



Title

Validity Arguments of the Speaking and Listening Modules of International English Language Testing System: A Synthesis of Existing Research

Author

Vahid Aryadoust

Nanyang Technological University

Biodata

Vahid Aryadoust is a PhD candidate in applied linguistics at the National Institute of Education (NIE) in Singapore. He has written numerous books and articles on English grammar instruction, applied linguistics, and language and psychosociological assessment. As co-principal researcher in the Spaan Fellowship Program of the University of Michigan, he investigated the construct validity of the Michigan English Language Assessment Battery (MELAB) listening test. His current research focuses on listening assessment, and particularly on the validity arguments of standardized listening tests. Correspondences should be sent to Vahid Aryadoust, ELL AG, NIE, 1Nanyang Walk, Singapore, 637616. Email: ng0666907m@stdmail.nie.edu.sg

Abstract

This study reviews research into the IELTS Speaking and Listening modules to build a validity argument for them. Based on Kane's (1992, 2001, 2004) validity argument framework, the researcher postulates seven assumptions to examine the two modules' interpretive arguments, as well as the sufficiency and efficacy of research conducted on them. The Speaking Module has been thoroughly studied in many respects, but its validity argument is nevertheless seriously compromised because IELTS has yet to articulate a constituent theory of second-language speech on which the module's analytic scoring system is based, and because a number of studies have shown very

limited correlations between performance on the module and performance in target language domains. The Listening Module is the least-researched module of the test, and is in urgent need of investigation before a validity argument can even be attempted.

Keywords: IELTS; interpretive argument; listening; speaking; validity argument

1. Introduction

International English Language Testing System (IELTS) was established in 1989. To date, fifteen rounds of research designed to improve the test have been jointly commissioned by the British Council, the University of Cambridge English for Speakers of Other Languages (ESOL) Syndicate, and the International Development Program of Australian Universities and Colleges (IDP). These projects have examined very diverse aspects of the test, from the language testing theory underpinning it to test taker attitudes and classroom training methods. As a result of this research, the test has undergone changes, and the IELTS of 2010 is quite different from the IELTS of 1989.

Today, many universities whose medium of instruction is English have adopted IELTS as an entry requirement for non-native English speakers. Many require a score of at least six out of nine on the test as a threshold of minimum proficiency. Students with lower scores may be refused admission or asked to take supplementary English language programs even if they have already demonstrated other qualifications.

In the present study, I review existing research on the IELTS Speaking and Listening modules, and create a unitary structure of validity arguments that organizes the results of this research. This synthesis will help spot weak areas and direct future attempts to bridge existing gaps.

This article proceeds as follows: It briefly reviews the history of the validity concept and validity arguments; provides an overview of the Speaking Module, including its analytic scoring system, the role of examiner judgment, its structure, the Revision Programs and impact studies, and some miscellaneous topics; highlights the paucity of research on the Listening Module; explains the interpretive arguments of the IELTS Speaking and Listening modules and investigates their plausibility using

the findings of existing IELTS research; and proposes conclusions and guidelines for future research for both the Speaking and Listening modules.

2. Theoretical Framework

The first references to validity arguments were made by Cronbach (1988) and Messick (1988), and several years later were used as a building block of Kane's (1992) validation framework. The concept of validity has continued to evolve since that time: whereas it was once considered a characteristic of measuring tools (see Chapelle, 1999), it is now considered a feature of score uses and interpretations, or, as Messick (1988, p. 3) put it, of the "inferences derived from test scores."

This new definition of validity accepts three types of supporting evidence: construct, content, and criterion (Messick, 1989). Construct-referenced evidence is evidence of the ability of the measuring device to assess what it is intended or claims to measure (Messick, 1989), and is typically based on statistical analysis. Content-referenced evidence is evidence of the degree to which a psychometric tool represents a psychological trait, and is typically based on expert judgment. Criterion-referenced evidence is evidence of the strength of the relationship between test scores and an external criterion, and is divided into concurrent and predictive classes: concurrent evaluations correlate test scores with scores from another standardized test, and predictive evaluations correlate test scores and future academic or work performance.

From this discussion it follows that there are two views of the validity of measuring tools: "theoretical construct" and "pragmatic ascription to ability" (Upshur, 1979, p.85). Predictive validity evaluations, for example, follow the pragmatic ascription to ability: they examine test takers' proficiency in future academic or work environments. Conversely, a study that targets the theoretical construct focuses on whether the test taker has the proficiency to perform the language task at hand. A common truism holds that any attempt to merge these two views would lead to an incoherent definition of validity.

However, Messick (1988) has called for a unitary concept of validation, based around construct-referenced evidence. Messick (1980) argues that construct-referenced evidence is a "unifying concept that integrates criterion and content considerations into a common framework" (p. 1015), and that content- and criterion-referenced evidence do not suffice in validation. Messick argues that content-referenced evidence has an element of subjectivity since it is mainly a function of

expert judgments, and leaves out the psychological processes of test takers, internal structures of the test, and differences in performance across test takers (Messick, 1988, p. 8); and that criterion-referenced evidence's correlation of test scores with future performance on a criterion may compound confusion, because the criterion will need to be validated like the test itself (Messick, 1988, p. 9).

In a later attempt to improve validation, Bachman (2005) argued that the conventional validity approaches are vague and do not provide a clear start and end point in validation. Drawing chiefly on Kane (1992, 2001, 2002, 2004), Bachman introduced a framework to assess the validity argument of any language measuring tool. In Bachman's framework, the researcher systematically examines supporting and attenuating data for a validity argument to reach a final conclusion that supports, weakens, or rejects the main validity claim. Bachman divides validation into two stages, which complement each other. In the formative stage prior to operationalizing the test, an interpretive argument (which Bachman calls an "assessment validity argument") is developed. This argument establishes the intended interpretations of test scores, sets up a web of inferences and assumptions that move from observed scores to the decisions made based on those scores, and "provides guidance as to the type of research needed" (Chapelle, 2008, p. 321). Interpretive arguments are based on assumptions that are temporary, explicit, and "defeasible in the sense that... they can be overturned in a particular case" (Kane, 2004, p. 147). In the summative stage of validation, a validity argument (which Bachman calls an "assessment utility/use argument") is created to back the interpretation. The validity argument comprises qualitative and quantitative data collected by the researcher to support the assumptions and claims made in the interpretive argument.

