms of difficulty, suitability of the topic, time allocation, and potential interferences with test performance. It was concluded that "The overall impression given about the IELTS test… was positive with a number of comments made about the appropriacy and effectiveness of the IELTS test for immigration purposes." (p. 36) However, this conclusion was drawn from the observation that the immigration-seeking group, like the other group, showed a general appreciation of the test's reliability. Indeed, the survey focused exclusively on the test per se with no direct reference to its use for assessing immigration eligibility. This conclusion contradicts the results of a qualitative study by Rumsey, Thiessen, Buchan, and Daly (2016). Based on interviews with health industry stakeholders and health professional immigrants in the Australian context, this study showed overall negative perceptions of the IELTS as a test for immigration. Concerns were raised about the IELTS's scoring protocols, which were believed to be not consistent, "like gambling" by some participants (p. 100). In addition, participants in both groups indicated that the IELTS test was not relevant to their work contexts, and thus, not a suitable testing tool for migrants working in healthcare. This was in agreement with Read and Wette (2009), which found a general perception among health professionals seeking permanent migration in New Zealand that "neither [the Occupational English Test (OET) or the IELTS] is, in any real sense, a test of their ability to communicative effectively in clinical contexts" (p. 3). In the same vein, Müller (2016) emphasised that achievement of satisfactory scores does not guarantee successful communication in clinical settings because language proficiency constitutes "a core pillar, rather than the sole contributor, of communicative competence" (p. 132). A similar argument was made about language requirement for visa in the meat production industry by Piller and Lising (2014), who rightly pointed out that

> Language at work is governed by a corporate regime and language in migration is governed by the state. These language regimes do not always operate in sync and sometimes even conflict... (p. 37)

The second group of studies on language testing in immigration context explore the rationales or motivations for introducing the language element in the processing of visa and citizenship applications, although they do not always link it to validity. Merrifield (2012) found that the immigration authorities of major English-speaking countries considered "easier settlement, integration into the host community and contribution to workforce knowledge" (p. 1) as main reasons for relying on standardised tests like the IELTS to screen intending immigrants. The usefulness and convenience of changing cut-off scores to manipulate the number of immigrants accepted and control immigration patterns was considered another important reason. Other studies (e.g., Berg, 2011; Blackledge, 2009; Capstick, 2011; Hunter, 2012) revealed similar motivations such as addressing skilled labour shortage, fostering destination countries' economic development, enhancing immigrants' active participation in the labour market, ensuring that they meet occupational health and safety standards, enabling them to access their full rights, and reducing costs for welfare systems.

Beyond economic reasons, a number of studies (e.g., Blackledge, 2009; McNamara, 2009; Shohamy, 2013; Shohamy & McNamara, 2009; Slade & ering, 2010) unpacked a range of social and political reasons behind language legislations for migration. They pointed out that requiring the newcomers to be proficient in the language of the receiving society is commonly claimed by politicians to uphold social cohesion, national identity and security often associated with monolingualism (Berg, 2011; Shohamy, 2013). As such, raising language requirements for immigrants appears to be an easy

and low-cost measure to deal with today's increasingly formidable challenges of balancing economic benefits with political stability. Ndhlovu (2008), based on a critical discourse analysis of Australia's language-in-migration policies throughout history, illustrated how language testing could be and has been often (ab)used for racial exclusion, explicit or subtle.

It is also notable that there remains little evidence to support the wide variety of rationales for language-in-migration policy. An Australian study by Chowdhury and Hamid (2016) challenges common public discourses about the vital role of English proficiency in immigrants' social integration and economic contribution. It clearly demonstrated that low English proficiency Bangladeshi immigrants in Australia were able to develop social and communication strategies to achieve satisfactory work, economic, and social life in the host society. On the other side of the story, Hoang and Hamid (2016) investigated two exceptional cases of prospective immigrants. Both had been residing in Australia for an extended period of time, secured good jobs, and enjoyed Australian social life. Yet they were unable to fulfil IELTS sub-score requirements for skilled visa after multiple attempts. The fundamental questions of the test's suitability for immigration and the "fairness" of language-in-migration policy were thus raised. Looking beyond the issue of language at the workplace, Gribble, Blackmore, Morrissey, and Capic (2016) discovered that non-linguistic factors deriving from the host community including workplace discrimination and isolation, rather than immigrants' language ability, played the most significant role in new immigrants' entry into the labour market and integration into the destination society. The disconnect between language proficiency and social integration and employability identified in the above studies suggest that the often-cited rationales for language requirements for immigrants might be untenable.

The third major strand of research delves into impact of the use of language testing in migration screening. Capstick (2011) documented the experiences of four learners of English struggled to meet the UK government's tightened language legislation for spousal visa applicants. The study showed the policy's differentiating effects on the immigrants and the receiving country: while it allowed the UK to benefit economically from immigrants' skilled low-wage labour and helped politicians gain electoral advantage by appearing to be tough on immigration issues, it denied "members of transnational families the right to marry by choice" and practically prevented many from uniting with their spouse already residing in the UK (Capstick, 2011, p. 3). Hoang and Hamid (2016) demonstrated significant financial, emotional-psychological, social-relationship, and other consequences of Australia's language-in-migration policy in two exceptional cases. Substantial affective impacts on immigration-seeking test-takers including depression, self-doubt, and negative self-perceptions were reported in the study by Rumsey et al. (2016). Inappropriate policies may backfire on immigration countries as well. Hoang and Hamid (2016) suggested that Australia's skilled migration scheme risked failing to achieve its goal of addressing skilled labour shortage if qualified, capable immigrants were denied visa solely on the basis of their language test results. Berg (2011) pointed out that a rigid language-in-migration policy had detrimental impact on the receiving country's cultural and linguistic diversity, which paradoxically, is considered in need of protection. She further argued that such use of language testing could lead to xenophobic attitudes and social and racial exclusion, as only those who speak that country's language are accepted into the society. Questions of human rights were also raised when proficiency in the immigration country's language prospective is made an entry requirement.

It is noteworthy that while previous studies have provided valuable evidence, few have explored both technical and social dimensions of the use of language testing for assessing migration eligibility, which is essential for a unified judgement of its validity.