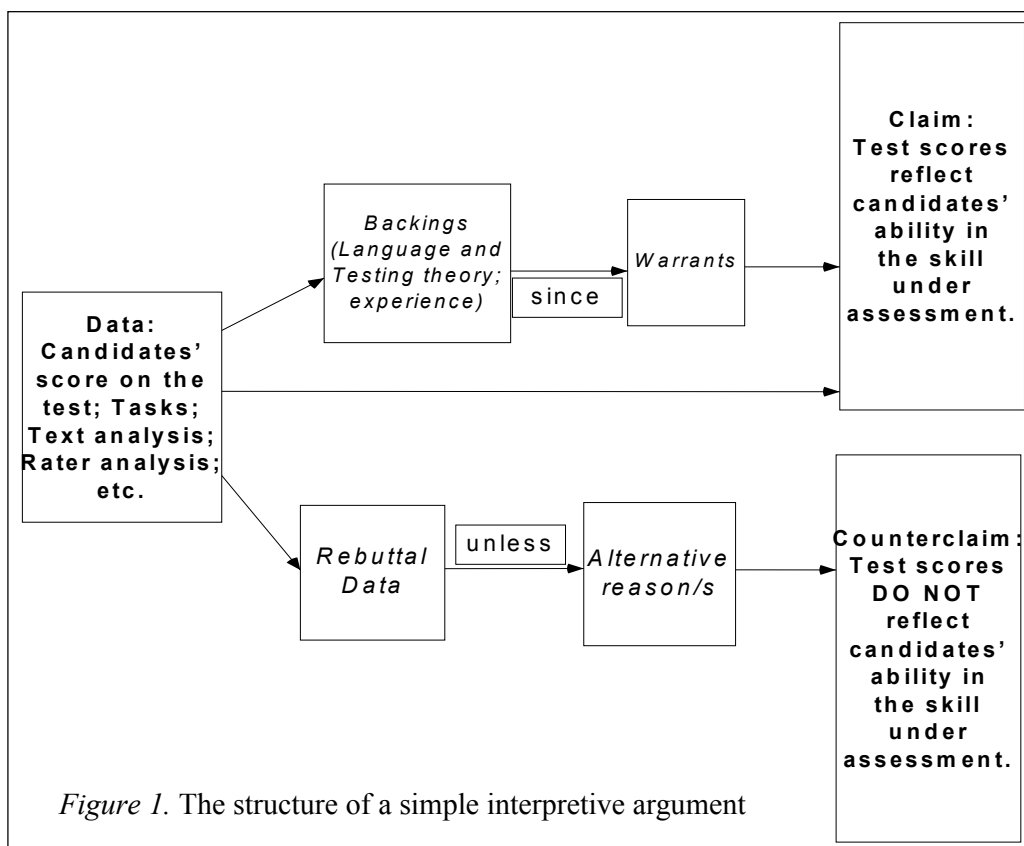
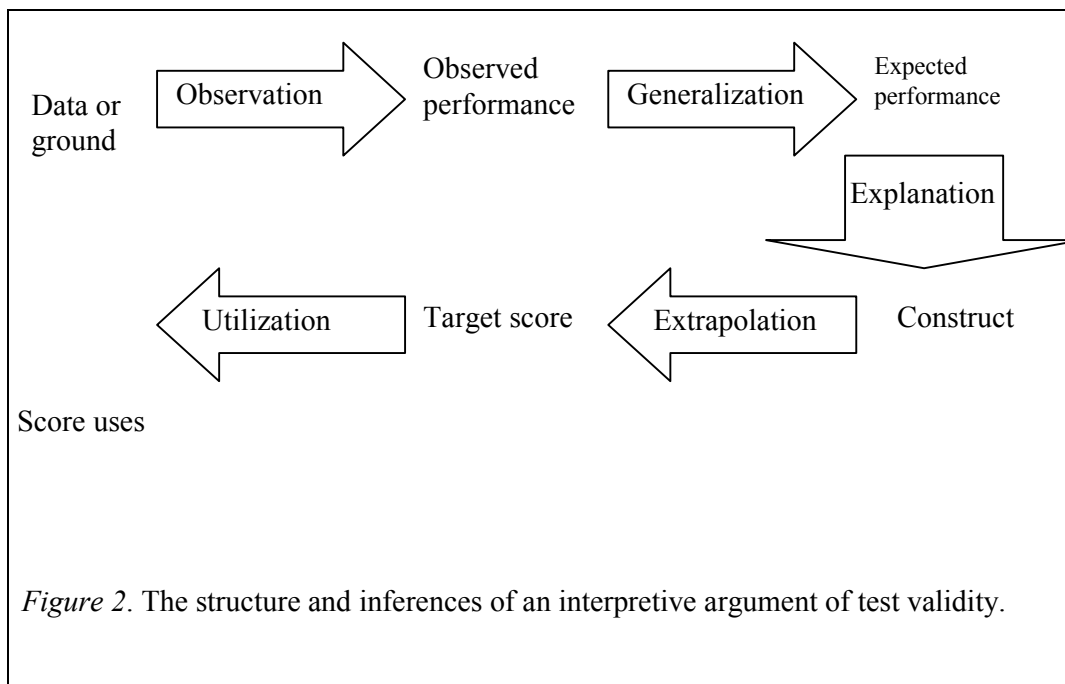


Figure 1 displays an interpretive argument structure comprising a claim, a warrant, data, backing, alternatives, and rebuttals. According to Toulmin (1958), claims are conclusions or interpretations that we intend to draw. A claim relies upon a warrant, one or more general statements that legitimize it (Toulmin, 1958). For example, syllogisms are strong warrants that guarantee a certain outcome. To provide warrants with authority and currency we need backings; these are any form of “assurance,” the “evidence supporting the warrant” (Kane, 2004, p. 148). These backings, in turn, should be drawn from data: expressions of what we have witnessed or observed corresponding to the inference or inferences being made (Mislevy, Steinberg, & Almond, 2003).

To establish a claim as false, a rebuttal is needed. Rebuttals weaken or reject a claim, its backings, or its warrants, unless alternatives are found to resolve them and support the claim.

As noted earlier, an interpretive argument of test validity is the first step in creating a validity argument for a test (Bachman, 2005). Kane (1992) proposed that a properly constructed interpretive argument for a test should follow four inferential

steps: observation, generalization, extrapolation, and utilization. More recently, Kane (2004, 2006), Bachman (2005), Chapelle (2008), and Bachman and Palmer (2010) argued that a fifth inference, called explanation, is needed to bridge the gap between generalization and extrapolation. Figure 2 incorporates this fifth inference, and displays the full structure of a test's interpretive argument.



The observation inference assesses the consistency of scoring methods with measurement processes. The process of measurement should be standardized irrespective of the measuring tool used. In assessment procedures where interpretations and evaluations are performed by an assessor, clear guidelines for the assessor's behavior ought to be stated. Irrelevant and contaminating factors, such as environmental variables (Bachman, 1990), can affect observation inferences seriously.

The generalization inference generalizes the observed scores and interpretations to a broader domain. In this inference, evidence should be collected to assess the degree of variance of test data across the domain. Chapelle (2008) proposed using G-theory and reliability analysis to warrant this inference.

The explanation inference associates test scores to the postulated construct theory. Smith (2005, p. 413) proposed that tests be subjected to a five-stage construct validation program to provide the construct-referenced evidence of validity needed to support this inference.

Because the explanation inference cannot extrapolate scores and observations to non-test behavior, Kane (1992) proposed the extrapolation inference, which involves the use of criterion-referenced evidence. This evidence can be either predictive or concurrent (Chapelle, 2008).

Utilization is the final inference in this framework. It legitimizes the uses to which scores are put. This inference is critical in high-stakes testing, because decision makers may use test scores to admit or reject applicants into their programs.

The present study employs this framework to investigate the validity of scores' use and interpretation in the IELTS Speaking Module, and to suggest the research that remains to be done before such an investigation becomes possible for the Listening Module.

3. Assumptions of the Study

Table 1 lists the assumptions made in this study, and the inferences required to make them. (The term "assumption" here is synonymous with the conventional term "hypothesis": "Assumption" is appropriate to this study because it is an underlying concept in validity arguments.) These assumptions are the basis of the interpretive argument phase of the study.

Table 1. Assumptions of the Study

Category	Definition
Assumption 1	Band scores in the Speaking and Listening Modules are clearly defined and testing conditions are standardized. (Observation) ^a
Assumption 2	Rater judgment does not affect scores in the Speaking Module. (Observation)
Assumption 3	Observed scores mirror expected scores across parallel task versions. (Generalization)
Assumption 4	Listening and Speaking skills have distinct construct definitions which are sufficiently operationalized and represented in the test. (Explanation)
Assumption 5	Construct irrelevancies are not detectable from data analysis. (Explanation)
Assumption 6	IELTS Speaking and Listening tasks portray real-life communication. (Extrapolation)
Assumption 7	Test scores are useful for decision making purposes. (Utilization)

Note. ^a = Every assumption underlies an inference which has been parenthesized.

4. Review of Research on the Speaking Module

Table 2 organizes twenty-eight studies on the Speaking Module into four broad categories, plus a “miscellaneous” category for two studies that did not fit in this framework but rendered relevant information about the module. The “examiners’ judgment” column lists five studies on the behavior of raters; the “analytic scoring and analysis of test and language structures” column includes eight studies that discuss the discourse produced in interviews and the functionality of the module’s analytic scoring system; the “revisions” column includes five studies that report on past and planned revisions to the structure of the Speaking Module; the “impact on future academic performance” column includes eight studies on the relationship of Speaking Module scores to future performance in the target language domain; and the “miscellaneous issues” column includes two relevant studies that fall outside the first four categories.

Table 2. Categorization of Research to Build a Validity Argument for the IELTS Speaking Module

Examiners' judgment	Analytic scoring and analysis of test and language structures	Revisions	Impact on future academic performance	Miscellaneous issues
1. Brown and Hill (1998)	1. Read (2005)	1. Tonkyn (1998)	1. Cotton and Conrow (1998)	1. O'Loughlin (2000)
2. Brown (2000)	2. Read and Nation (2006)	2. Lazaraton (1998)	2. Kerstjen and Nery (2000)	2. Issitt (2008)
3. McDowell (2000)	3. Seedhouse and Egbert (2006)	3. Taylor (2001)	3. Hill, Storch, and Lynch (1999)	
4. O'Sullivan and Lu (2006)	4. Brown (2006a)	4. Taylor and Jones (2001)	4. Allwright and Banerjee (1997)	
5. Merrylees (1999)	5. Brown (2006b)	5. Brooks (2003)	5. Paul (2007)	
	6. Elder and Wigglesworth (2006)		6. Rea-Dickins, Keily, and Yu (2007)	
	7. Weir, O'Sullivan, and Horai (2006)		7. Banerjee (2003)	
	8. Stoyhoff (2009)		8. MacNamara (2006)	

The review presented in this section synthesizes previous research to investigate the validity of test score uses and interpretations in the IELTS Speaking Module. The twenty-eight studies in the five categories presented in Table 2 are the dataset of this section, which I use to provide warrants and backings for the assumptions articulated in Table 1 and build a validity argument for the module.

4.1. Examiners' Judgment

Because examiner judgment is a crucial element of performance tests, several studies on the IELTS Speaking Module have focused on examiner behaviour. Brown and Hill (1998), for example, investigated the effects of examiners' rapport with test takers and the extent to which examiners offered help during the test (for example, by rephrasing test questions) on the structure of the discourse elicited from test takers. They used two discourse levels of interviews, easy and difficult, and observed that examiners who tended to use simpler questions had more problems changing discourse levels.