Methodology

The Cases: IELTS and TOEFL

IELTS and TOEFL can be seen as archetypes of originally academic English tests used for immigration purposes, taken annually by millions of people over the globe (Educational Testing Service, 2018; IELTS Partners, 2018). IELTS claims to be "the high-stakes English test for study, migration or work" (IELTS Partners, 2018) [i]. It is the only internationally available English proficiency certification accepted by Citizenship and Immigration Canada. It also remains the preferred test by the immigration authorities of Australia, New Zealand, and the UK, although a few other tests are also accepted (IELTS Partners, 2018; Merrifield, 2012). These other tests are also not specifically designed for immigration, and thus validity issues encountered they are used for immigration screening should be similar to when IELTS and TOEFL are. TOEFL is not officially stated to be a test for migration and was not used for this purpose when the data for the current study were collected yet is now accepted for skilled migration in Australia, New Zealand, and the UK[ii]. Score requirements for skilled migration vary from one country to another. Australia requires a minimum IELTS score of 6.0 across all components or a TOEFL score of at least 12 in listening, 13 in reading, 21 in writing, and 18 in speaking as proof of competent English[iii]. For New Zealand, it is 6.5 in all IELTS components or a total score of 79 in TOEFL[iv]. The UK requires 6.5 across the IELTS components or 110 in TOEFL[v]. Canada appears less strict, as a score of 4.0 for speaking and 4.5 for listening on the IELTS general training module is acceptable[vi].

Participants

The research reported in this paper is drawn from a larger mixed-methods study. The parent study involved 517 people coming from and residing in over 50 countries/territories who took IELTS and TOEFL for different purposes (e.g., higher education, scholarship application, professional registration or employment). The current paper examines only the use of these tests in the immigration domain.

It involves 39 test-takers (16 female and 23 male), who reportedly had taken IELTS or TOEFL for immigration. They came from 14 countries including Vietnam, India, the Philippines, Germany, Italy, and the UK. Five participants identified themselves as native speakers of English, who sat the test to gain bonus points in the point-based system for skilled migration. The sample was reasonably homogeneous in terms of social class, with the majority of participants belonging to middle or higher-middle classes. Most of them were high scorers but over one third remained unsuccessful in obtaining the scores they targeted for migration.

Data Collection and Analysis

All the participants completed an online survey (phase 1) and six continued to follow-up individual interviews (phase 2). The survey sought information about 1) the participants' demographic details and experiences of taking the tests; 2) their perceptions of issues related to test reliability; and 3) their perceptions of test use and its consequences. Most of the survey items were constructed on a Likert scale, but there was also an open question (optional) at the end of each major sections asking for further comments, explanation, or elaboration. In total, 37 open comments were received. The in-depth interviews were semi-structured to ensure that the main topic was maintained while the informants had the opportunity to freely express themselves (Creswell & Plano Clark, 2011; Lichtman, 2010). This means that many questions were not prepared in advance but emerged from the participants' answers to earlier questions in the interview. Thus, the set of questions differed from one interview to another

(see a sample interview protocol in Appendix A). Each interview lasted from 1.5 to 2.5 hours. Both the survey and the interviews used lay language, taking heed of the common concern that a typical test-take might not be familiar with linguistic and assessment-specific terminologies and highly technical concepts. Where necessary, efforts were made to explain these concepts to make sure that the test-takers understood them properly before offering their views.

The qualitative data were analysed using content analysis with the help of NVivo. The analysis followed the six-stage procedure for systematic qualitative data coding proposed by Strauss and Corbin (1990). Specifically, after the data were gathered (stage 1), the interviews were transcribed, pseudonyms assigned to the participants, and data imported to NVivo (stage 2). The data were then fragmented (i.e., broken down into smaller chunks or meaningful parts and coded as free nodes in NVivo – stage 3) before they were categorised using axial coding strategy (stage 4). For the purpose of this study, the codes were aligned with the theoretically drawn components of validity (i.e., the overarching themes of test reliability and test score use, and the themes subsumed under them). As the study focused on validation, the data were further categorised as positive, negative, or neutral, which represented the participants' perceptions (i.e., whether they supported or rejected those particular elements of the tests and their use). Next (stage 5), they were linked (i.e., establishing the relationships between the codes through *inductive* process) and in the final stage, themes were generated. Due to the limited scope of this paper, the discussion of the results focuses only on validity-related themes.

Results and Discussion

Perceptions of Test Reliability

Perceived test reliability was conceptualised in consistence with the three inferential links concerning score reliability in Kane's (2006) validation framework: evaluation, generalization, and extrapolation. As such, three survey items were used to seek the test-takers' perceptions of: 1) how effectively the tests measured their English ability at the time of taking them; 2) how well the scores reflected their test performance; and 3) how well the scores predicted their English ability in the target context. The responses are presented in Table 1.

Table 1 *Test-takers' perceptions of the tests' reliability*

Aspects of reliability	(Strongly) agree	Neutral	(Strongly) disagree	Don't know/don't remember/non-response
Effective measure	19 (49%)	8 (21%)	12 (31%)	0
Accurate scores	18 (46%)	11 (28%)	10 (25%)	0
Predictivity	14 (36%)	9 (23%)	14 (36%)	2 (5%)

As the table shows, nearly half of the participants believed that the tests effectively measured their English proficiency and that the scores accurately reflected their test performance but just over one third of them found the scores predictive of how well they would use English in the target context. The low ratings for the test scores' predicting power could signify test-takers' perceptions of the mismatch between the domain of use intended by the tests (i.e., mainly academic) and that of their actual use (i.e., immigration). There was a clear tendency to consider the tests as reliable but not completely so. The reasons for this general perception were further examined by a survey item aiming to ascertain whether test performance and scoring were affected by the various factors identified in the literature. Table 2 displays responses to this question.