They also found that examiners substantially influenced test takers' performance. In response, they suggested a behavioural framework for examiners to follow in order to minimize this influence, and recommended that IELTS establish a more restricted code for examiner behaviour. It is worth noting that there was no standardized behaviour set for examiners until that time.

Merrylees (1999) investigated behavior among IELTS examiners in Australia, Honk Kong, Indonesia, Malaysia, the Philippines, and Thailand. At the time, the module had five sections, and Merrylees found that examiners did not elicit enough discourse from test takers in sections four and five. He proposed that "since the object of the IELTS interview is to elicit assessable discourse from the candidates but not to assess their listening comprehension skills, this is clearly a problem" (p. 34). He also reported some discrepancy between prescribed and observed interview durations. This was especially observed of low-proficiency test takers (which is acceptable because they could not produce lengthy stretches of discourse), but Merrylees also found that many test takers assigned a Band 5 score in the test's fourth section were given more time by raters to display their weaknesses than Band 6 scorers. Merrylees found that "these findings call into question the reliability of the test in the hands of examiners who chose not to extend their candidates fully" (p. 34).

Brown (2000) studied examiners' decision making processes to rate them, the relationship between "linguistic and non-linguistic aspects of [test] performance" (p. 49), the salience of the rating criteria, and the probability of a single test taker being assessed differently by different examiners. She used the tape-recorded data she had collected in her previous study on the Speaking Module (Brown & Hill, 1998) and recruited eight raters to assess performance in each interview. She found that topic choice, test condition, and the fact that raters were rating taped interviews affected the scores they assigned to the interviews. Brown's conclusion was almost the same as in the 1998 study: since the rating criteria were holistic and vague, it was not logical to expect consistent examiner behaviour (Brown, 2000).

Similarly, McDowell (2000) examined the efficacy of and discrepancies between IELTS examiners' evaluations of test performance. This study showed that examiners' effectiveness was influenced considerably by their choice of training materials. At the time, examiners could select either self-access to training packs or face-to-face training. Most preferred the self-access method, and McDowell observed behavioural discrepancies based on this choice. McDowell recommended practical

measures to amend the training guidelines and augment the quality of the training packages.

O'Sullivan and Lu (2006) examined effects of examiners' deviations from "the interlocutor frame," and the effects of these deviations on the performance of test takers in the Speaking Module. They found that there were few deviations in the first sections of the module and many in the last part—especially in paraphrasing the questions for test takers—but that these deviations did not influence test takers' performance much.

4.2. Analytic Scoring and Analysis of the Test and Language Structures

Following the 2001 replacement of the holistic scoring system in the Speaking Module by an analytic scale, eight IELTS studies addressed the features and scoring patterns of the new module. Read (2005) used a mini-corpus of 88 speaking performances of IELTS test takers to investigate different aspects of their vocabulary resources. He calculated the token (the total number of words used) and the type of words. More proficient test takers typically used a wider range of vocabulary, but tokens varied widely within each band score. In general, Read found no significant correlation between the use of vocabulary and the band scores assigned to test takers.

Following this study, Read and Nation (2006) attempted to investigate test takers' lexicon and use of "formulaic" language in the module. They found considerable variance in lexicon within band score levels, indicating the limited significance of vocabulary use in assessment, which agreed with Read's (2005) report. They also found that fluency and use of formulaic language and common words were decisive in helping test takers attain high band scores. For instance, among the candidates who were awarded a band score of four (indicating poor performance), the use of formulaic language was not common.

Addressing the interactional make-up of the module, Seedhouse and Egbert (2006) investigated its pragmatic features, such as "turn-taking, sequence, and repair" (p. 164). They found that most examiners adhered to the interview framework. They also contended that the interaction in the Speaking Module is unique in that it resembles both second-language interactions and interactions between students and academic staff at university. They argued that these similarities supported the construct representation of the module on the one hand, and the authenticity of the module and its tasks on the other.

Brown (2006a) made a general assessment of the module's analytic scale. She investigated the degree to which raters were able to differentiate between band scores, their consistency in scoring, and the "salient features" they used to judge test taker performance. Brown found evidence that the analytic scales were effective, but she also highlighted the problems of inference caused by the vagueness of some descriptors, especially in the Fluency and Coherence scale. In addition, she reported many examiners' insistence that the Pronunciation scale was not clear enough for them to confidently assign band scores.

In another study, Brown (2006b) examined the "Speaking Test band descriptors and criteria key indicators" to find what scoring criteria examiners used. She found that as test takers' overall level of proficiency increased, their awarded scores on individual assessment criteria (such as grammar and vocabulary) also increased, supporting the test's construct validity argument. However, the intervals between performance levels were found to be too large, and the test takers within individual levels formed non-homogenous groups. On most measures, the overlarge intervals between band scores meant that test takers falling near band score boundaries were assessed poorly relative to test takers falling squarely within band scores. Brown concluded that examiners were using the assessment criteria in combination, and not in isolation as intended.

The format of the Speaking Module was changed concurrently with the change in scoring. The module, which had been administered in five consecutive stages, was modified into a three-stage format. Two studies examined this new format. Elder and Wigglesworth (2006) investigated the effect of timing on the second phase of the module, which evaluates test takers' planning skills and language proficiency on a timed speaking task. In this phase, test takers are given a topic card containing a main question and a few sub-questions. They are then given one minute to take notes on the questions and prepare a short response. Elder and Wigglesworth's results did not show any essential difference in performance under timed and untimed conditions. The researchers recommended continuing to allot one minute of planning time in the interest of fairness and "face" (perceived) validity.

Weir, O'Sullivan, and Horai (2006) also investigated the effects of changes in preparation time on the difficulty of the second phase of the module and the "amount of scaffolding" given to test takers. ("Scaffolding" is a metaphorical concept that refers to the visible or audible assistance that a more expert member of a culture can

provide in any social setting; it is often analyzed in classroom discourse [Aryadoust, 2006, p. 149].) Weir et al. (2006) worked “within the socio-cognitive perspective of test validation” (p. 119) and employed both qualitative and quantitative research methods. They chose four tasks of the same difficulty level, three of which were altered in terms of “planning time, response time, and scaffolded support” (Weir et al., 2006, p. 143). They found that the unaltered test version produced the highest band scores, and that test takers of different abilities responded differently to the alterations to the other three tests. Contrary to Elder and Wigglesworth’s (2006) research, Weir et al. found that eliminating preparation time could negatively affect test takers’ performance, depending upon their level of proficiency.