Table 2 *Potential interferences with test performance and scoring (n = 39)*

Factors	No interference	Slight interference	Heavy interference	Don't know/don't remember/non-response
Perceived inferences with test performance				
Unfamiliarity with tests	16 (41%)	14 (36%)	8 (21%)	1 (3%)
Testing condition	25 (64%)	6 (15%)	8 (21%)	0
Test administration	19 (49%)	15 (39%)	5 (13%)	0
Test structure	18 (46%)	13 (33%)	8 (21%)	0
Test content/topics	10 (26%)	16 (41%)	13 (33%)	0
Question types	12 (31%)	16 (41%)	11 (28%)	0
Feelings while taking tests	7 (18%)	12 (31%)	20 (51%)	0
Perceived inferences with test score				
Scoring system	5 (13%)	17 (44%)	15 (39%)	2 (5%)
Consistency between raters	5 (13%)	17 (44%)	14 (36%)	3 (8%)

It appears that in the test-takers' view, the following factors did not significantly affect test reliability: 1) the testing condition (specified in the survey as factors such as room configuration, noise and light in the test room and sound quality); 2) test administration procedure (e.g., checking identity, ushering examinees to test rooms and seats, distributing and collecting test materials, and instructions for test-takers); and 3) test structure (e.g. constituent sections of each test, number of questions per section, order of questions, and time allocations). Lack of familiarity with the tests could also be considered an insignificant factor, as only eight test-takers (21%) reported considerable interference. The remaining factors were perceived to compromise the tests' reliability to varying degrees, as will be discussed in the following sections.

Test Content/Topics

While some test-takers stated that topical knowledge largely determined one's performance on the tests, others posited that unfamiliarity with or lack of knowledge of the test topics would put the test-taker at a disadvantage. All the interviewees indicated that they would have performed better if the topics had been related to their field of study or work. However, in IELTS and TOEFL, test-takers are not given choices over the test topics in any sections. Thus, many of them believed "luck" (in the form of having a familiar topic) could largely affect their ability to demonstrate their language ability. This view is consistent with the findings of many studies on the potential effect of subject/topical knowledge (either alone or in interaction with other factors such as one's language proficiency) on test performance (Alderson & Urquhart, 1985; Bachman & Palmer, 2010; Huang, Hung, & Plakans, 2018; Jensen & Hansen, 1995; Karimi, 2016), which could be considered a source of invalidity (Jennings, Fox, Graves, & Shohamy, 1999).

Question Types

Some test-takers identified certain discrepancy between the tests' intended and their actual domains of use. For instance, I34, who took the academic module of IELTS, stated that the test questions were too general for its purpose (i.e., academic). Yet, I33, a test-taker of IELTS general training module

maintained that the test tasks were too complex for real-life language encounters of a typical immigrant. Interestingly, apart from these comments, very few references to question types were made.

Feelings while Taking the Tests

It is not surprising that feelings were most commonly reported to affect the test-takers' performance, given the high-stakes nature of these tests. Test-takers' feelings were investigated through a survey item asking the respondents to use at least three words or phrases to describe how they felt while taking the tests. The question was responded with a considerable number of words that denote positive feelings such as *confident, calm*, and *relaxed*. However, these were outnumbered by those conveying negative feelings including *anxious, tired, stressed, scared, nervous, uncomfortable, annoyed, and angry*. The main reasons for these feelings, as self-reported by the test-takers, included the time, effort, and money they had invested in the tests and the anticipated consequences of failure to achieve desired scores. I9's reflection on her eight times sitting IELTS without success illustrates this impact most clearly:

> [Because IELTS]'s gonna change your life, it really causes you a lot of pressure and worry [...] and sometimes you don't focus on the test, you just keep telling yourself "I need to pass, I need to pass" and then you don't pass! [...] If your body reacts to this kind of thing, you can't think clearly. You just know [...] you need to pass IELTS otherwise you have to go home. And you can't concentrate although you have studied for it.

This quote reflects Shohamy's (2001b) observation that test-takers have a clear sense of the gatekeeping role and power of these tests in their life, which invokes anxiety, fear, and a feeling of helplessness. The participants were fully aware of how these negative feelings impacted on their performance yet failed to control them. While impact of psychological state on test performance has been documented in an extensive body of research er & Pekrun, 2004; von der Embse & Witmer, 2014; Zeidner, 1998), the current study further indicates that it tends to be more severe when the test results are used to make such life-changing decisions as granting migration permission.

Scoring

Although the scoring of IELTS and TOEFL is routinely inspected and "endorsed" by considerable research mostly by in-house research teams and external researchers [vii], the test-takers in the present study did not display a high trust in it. Nearly four fifths of them believed that the scoring system and the marking consistency to some extent affected their scores. The transparency of the marking process was frequently questioned, probably because IELTS and TOEFL do not provide feedback to test-takers. Survey respondent I16 made a strong point that, "The speaking test and writing test are subjective. We don't know the exact result of how we were going on the test. Need to have a specific result explained to the examinees." Some IELTS test-takers believed that they did not receive a right/fair score for the speaking part due to the lack of professionalism of speaking examiners. I34, who speaks English as the first language, raised the issue of the examiner's inappropriate attitude and behaviour on discovering that she was taking the IELTS for Australian migration. It was on this occasion that she received a significantly lower score for speaking (7.5) than on all other sittings (8.5). Another IELTS test-taker, I37, indicated that he performed worse than expected because of the "apparently sluggish and bored" examiner's attitude. These perceptions corroborate findings of previous studies that language speaking test examiners vary in their elicitation of test-takers' response which affects test-takers' performance as well as examiners' judgement of their language ability (e.g., A. Brown, 2003). The lack of standardisation across examiners echoed in this test-taker study signifies potential threats of test bias which need to be considered and rectified.

Notably, like participants in Rumsey et al.'s (2016) study, some test-takers attributed what they perceived as scoring problems to the commercial nature of the tests. In particular, the pattern of failing to obtain the required score in one test component in a test sitting and another component in the next among many test-takers seeking Australian immigration, was thought to be a mechanism to financially exploit test-takers.

In short, the test-takers pointed out a number of factors as hinderances to their optimal performance on the tests and/or contamination of their "true score." Though it was impossible to verify the exact nature and magnitude of the influence of these factors with the available data, the perceptions themselves affected the test-takers' psychological state and cognitive functionality during the test. Scores obtained under these conditions were unlikely to reflect their true ability. Consequently, inferences about test-takers' language ability based on the test scores might not be entirely sound, which could lead to unfair and unreasonable decisions about aspiring immigrants. Beyond concerns about the reliability of IELTS and TOEFL scores, the following section discusses findings about their use for immigration from the test-takers' perspectives.

Score Interpretation and Use

The test-takers' perceptions of the use of their test results by immigration authorities were explored with regards to four key aspects: 1) the extent to which the test scores were relied on in this decision-making process; 2) the cut-off scores; 3) consequences of this test score use; and 4) overall appropriateness of the score-based decisions. The findings are discussed below drawing on both survey and interview data.

Extent of Reliance

Table 3 summarises the survey response concerning how much the test scores determined one's visa application outcome.

Table 3 *Test-takers' perceptions of the extent of reliance on of IELTS and TOEFL in processing skilled migration visa*

Aspect of score use	Test-taker perceptions	Frequency & percentage
Extent of reliance	Appropriate	22 (56%)
	Inappropriate:	16 (41%)
	• Too heavy	• 6 (38%)
	• Should not rely on the score	• 4 (25%)
		• 6 (38%)
	• No comment	1 (3%)

While the majority of participants considered the extent to which immigration authorities relied on their test scores to make immigration decisions as appropriate, over a quarter considered it an overreliance. These latter group of test-takers particularly criticised what they viewed as "rigid" policy of accepting only a limited number of tests, disregarding other potential evidence of their language proficiency. The strongest critics were I9 and I33, who failed to obtain desired sub-scores for Australian permanent residence visa after multiple attempts. Both of them argued that the decisions on their immigration eligibility would have been more reasonable if other indicators of English proficiency had been also considered.

Cut-Off Scores

The test-takers' views about the appropriateness of the scores required by immigration departments varied, as Table 4 shows.

Table 4 *Test-takers' perceptions of the use of IELTS and TOEFL by immigration authorities*

Aspect of score use	Test-taker perceptions	Frequency & percentage
Cut-off scores	Appropriate	20 (51%)
	Too high	10 (26%)
	Too low	3 (8%)
	Some too high, others too low	4 (10%)
	Don't know/ Non-response	2 (5%)

While over half of the test-takers advocated these cut-off scores, most of the remaining found them unrealistically high. Some problematised specific requirements of sub-scores rather than those of overall or total score. I39, I33, and I9 strongly voiced against Australian's requirement of 6.0 in all IELTS components to meet minimum language requirement or 7.0 to gain 10 bonus points. This legislation practically forced them to repeat the test many times and to waste an unreasonable amount of money but more importantly, they argued, it was irrelevant to the reality of immigrants' language demands. I39, who had resided in Australia for five years, contended that lower scores in reading and writing would suffice but higher speaking and listening scores were essential for a full integration into Australian life. I34 argued that this score requirement was both fair and unfair. According to her, it was fair because it benefited Australia by allowing the country to "pick the best of the very best." It was, in her view, unfair because it effectively filtered out a great many highly skilled migrants who would otherwise be accepted, especially those in professional fields where high English proficiency is not crucial. Like I32, she maintained that the tests were used as a tool to control the migration flow and so the standard setting was arbitrary. These beliefs are in line with Merrifield's (2012) findings and highlight ethical issues that might arise when the score users do not have expertise in assessment (Kane, 2012) or intentionally abuse tests (Shohamy, 2001b).

Consequences

Consequences of test use are integral to validity in any contexts but are more so in high-stakes ones (Bachman & Palmer, 2010; Shepard, 1997). Examining the full range of consequences is thus critical in validation. To this end, the present study scrutinises all positive and negative, intended and unintended, short-term and long-term effects of the use of IELTS and TOEFL for immigration in terms of:

- learning of English (also called washback effects, see Tsagari & Cheng, 2016)
- finance (Templer, 2004)
- motivation, self-efficacy, self-image, confidence and pride (Crooks, Kane, & Cohen, 1996; Kirkland, 1971; Ockey, Koyama, & Setoguchi, 2013; Slomp, Corrigan, & Sugimoto, 2014)
- psychological and social-emotional wellbeing (Ahern, 2009; Bachman, 2005; Crooks *et al.,* 1996; Kirkland, 1971; Shohamy, 1998, 2001a; Templer, 2004) and
- social relationships (Crooks et al., 1996; Nevo & Sfez, 1985)

Table 5 summarises these types of impact based on the survey data.

Table 5 *Reported impacts of the use of IELTS and TOEFL scores as a requirement for skilled migration*

Types of impact	Test-taker responses	Frequency & percentage
Washback	Positive	30 (77%)
	No influence/both positive and negative	7 (18%)
	Negative	2 (5%)
Financial	Money spent	
	• Over USD 2,000	• 3 (8%)
	• USD 1,000 - 2,000	• 5 (13%)
	• Up to USD 1,000	• 18 (72%)
	• Don't know/ Non-response	• 3 (8%)
	Financial costs being significant	
	• (Strongly) agree	• 23 (59%)
	• Neutral	• 12 (31%)
	• (Strongly) disagree	• 3 (8%)
	• Don't know/ Non-response	• 1 (3%)
	Happy about costs	
	• (Strongly) agree	• 6 (15%)
	• Neutral	• 10 (26%)
	• (Strongly) disagree	• 22(56%)
	• Don't know/ Non-response	• 1 (3%)
Psychological/emotional impact	(Strongly) agree	23 (59%)
	Neutral	6 (15%)
	(Strongly) disagree	10 (26%)
Positive consequences outweighing negative consequences	(Strongly) agree	17 (39%)
	Neutral	7 (18%)
	(Strongly) disagree	15 (44%)

The following section discusses in greater detail how these types of impact were experienced drawing on insights from the qualitative data.

Washback

Over three quarters of test-takers reported positive washback effect, in consistence with the numerous studies such as Wall and Horak (2006) and Hawkey (2001, 2006). As I5 explained, he had to study hard for the test to avoid turning the considerable amount of money he had spent on test registration into a complete waste, and his hard work eventually improved his language proficiency. This explanation relates strongly to Shohamy's (2001) note that "tests do have the role of creating pressure and motivating [test-takers] to study, mostly out of fear of their consequences" (p. 13). However, for I32, I34, and I39, their test preparation was heavily focused on strategies and tricks to obtain the desired scores quickly. This process did not result in substantial improvement of their English proficiency, which, they argued, required methodical practice over a sustained period.