Recently, Stoyhoff (2009) criticized the IELTS scoring system, including scoring of the Speaking Module, on the grounds that IELTS scoring lacks an underlying theoretical construct: “At present, the IELTS does not include an explicit theoretical rationale to support the interpretation and use of test scores and one must infer the conceptualizations underlying the test constructs” (p. 18).

4.3. Revisions

The results of commissioned IELTS studies have been applied in a number of IELTS Revision Programs. Several reports have narrated and analyzed these programs. In particular, Taylor (2001) and Taylor and Jones (2001) reviewed five major phases that the IELTS Speaking Module has gone through.

In phase one, “consultation, initial planning and design,” an extensive study was commissioned in June 1998. Taylor and Jones (2001) mention Tonkyn (1998) and Lazaraton’s (1998) articles in this stage. For example:

...along with findings from earlier studies, [Lazaraton’s work] raised the question of how well the existing holistic IELTS rating scale and its descriptors were able to articulate key features of performance at different levels or bands. It was felt that a clearer specification of performance features at different proficiency levels might enhance standardization of assessment. (p. 8)

In the second phase, “development”, the “prototype of the revised test format” was trialled (Taylor & Jones, 2001, p.1). This phase lasted from January to September

1999. Second and third drafts of the test's assessment criteria were then constructed to help examiners understand the rating scale descriptors.

The third phase, "validation," proceeded from October 1999 to September 2000, and generated the following results, as reported by Taylor and Jones (2001):

- A. Pronunciation is a separate construct, but grammatical and vocabulary scales are highly correlated.
- B. There was no need to collapse the rating scales, as the Rasch model fit the data.
- C. Raters' scoring reflected a largely unified interpretation of performance, although some scores were found to misfit.
- D. The test was highly reliable and generalizable. In particular, its G- and Phi-coefficients increased as the number of subscales increased.

The fourth phase, "implementation," lasted from October 2000 to June 2001. In this phase, retrospective studies further compared the new and old rating methods and showed the revised rating methods to be of good quality, and a worldwide examiner retraining was undertaken to implement the new version of the test. Finally, the fifth or "operational" phase included the full operationalization of the revised IELTS Speaking Module worldwide.

Taylor (2001) also explained the process of retraining examiners worldwide. After the first revisions, some Senior Trainers were retrained again at a regional level. The new materials contained "an IELTS Examiner Induction Pack with accompanying video and worksheet [and] an IELTS Examiner Training Pack, with two accompanying videos and detailed Notes for Trainers" (Taylor, 2001, p. 9). After the training session, ninety-nine percent of examiners and trainers described the package contents as "very good" or "fairly good." Their suggestions were to be incorporated into future versions of the packs. The entire process of retraining IELTS examiners worldwide took between four and five months.

Brooks (2003) explained how an observation checklist (OC) used in the speaking section of the First Certificate in English (FCE) and Cambridge Proficiency in English (CPE) tests was converted into a checklist for the IELTS Speaking Module. Since IELTS speaking tests are audiotaped, Brooks described the advantages of audiotaped samples that did not have any animated visual input to distract the users of the OC.

Although the scholarly literature on IELTS from 2007 and 2008 did not focus on the Speaking Module, its results impacted the module heavily: they convinced the Cambridge ESOL Examinations, the British Council, and IELTS Australia to adopt half-band scores, significantly changing the Speaking Module's scoring system; and they led those organizations to resume study of stakeholder attitudes and test impacts as their main research agenda, some of which focused on the relationship between performance on the Speaking Module and future academic performance.

4.4. Impact on Future Academic Performance

The major backing for the extrapolation inference is future performance, the role of the theoretical construct in the task domain. Predictive validity research can assess this relationship by correlating test results with academic performance, as indicated by grade point average (GPA) in the task domain.

Investigating the predictive power of IELTS at the University of Tasmania, Cotton and Conrow (1998) found no significant correlation between GPA and performance on the Speaking Test. Academic staff ratings also did not significantly correlate with test results. More than half of interviewed students believed that IELTS could not bring into focus their areas of weakness in English, and many students reported language problems in their courses despite having received acceptable IELTS scores. The researchers pointed to the subjectivity of the interviews and questions as problems with the Speaking Module.

Two more similar studies were conducted, one by Hill, Storch, and Lynch (1999), and the other at RMIT University in Australia by Kerstjen and Nery (2000). In these studies, IELTS scores did not predict academic performance strongly, although scores on the Reading and Writing Modules were significantly correlated with it in the latter study. The Speaking Module had no significant correlation with students' academic performance, nor did it account for any portion of the observed variance in academic performance in regression analysis. Kerstjen and Nery further argued that the predictive power of the IELTS scores relied on the test takers' field of study. Allwright and Banerjee (1997) also found that IELTS scores above 7.0 indicated a low risk of failure in academia, but they reported no obvious relationship between IELTS scores and academic performance. Likewise, Paul (2007) suggested that "Language production at a micro level similar to that in IELTS tasks is not necessarily an indicator of overall language adequacy at a macro level or successful task

completion [in the IELTS Speaking Test]” (p. 205). Most directly, Rea-Dickins, Keily, and Yu (2007) found the Speaking Module to be a poor predictor of test takers’ future academic performance.

The studies reviewed above generally reported that academic performance is not solely a matter of language production and proficiency, but depends on other significant variables such as personality, socioeconomic background, nationality, and affective/cognitive factors. Rea-Dickins et al. (2007) found that acceptable scores on the Speaking Module do not guarantee a lack of communication problems in tutorials, since other intervening factors such as stress and the use of specialised academic language may impede students’ ability to communicate effectively.