Financial Impact

The costs of preparing for and taking the tests emerged as one of the test-takers' greatest concerns. These activities were reported to have cost three test-takers more than USD 2,000; five others between USD 1,000 and 2,000; and the rest up to USD 1,000. Regardless of socio-economic background, all except three test-takers considered these costs as significant. I5, an undergraduate student who had to

depend on his parents financially, exclaimed that the test fee was "intolerably exorbitant!" and unaffordable to the majority of people in his country. However, to stand a chance for migration, test-takers like him had to pay a "high price" (Ahern, 2009; H. D. Brown & Abeywickrama, 2010), willing or not. Indeed, only six test-takers were happy with the expended amount. Among these few exceptions, I34 elaborated on how profitable and "worthwhile" her financial, time, and effort investment became once her visa was granted. She listed the numerous benefits including access to high-quality healthcare, tuition fee subsidy, better scholarship and employment opportunities, and "so many other benefits." By contrast, I9 was anguished and furious that her hard-earned AUD 20,000 investment in the IELTS turned out to be a complete loss.

Of particular concern is the link between financial impact and test-takers' performance, as signalled by many test-takers. I5 asserted that his main source of stress and anxiety experienced during the test preparation process and on the test day was the test-related costs. For I39, it was a feeling of extreme regret over "wasting a big amount of money" in previous failures compounded by the fear of failing again. The interaction between different types of impact and their aggregate interference with test reliability again underscore the need to examine not just the test but also its use, which has real bearings on score-based decisions.

Impact on Motivation, Self-Efficacy, Self-Image, Confidence, and Pride

This impact was examined qualitatively due to its complex nature. Analysis showed that a few test-takers including I5 who succeeded from their first test attempt tended to associate themselves with a very positive self-image. Successful "conquest" of a test commonly considered challenging effectively boosted their confidence, pride, and motivation for further endeavours. By contrast, other test-takers such as I9, I33, and I39, reported self-doubt, loss of motivation, and a feeling of despair upon repeated failures. Having used English over extended periods of time in study, work, and daily life, they were unable to make sense of their unexpectedly low results. The only reason they could find was "luck," which suggested a complete loss of control over the tests, themselves, and their chance of success in future attempts. This feeling was most pronounced in I9's case. After eight failed attempts, she retained no motivation and declared that she would never apply for Australian permanent residency again, even if the regulation changed.

Psychological/Emotional Impact

A small number of test-takers who gained early success in the tests enjoyed very positive psychological effects manifested in their feelings such as "happy," "contented," "delighted," and "excited" upon their achievement. For example, I5 recounted that:

> I was over-excited on receiving the results. It was at mid-night, but I screamed so loudly that my house felt it would break to pieces. I rushed to my parents' room to tell them the news and we were all over the moon.

The remaining test-takers reported impact on their psychological/emotional wellbeing in various ways. Those who finally succeeded after multiple test sittings tended to experience both positive and negative psychological impact. For example, I34 described how "shocked," "disappointed," and "nervous" she was during her first two failed attempts, but the exact opposite was felt on her success in the third sitting. However, the most serious psychological impact was felt by test-takers who remained unsuccessful. It was common for these test-takers to feel frustrated and furious (I9 and I33), guilty and regretful (I3), and detrimentally stressed (various test-takers). I3, whose wife had taken most of the housework so he could have more time for test preparation, reported feeling "guilty about letting her

down" on receiving the unsuccessful outcome. He took over the housework and "punished" himself by cutting time and money on leisure activities for a few months. While the feeling of guilt could be alleviated, stress appeared to be out of control to him. It caused I39 to have serious sleep problems. The mere thought of IELTS nauseated him every morning days before the test date. Yet he "didn't allow [himself] to take even a short break to think about my feelings" and could only realise how stressed and exhausted he had been the moment he received the successful result. In an imagined scenario that he failed that third attempt, I39 thought he "would have exploded with stress and would need to retreat from the test for a couple of years before I could recollect my courage to take it again." The psychological impact of the test was so profound that at the beginning of the interview, which was conducted three years after the event, the notion of IELTS still made him "shiver in horror" (in his own words). When tests are associated with psychological impact of this intensity, measures need to be taken to protect the safety and wellbeing of test-takers (Hopfenbeck, 2017).

Impact on Social Relationships

The use of language test scores for immigration processing was believed to influence the test-takers' relationships with their family, friends, and acquaintances, as the qualitative data revealed. Positive effects were reported when the achieved success. I5, for example, had his extreme pride and happiness shared by his parents. In addition, hundreds of his friends and classmates "liked" his Facebook status about the result and "sounded as if it was also their success." The feeling of relief and elation over test success fostered the relationships between the test-takers and people close to them. On the contrary, not yet successful and unsuccessful test-takers chose to hide away from acquaintances and even close family members like parents (I33) to avoid mention of the tests. Before he was successful in the test, I33 had frequent quarrels with his wife, who was anxious about their visa application and grew doubtful of his ability.

Outside their circle of social acquaintances, unsuccessful test-takers often developed negative attitudes towards the test makers and immigration authorities. I9, for instance, resented IELTS and Australian immigration authorities for the hard-line policy for having taken her time, money, job but more gravely, "youth," "energy," and "life".

Weighing Positive Consequences against Negative Consequences

Kane (2002) asserts that in high-stakes contexts, score use can be deemed valid only if its positive consequences outweigh its negative consequences. However, an overall comparison of impacts of language test use in migration is rare to find in the published literature, particularly with regards to the immigration-seeking test-takers' perspective. In the current study, which aims to shed some light on this matter, 17 test-takers (39%) indicated that the former outweighed the latter, 15 (44%) indicated the opposite, while 17 stated a neutral position. This was well explicable by the test-takers' perceptions of test consequences (as discussed in preceding sections) relative to their success or failure to meet language requirements for migration.

Decision Fairness

Responses to the survey question about the fairness of the score-based decisions are presented in the following table.

Table 6 *Test-takers' perceptions of the use of IELTS and TOEFL by immigration authorities*

Aspect of score use	Test-taker perceptions	Frequency & percentage
Appropriateness of score-based decision	(Strongly) agree	26 (67%)
	Neutral	3 (8%)
	(Strongly) disagree	9 (23%)
	Don't know/ Non-response	1 (3%)

Two thirds of the test-takers believed that the decision made about them was fair, whereas less than a quarter disagreed with this statement. It is interesting to note that the number of test-takers supporting decision fairness (n = 26) was greater than the number of test-takers who had obtained the desired scores (n = 22) and the number of those who experienced more benefits than losses from the tests (n = 17). This difference suggests that Kane's (2002) idea of basing validity judgement on consequences might not apply to individual stakeholder groups but rather, to all relevant groups collectively (i.e., based on aggregate evidence concerning consequences for all stakeholders). Furthermore, it shows that test-takers do go beyond their self-interests and are concerned about fairness for the population in making their judgement about tests and test use. With this characteristic, test-takers might constitute a source of reliable information in test evaluation.

The Unified Validity Judgement

The participants in this study offered a mix of evidence that both supported and rejected the validity of the use of standardised academic test scores for immigration. The supporting evidence included generally positive perceptions of the scores' reliability, washback effect, and fairness of score-based decisions. The refuting evidence concerned factors perceived to interfere with test-takers' performance and the complex consequences for the test-takers in aspects other than washback. However, overwhelmingly, as more test-takers found the score-based decisions fair, the validity judgement appeared tilted towards the positive side. Thus, it could be suggested that from the perspectives of immigration-seeking test-takers, the use of standardised language test scores for assessing skilled migrants is moderately to largely valid. This is somewhat surprising given the theoretically identifiable gap between the intended (i.e., academic) and actual domains of test use as well as strong professional voices against the use of these proficiency tests for immigration purposes (e.g., Australian Council of TESOL Associations, 2017). A possible reason is that most of the participants were residing in English-speaking countries and possibly many of these aspiring skilled migrants were able to use their academic skills (which the tests measured) in their professional field in the destination country. However, the exact reasons cannot be ascertained until further investigation.

Nevertheless, this test-taker-based evaluation should not be considered ultimate. First, the small sample size of this study, which was due to remarkable difficulty in participant recruitment, restricts the generalisability of the findings. This means that a larger sample might end up with a different validity evaluation outcome. The more important reason is that the overall validity judgement must reflect the perspectives of all key stakeholders, which also include but are not limited to test developers, representatives of immigration departments, test-takers' language teachers, and employers in the destination country.

Concluding Remarks

This study examined the validity of the use of standardised language tests to assess skilled migration eligibility through the lens of immigration-seeking test-takers. Despite the small sample size, rich and unique data were gathered that provided new and important insights into the matter. The evidence

generated suggests that the validity of this use, though only "a windfall for the test owners" (Australian Council of TESOL Associations, 2017, p. 23), was supported by most test-takers. The results of the study have important implications for educational assessment development, administration, and use.

First, as indicated by a number of test-takers, test performance could be affected by a complex combination of test-takers-related, test-related, and score use-related factors. Most notably, the use of the tests for high-stakes decisions and the high test fees can trigger tremendous test anxiety. It is also recognised that most of the validity issues identified were attributable to the high stakes attached to the tests rather than to their inherent features.

Second, when the stakes attached to the tests are as high as in the context of immigration, the impacts, either positive or negative, can range widely. The concept of consequences in validation should be extended beyond washback to include, for example, financial and psychological-emotional effects. Awareness of the full array of possible and actual effects of test use would be beneficial for migration policymakers who wish to control potential harms for intending migrants.

Furthermore, despite a common concern that without specialist training in testing and assessment, the average test-taker lacks the capacity to form and express defendable opinions about tests (Wall, Clapham, & Alderson, 1994), the present study has provided more optimistic findings. It has shown that test-takers possess a strong ability to articulate consistent views and support them with relevant reasoning and/or experiential evidence, albeit without using technical terminologies. This could be explicable by the fact that applicants of skilled migration visa are likely to be more highly educated than an average test-taker of English language tests. As they also tend to be high scorers of language tests, they are likely to communicate their opinions clearly and effectively. The finding suggests a great potential of utilising immigration-seeking test-takers' experiences and perceptions in high-stakes testing validation. Test-takers' voices appear to be most useful in identifying perceived interferences with test performance and the complicated psychological and social impact of test score use. As these issues are not obtainable from or might be overlooked by other stakeholders, engaging test-takers in validation is vital and highly valuable.

Finally, this case study has highlighted that the use of academic language tests, even those of high standards, for immigration purpose could be associated with social and technical issues. However, the current lack of immigration-specific language tests means that this use is unlikely to cease in a near future (Rumsey et al., 2016). Therefore, solutions need to be sought to maximise the reliability of the scores and ultimately the fairness of score-based decisions in the existing system. In an orchestrated endeavour to make the use of language testing for immigration purposes more valid, this study offers the following set of recommendations.

For Score Users and Administrators

It is recommended that before considering a test for a particular purpose, score users, in this case immigration authorities, specify a set of language skills and knowledge required of immigrants and justify it. The test selected needs to be one that demonstrates a close match with the skills and knowledge specified. It is crucial that score administrators (who directly apply score use guidelines to process skilled migration visa applications) be aware that however robust the test is, the scores obtained under tremendous pressure in such a high-stakes setting are unlikely to be perfectly accurate indicators of one's language ability. Therefore, the scores may need to be considered in conjunction with other possible indicators or evidence of language proficiency, especially in processing applications on the borderline of acceptance/rejection. These steps are probably best carried out in collaboration with test experts, given that score users are usually non-experts (Hamp-Lyons, 2000).

The entire score-based decision-making process needs to be reasonable, consistent, and transparent. It should be documented for the purpose of collecting continuously growing evidence to validate a new use of language tests.