Further, Banerjee (2003) reported that many university admissions decision makers lack a clear understanding of the meaning and interpretation of IELTS scores. Rea-Dickins et al. (2007) confirmed this finding, and added that even when students’ scores on the sub-skills of the test satisfy a program’s admission requirements, students are found to “lack critical thinking and evaluative skills” (p. 117).

Finally, McNamara (2006) raised the concern that IELTS scores are being interpreted and used in use contexts where the test has not been validated yet, such as “immigration selection in Australia and other countries” (p. 30).

4.5. Miscellaneous Issues

Two issues are discussed here: the effect of gender on examiners, and methods for enhancing candidates’ performance. The former has bearing on the test’s construct-referenced evidence of validity and the latter on its consequential validity—that is, on the consequences of the test on the educational system.

Gender has been a longstanding topic of discussion in performance tests. In high-stakes tests such as IELTS, its effects on both examiners and test takers may cause bias and unfairness. O’Loughlin (2000) investigated the effect of gender on the Speaking Module rating process and the discourse produced during the interview, and found no systematic differences.

Some IELTS research has also focused on the possibility of enhancing candidates’ performance. In an uncommissioned study, Issitt (2008) recommended using three strategies in IELTS preparation courses for the Speaking Module: materials to build candidates’ confidence, materials to encourage candidates to “think

critically,” and “close inspection and utilization” of the rating criteria available from IELTS materials. Issitt argued that these strategies are especially effective for alleviating excessive test taker stress, although he cautioned that they had not been tested on large samples of test takers.

5. Review of Research on the Listening Module

The IELTS Listening Module is the least-researched IELTS module. Until very recently, no published studies focused exclusively on the module.

Coleman and Heap (1998) examined IELTS Listening Module and Reading Module rubrics for evidence of confusion or misunderstanding resulting from the concurrent processing of auditory and visual input. They found the test rubrics to be clear, with “only minor areas which need tightening up” (p. 70), but that some test items were not as clearly written as the rubrics.

In a survey study, Merrylees (2003) showed that 78% of IELTS test takers had no problem understanding rubrics, instructions, and items in the Listening Module. Some individuals had trouble with the subject of the stimuli; 23% believed that the pace of speech delivery was too fast, whereas 44% believed it was not; participants categorically agreed that the more they moved through the test, the more the test became difficult; and some participants had difficulties understanding the British accent.

A number of predictive validity studies have examined the Listening Module alongside the other modules. Kerstjen and Nery (2000) found no correlation between test takers’ Listening Module scores and their academic performance in business courses. They argued that, “It may be that, as the Listening test does not test academic listening skills, it does not accurately predict the kind of listening skills that students in these Business courses are required to develop” (p. 96). Cotton and Conrow (1998) found that Listening Module scores were actually negatively correlated to academic performance for two groups of students: the correlation was weakly negative (-0.19) for a first group of twenty-six test takers after a full academic year, moderately negative (-0.58) for a second group of seventeen students after their first semester, and moderately negative (-0.56) for the same seventeen test takers in their second semester. Cotton and Conrow also interviewed students on the face validity of the Listening Module. Some commented that the module was too general and did not suit academic purposes, others that the test was easy, and others

that performance was affected by test conditions such as the distance of the test taker from the tape player.

Working in Australia, Ingram and Bayliss (2007) reported a weakly positive correlation between test takers' Listening Module scores and their self-evaluation of their lecture comprehension skills, but that "there was not, in fact, any obvious relationship between IELTS Listening scores and the amount of understanding of classes and discussions, according to the participants" (p. 162). Students experienced difficulty when interacting with their peers, although they felt better able to comprehend teaching staff. They further observed that most students did not take many notes in lectures and that their notes did not contain much "substantive content" (p. 179). All students had problems with the speed of speech delivery and the Australian accent, particularly in their first semester.

IELTS has also commissioned four studies on the Listening Module, two of which have only recently been published. Table 3 presents these studies.

Table 3. Unpublished IELTS-Commissioned Studies on the Listening Module

Researcher	Study title	Comments
J. Field	A cognitive validation of the lecture-listening component of the IELTS listening paper.	Round 11, 2005
R. Badger and O. Yan	The use of tactics and strategies by Chinese students in the listening component of IELTS.	Round 12, 2006
R. Breeze and P. Miller	Predictive validity of the IELTS Listening Module as an indicator of student coping ability in English-medium undergraduate courses in Spain.	Round 14, 2008; Unpublished
F. Nakatsuhara	The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking test	Round 15, 2009; Unpublished

Field (2009) found that gap filling and MCQ items in the fourth section of the test engage construct-irrelevant skills. He stated that:

The former [i.e., gap filling] has unfortunate effect on focusing candidate attention at word level and providing *gratis* a great deal of structure of the lecture which it should be the listener's responsibility to construct. The latter [i.e., MCQ] imposes heavy reading demands. Both foster a practice of switching attention away from the recording to the written modality. (Field, 2009, p. 47)

Field (2009) pointed out that IELTS listening test format confines test takers' cognitive processing because they hear the audio input a single time—which has been a long-established tradition in many Cambridge exams—and have to rely heavily on their reading skills while listening to the audio material. Shifting to different modes of input (i.e., listening and reading), test takers can become overanxious if they miss an item. This invites test-wiseness strategies and those who are test-wise stand a better chance to answer the item successfully. In addition, the test structure and the construct irrelevant processes jointly lead test takers to a superficial comprehension of the aural text. The lower level comprehension processes coupled with the limited redundancy of the audio materials and their richness in terms of detailed information make a high cognitive demand of test takers, which is beyond the discourse structure of the university lectures.

The findings of Field's (2009) qualitative research resonate with Aryadoust's (in press) Rasch-based differential item functioning study of the IELTS listening test. Aryadoust found that the listening construct in the IELTS test was underrepresented, which "is probably an important cause of the lack of significant correlation between test results and academic performance observed in previous studies" (Aryadoust, in press, p. ##). He further reported that "short answer items, which feature prominently on the test, are likely to be biased in favor of higher-ability listener subgroups in listening comprehension because of these test takers' ability to apply swiftly what they have understood" (Aryadoust, in press, p. ##). Like gap filling and MCQ items, these items are not effective to measure the listening construct due to the cognitive limitations they impose on listeners. Finally, that low-ability test takers who are test-wise appear to be taking advantage of this fact leads to flawed test results.

In another study, Badger and Yan (2009) found no significant differences between pre-undergraduate and pre-postgraduate students taking the IELTS listening test in using strategies and tactics. Although Badger and Yan believed that this finding provides a piece of evidence supporting the validity arguments of the IELTS listening test, they argued that some of the written text in the examination should be replaced by listening texts, which is in agreement with Field's (2009) finding. They agreed with Field in that an efficient alternative to the present IELTS system is a test system where the test takers hear the audio materials twice. Another solution, as Badger and

Yan argued, would engage maintaining the current practice where test takers listen once without reading the test items. They are encouraged to take notes and use them to answer the test items which are displayed or read to them later. This alternative, which would lead to a major change in the structure of the test and the way the construct is operationalized, is the testing method which has been recently adopted by the Test of English as a Foreign Language (TOEFL).

Because of the paucity of published research on the IELTS Listening Module, and the findings of the studies above which generally do not support its functionality, many outstanding questions preclude the establishment of a comprehensive validity argument for the module.

6. Results and Discussion

IELTS Speaking Module studies have prioritized improving the assessment power of the module and improving the theoretical construct. Findings from these studies have helped test developers devise newer formats of assessing speaking. Now we are in a position to set forth the validity argument for the IELTS Speaking Module. This validity argument is presented in Table 4. All components in this argument, and its division into seven inference classes, are extracted from the main research assumptions that have guided IELTS researchers.

Table 4. Validity Argument for the IELTS Speaking Module, Based on Data from IELTS Studies and Revision Programs

Inference	Warrants	Backings/rebuttals
<i>Observation (A)</i> Band scores in the Speaking Module are clearly defined.	According to IELTS descriptions in Research Notes, band scores are intended to precisely describe speaking ability.	<i>Backings:</i> Although the findings of Brown and Hill (1998) and Brown (2000) questioned the module's scoring system, the later adoption of an analytic rating scale largely resolved these concerns. Tonkyn (1998), Lazaraton (1998), and Taylor (2001), reported on a Revision Program and retraining of examiners; McDowell (2000) recommended ways to improve the training packages; Brown (2006b) concluded that the Speaking Test and its scores are valid and the band scores are clear. Half-band scores were introduced to the scoring system in 2008 to clarify the scoring process.
<i>Observation (B)</i> Rater judgment does not significantly affect	IELTS has made a concerted effort to neutralize the effect of examiner judgment on	<i>Backings:</i> O'Sullivan and Lu (2006) found that examiner deviations did not adversely affect test taker performance. Seedhouse and Egbert (2006) found that most examiners adhered to the

scores.	scores.	interview framework.
<i>Generalization</i> Observed scores approximate expected scores across parallel test/task versions.	Significant G-theory and high reliability coefficients support the generalizability of the module.	<i>Backings:</i> Merrylees's (1999) study, training, and revision programs support the generalization inference. The IELTS official Web site reports an inter-rater correlation of 0.77 and a G-coefficient of 0.86 for the test as a whole.
<i>Explanation (A)</i> Second-language speaking as operationalized in IELTS entails representative components which are measured in the test.	The test assesses speaking as a component of language proficiency using an analytic scoring system comprising vocabulary, grammar, discourse, and cohesion/coherence of ideas.	<i>Backings:</i> Read and Nation (2006) found that fluency and the use of formulaic language and common words affect scores. O'Sullivan and Lu (2006) found that raters mainly followed the module's analytical framework, comprising vocabulary, grammar, discourse, and cohesion/coherence. <i>Rebuttals:</i> Read (2005) found that proficient students often had wider vocabularies, but otherwise found, with Read and Nation (2006), that vocabulary has little relationship to scores. Stoyhoff (2009) found that the module's analytical framework is not based on any explicit underlying theory of second-language speaking.
<i>Explanation (B)</i> The current practices in the IELTS Speaking Test do not impede candidates' real performance.	Standardized tests should not expose students to impeding variables: excessive stress and construct irrelevancies should be eliminated.	<i>Backings:</i> Weir, O'Sullivan, and Horai (2006) showed that test takers performed better under the module's standard time and support constraints than under any of three alternatives. Conversely, Elder and Wigglesworth (2006) found that preparation time did not significantly affect test taker performance. The IELTS Revision Programs unified the test's rating system.
<i>Extrapolation</i> IELTS Speaking tasks which operationalize the construct can predict the future performance of test takers.	The IELTS Speaking Module predicts candidates' ability to complete tasks they may encounter in academia and social life.	<i>Backings:</i> Seedhouse and Egbert (2006) concluded that IELTS Speaking tasks were quite similar to both university and second-language interactions. <i>Rebuttals:</i> Most impact studies—by Allwright and Banerjee (1997), Banerjee (2003), Cotton and Conrow (1998), Hill, Storch, and Lynch (1999), Kerstjen and Nery (2000), Paul (2007), Rea-Dickins, Keily, and Yu (2007)—showed no correlation between academic performance and Speaking Module scores.
<i>Utilization</i> Test scores are useful for making admissions and remedial program placement decisions.	Standard setting, available score interpretation materials, and positive attitudes of stakeholders are evidence of the utility of scores.	<i>Backings:</i> Impact studies showed that the majority of candidates, academic staff, and decision makers considered IELTS a fair test. Also, regular Revision Programs, which retrain examiners and rethink benchmarks, represent a policy of standard setting. <i>Rebuttals:</i> Banerjee (2003) and Rea-Dickins et al. (2007) reported that many admissions decision makers are unqualified to interpret IELTS scores, and McNamara (2006) reported the use of IELTS scores in untested domains, such as immigration.

The two observation inferences are attenuated by studies by Brown and Hill (1998), Brown (2000), and MacNamara (1996), but these concerns have been largely

addressed by the new analytic scoring system developed in 2001 and modified in 2008, and by the introduction of a number of guidelines for examiner behaviour.