For Test Developers and Administrators

As language assessment experts, test developers are responsible for providing exhaustive guidelines for test score interpretation and use and acting within their capacity to control sources of avoidable interferences with test performance. They are also in the best position to offer training in assessment literacy for other major stakeholders in language testing. Test makers may intervene but should not be held accountable for test misuse and abuse. The test construction and administration ideally incorporate test-takers' expectations and desires, to the extent technically possible, to create a test-taker-friendly environment and mitigate test anxiety. Responding to test-takers' diverse needs also helps develop test-takers' positive attitudes towards the test, which facilitate them in demonstrating their language ability.

For Test-Takers

To improve test reliability, test-takers are expected to familiarise themselves with the test and the test administration process. They need to understand what the scores represent and how they are used in visa processing, in part to control their own test anxiety. Also, their active engagement in test evaluation is crucial as they can offer unique validity evidence.

Further studies can be conducted on a larger scale, involving other key stakeholders so as to provide an overall view of language-in-migration policy. Also important is research into the relationship between achievement on standardised academic language tests and social integration and economic contribution in the destination society.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (2014 ed.): Washington, DC: American Educational Research Association.

Ahern, S. (2009). "Like cars or breakfast cereal": IELTS and the trade in education and immigration. *TESOL in Context, 19*(1), 39-51.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*(2), 192-204. https://dx.doi.org/10.1177/026553228500200207

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34. https://dx.doi.org/10.1207/s15434311laq0201_1

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

Berg, L. (2011). 'Mate speak English, You're in Australia now': English language requirements in skilled migration. *Alternative Law Journal, 36*(2), 110-115. https://dx.doi.org/10.1177/1037969x1103600208

Blackledge, A. (2009). "As a country we do expect:" The further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly, 6*(1), 6-16. https://dx.doi.org/10.1080/15434300802606465

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25. https://dx.doi.org/10.1191/0265532203lt242oa

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices*. New York: Pearson Education.

Capstick, T. (2011). Language and migration: The social and economic benefits of learning English in Pakistan. In H. Coleman (Ed.), *Dreams and realities: Developing countries and the English language*: British Council.

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104-122. https://dx.doi.org/10.1080/10627197.2011.584042

Chowdhury, F., & Hamid, M. O. (2016). Language, migration and social wellbeing: A narrative inquiry into the lives of low English proficiency Bangladeshi migrants in Australia. *Australian Review of Applied Linguistics, 39*(1), 8-30. https://dx.doi.org/10.1075/aral.39.1

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks: SAGE Publications.

Crooks, T. J., Kane, M., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*(3), 265-286. https://dx.doi.org/10.1080/0969594960030302

Educational Testing Service. (2018). *TOEFL*. Retrieved from http://www.ets.org/toefl

Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. London: Continuum.

Gribble, C., Blackmore, J., Morrissey, A.-M., & Capic, T. (2016). *Investigating the use of IELTS in determining employment, migration and professional registration outcomes in healthcare and early childcare education in Australia* (4). Retrieved from https://www.ielts.org/teaching-and-research/research-reports?fr=4836182a-bae5-4314-831a-2f540352d3d5|68044d93-8326-4704-b680-191904efc373|#sthash.LPclLmsr.dpuf

Hamid, M. O., & Hoang, N. T. H. (2018). Humanising language testing. *Teaching English as a Second or Foreign Language - Electronic Journal, 22*(1), 1-20.

Hawkey, R. (2001). *The IELTS impact study: Development and implementation*. Retrieved from https://www.cambridgeenglish.org/Images/23117-research-notes-06.pdf

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000* (Vol. 24.). Cambridge, UK: Cambridge University Press.

Hoang, N. T. H., & Hamid, M. O. (2016). 'A fair go for all'? Australia's language-in-migration policy. *Discourse: Studies in the Cultural Politics of Education, 18*(1), 1-15. https://dx.doi.org/10.1080/01596306.2016.1199527

Huang, H.-T. D., Hung, S.-T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27-49. https://dx.doi.org/10.1177/0265532216677106

Hunter, J. (2012). Language and literacy on the ground: Disconnects between government policy and employer perspectives. *Discourse: Studies in the Cultural Politics of Education, 33*(2), 299-311. https://dx.doi.org/10.1080/01596306.2012.666082

IELTS Partners. (2018). IELTS. Retrieved from http://www.ielts.org/default.aspx

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456. https://dx.doi.org/10.1177/026553229901600402

Jensen, C., & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing, 12*(1), 99-119. https://dx.doi.org/10.1177/026553229501200106

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41. https://dx.doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*. Westport, Conn.: Praeger Publishers.

Karimi, M. N. (2016). Prior topical knowledge and L2 proficiency as determinants of strategic processing in English for Academic Purposes multi-texts comprehension. *Innovation in Language Learning and Teaching*, 1-12. https://dx.doi.org/10.1080/17501229.2016.1177058

Kirkland, M. C. (1971). The effects of tests on students and schools. *Review of Educational Research, 41*(4), 303-350. https://dx.doi.org/10.2307/1169441

Koch, M. J. (2013). The multiple-use of accountability assessments: Implications for the process of validation. *Educational Measurement: Issues and Practice, 32*(4), 2-15. https://dx.doi.org/10.1111/emip.12015

Lichtman, M. (2010). *Qualitative research in education: A user's guide*. Thousand Oaks, Calif: Sage Publications.

McNamara, T. (2009). Australia: The dictation test redux? *Language Assessment Quarterly, 6*(1), 106. https://dx.doi.org/10.1080/15434300802606663

Merrifield, G. (2012). *The use of IELTS for assessing immigration eligibility in Australia, New Zealand, Canada and the United Kingdom*. Retrieved from http://www.ielts.org/PDF/vol13_Report1.pdf

Merrylees, B. (2003). *An impact study of two IELTS user groups: Candidates who sit the test for immigration purposes and candidates who sit the test for secondary education purposes*. Retrieved from https://www.ielts.org/teaching-and-research/research-reports/volume-04-report-1

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan.

Müller, A. (2016). Language proficiency and nursing registration. *International Journal of Nursing Studies, 54*, 132-140. https://dx.doi.org/10.1016/j.ijnurstu.2015.01.007

Ndhlovu, F. (2008). A critical discourse analysis of the history of the language question in Australia's migration policies: 1901 – 1957. *Australian Critical Race and Whiteness Studies Association (ACRAWSA) e-Journal, 4*(2), 17-33.

Nevo, B. (1995). Examinee feedback questionnaire: Reliability and validity measures. *Educational and Psychological Measurement, 55*(3), 499.

Nevo, B., & Sfez, J. (1985). Eaminees' feedback questionnaires. *Assessment & Evaluation in Higher Education, 10*(3), 236-248. https://dx.doi.org/10.1080/0260293850100305

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10*(3), 292.

Piller, I., & Lising, L. (2014). Language, employment, and settlement: Temporary meat workers in Australia. *Multilingua: Journal of Cross-Cultural and Interlanguage Communication, 33* (1-2), 35-59. https://doi.org/10.1515/multi-2014-0003.

Read, J., & Wette, R. (2009). *Achieving English proficiency for professional registration: The experience of overseas-qualified health professionals in the New Zealand context* (4). Retrieved from https://www.ielts.org/teaching-and-research/research-reports/volume-10-report-4

Rumsey, M., Thiessen, J., Buchan, J., & Daly, J. (2016). The consequences of English language testing for international health professionals and students: An Australian case study. *International Journal of Nursing Studies, 54*, 95-103. doi:https://doi.org/10.1016/j.ijnurstu.2015.06.001

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8. https://dx.doi.org/10.1111/j.1745-3992.1997.tb00585.x

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331-345. https://dx.doi.org/10.1016/s0191-491x(98)00020-0

Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391. https://dx.doi.org/10.1177/026553220101800404

Shohamy, E. (2001b). *The power of tests: A critical perspective on the uses of language tests*. Harlow: Longman.

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication, 13*(2), 225-236. https://dx.doi.org/10.1080/14708477.2013.770868

Shohamy, E., & McNamara, T. (2009). Language tests for citizenship, immigration, and asylum. *Language Assessment Quarterly, 6*(1), 1. https://dx.doi.org/10.1080/15434300802606440

Slade, C., & Möllering, M. (Eds.). (2010). *From migrant to citizen: Testing language, testing culture*. Basingstoke, Hampshire: Palgrave Macmillan.

Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English, 48*(3), 276-302.

Stöber, J., & Pekrun, R. (2004). Advances in test anxiety research. *Anxiety, Stress, & Coping, 17*(3), 205-211. https://dx.doi.org/10.1080/1061580412331303225

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

Templer, B. (2004). High-stakes tests as high fees: Notes and queries on the international English assessment market. *Journal for Critical Education Policy Studies, 2*(1).

von der Embse, N. P., & Witmer, S. E. (2014). High-stakes accountability: Student anxiety and large-scale testing. *Journal of Applied School Psychology, 30*(2), 132-156. https://dx.doi.org/10.1080/15377903.2014.888529

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321-344. https://dx.doi.org/10.1177/026553229401100305

Wall, D., & Horak, T. (2006). *The TOEFL impact study: Phase 1 The baseline study. TOEFL Monograph 34*. Princeton, NJ: Educational Testing Service.

Zeidner, M. (1998). *Test anxiety: The state of the art*. Dordrecht: Springer.

Author biodata

Ngoc T. H. Hoang recently obtained her PhD in Education from the University of Queensland. Her teaching and research activities have been in the field of language education and assessment. She is particularly interested in democratic, participatory assessment development and evaluation in which students/test-takers are actively engaged.

Appendix A: Interview Protocol

(To be adapted for different interviews)

Introduction

Researcher introduces herself and the research, answers questions from interviewee, and obtains interviewee's signature on consent form.

Content Area 1: Background Information and Motivations to Take the Test

1. Can you please tell me about your background?
2. Why did you decide to take the test?
3. How important was the test score to you?

Content Area 2: Experiences and Perceptions of the Test and the Score

1. You said in the survey that you have taken the test ... times. Was it a pleasant experience every time you took it? Why/Why not?
2. Was everything alright with the test and its administration, etc.?
3. Did you have chance to do your best on the test? If not, what interfered with your performance?
4. In your opinion, what are the key factors that determine one's score on the test? Please rank them from the most important to the least important.
5. In the survey, you (agreed/disagreed) that the test effectively measures your English ability at the time of taking the test. Can you tell me why?
6. In the survey, you (agreed/disagreed) that the test score reflected your performance. Can you tell me why?
7. In the survey, you (agreed/disagreed) that the test score can help predict one's ability to understand and use English in their (study/work/life) in an English environment. Can you tell me why?

Content Area 3: Experiences and Perceptions of Score Interpretation and Use

1. You answered in the survey that you took the test for the purpose of..... How important was the test score compared to the other requirements? Was its weighting appropriate?
2. What could be the institution's purposes of setting this score requirement?
3. What do you think of the minimum required (overall/total) score and sub-scores?
4. Did the institution that asked you to submit that test score (university/scholarship committee/immigration department) also considered other evidence of English proficiency that you possessed? Is it fair for them to (not) do so? Why/why not?
5. Did this whole process have any positive or negative impact on you?
6. How do you compare its positive impact with its negative impact? /In general, did you gain more than lose?

Further comments and wrap-up

Do you have any further comments or suggestions?
Thank you very much for your time and your kind support!

---

[i] https://www.ielts.org/

[ii] https://www.ets.org/toefl/ibt/about/who_accepts_scores

[iii] https://immi.homeaffairs.gov.au/help-support/meeting-our-requirements/english-language/competent-english

[iv] https://www.immigration.govt.nz/about-us/media-centre/news-notifications/new-zealand-residence-programme-changes/nzrp-smc

[v] http://webarchive.nationalarchives.gov.uk/20110413140733/http://www.ukba.homeoffice.gov.uk/sitecontent/applicationforms/pbs/approvedenglishtests.pdf

[vi] https://www.canada.ca/en/immigration-refugees-citizenship/services/canadian-citizenship/become-canadian-citizen/eligibility/language-proof/step-2.html

[vii] See https://www.ielts.org/teaching-and-research/research-reports and https://www.ets.org/toefl/research/