The generalization inference is made based on observed scores. G-theory, reliability coefficients, and similar analyses provide support for this inference.

The explanation inference is supported by pertinent warrants which are based on the definition of construct-referenced evidence of validity. However, the inference is compromised by the lack of an IELTS-sanctioned model of the constituent structure of second-language speech. The Speaking Module's analytic scoring system assumes that second-language speech comprises vocabulary, grammar, discourse, and cohesion/coherence, but IELTS provides no explicit theoretical model to support this assumption. The lack of construct models is a problem extending to the entire test (Stoynoff, 2009).

The extrapolation inference is also highly problematic for the Speaking Module. Supporting evidence for this inference is admittedly difficult to collect through predictive validity studies, since academic performance is a function of many factors other than language proficiency; nevertheless, tested language proficiency should be able to predict a significant portion of future academic performance. The preponderance of existing literature indicates that the Speaking Module is inadequate in this regard.

The utilization inference also has significant complications. It is warranted by institutionalized standard-setting procedures, available score interpretation materials, and the largely positive attitudes of stakeholders toward IELTS generally, but rebutted by studies that show that admission officers and academic staff are poorly equipped to interpret IELTS scores, and by the misuse of IELTS test results for decision making in unapproved contexts such as immigration and job selection.

7. Conclusion

This article reviews and organizes research on the IELTS Speaking and Listening modules. The existing literature lends some support to the Speaking Module's validity argument, but also presents important evidence against it: the module's scoring system does not refer to any explicit underlying theory of second-language speech; performance on the module does not appear to predict future performance in the target language domains; and test results are often interpreted in unintended contexts, or by decision makers with limited understanding of the test. In the five-stage inferential

framework adopted in this article, the module's explanation, extrapolation, and utilization inferences need improvement. Focusing IELTS studies on these three stages of validation and inference-making could help improve the test.

There is also an urgent need to support the test structure and validity argument of the Listening Module, whose almost total lack of scholarly attention contrasts sharply with its complexity and the complexity of the listening skill.

Acknowledgements

Acknowledgements are gratefully extended to Professor Philippa Mungra and two anonymous reviewers for their comments on an earlier draft of this article.

References

- Allwright, J., & Banerjee, J. (1997). *Investigating the accuracy of admissions criteria: A case study on a British university*. Lancaster: Institute for English Language Education, Lancaster University.
- Aryadoust, S. V. (2006). *A dictionary of sociolinguistics, plus pragmatics and languages*. Shiraz: Faramatn Publication.
- Aryadoust, S. V. (in press). Differential item functioning in while-listening performance tests. *International Journal of Listening*.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in the real world: Designing language assessments and justifying their use*. Oxford: Oxford University Press.
- Badger, R., & Yan, X. (2009). The use of tactics and strategies by Chinese students in the Listening component of IELTS. In Thompson, P. (Ed.), *IELTS research report*, (Vol. 9), (pp. 67-96). UK and Australia: British Council and IELTS Australia.
- Banerjee, J. (2003). *Interpreting and using proficiency test scores*. Unpublished doctoral dissertation, Lancaster University, Lancaster, UK.
- Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. *Research Notes*, 11, 20-21.

- Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS Research Reports* (Vol. 6) (pp. 49-84). Canberra: IELTS Australia, Pty Ltd.
- Brown, A. (2006a). An examination of the rating process in the revised IELTS speaking test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Report* (Vol. 6) (pp. 41-70). Canberra: IELTS Australia, Pty Ltd & British Council.
- Brown, A. (2006b). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports* (Vol. 6) (pp. 71-90). Canberra: IELTS Australia, Pty Ltd & British Council.
- Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Wood (Ed.), *IELTS Research Reports* (Vol. 1) (pp. 1-19).
- Chapelle, C. (1999). Validity in language testing. *Annual Review of Applied Linguistics*, 19, 254-274.
- Chapelle, C. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson, (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 319-350). New York: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C.A., Chapelle, M.K., Enright, & J.M., Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York: Routledge.
- Coleman, G., & Heap, S. (1998). The misinterpretation of directions for the questions in the Academic Reading and Listening sub-tests of the IELTS test. In S. Wood (Ed.), *IELTS Research Reports* (Vol. 1) (pp. 38-71). Canberra: IELTS Australia, Pty Ltd.
- Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a sample of international students studying at the University of Tasmania. In S. Wood (Ed.), *IELTS Research Reports* (Vol. 1) (pp. 72-115). Canberra: IELTS Australia, Pty Ltd & British Council.
- Elder, C., & Wigglesworth, G. (2006). *An investigation of the effectiveness and validity of planning time in Part 2 of the IELTS Speaking Test*. In P. McGovern, & S. Walsh (Eds.), *IELTS Research Reports* (Vol. 6) (pp. 13-40). Canberra: IELTS Australia, Pty Ltd & British Council.

- Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS Listening paper. In Thompson, P. (Ed.), *IELTS research report*, (Vol. 9), (pp. 17-66). UK and Australia: British Council and IELTS Australia.
- Hill, K., Storch, N., & Lynch, B. (1999). A comparison of IELTS and TOEFL as predictors of academic success. In R. Tulloh (Ed.), *IELTS Research Reports*, (Vol. 2), (pp. 52-63). Canberra: IELTS Australia Pty Limited.
- Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, part 1. In L. Taylor (Ed.), *IELTS Research Reports* (Vol. 7) (pp. 137-204). Canberra: IELTS Australia, Pty Ltd & British Council.
- Issitt, S. (2008). Improving Scores on the IELTS Speaking test. *ELT Journal*, 62(2), 131-138.
- Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31-41.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135-170.
- Kerstjen, M., & Nery, C. (2000). Predictive validity in the IELTS test. In R. Tulloh (Ed.), *IELTS Research Reports* (Vol. 6), (pp. 85-108). Canberra: IELTS Australia, Pty Ltd.
- Lazaraton, A. (1998). *An analysis of differences in linguistic features of candidates at different levels of the IELTS Speaking Test*. Unpublished study commissioned by UCLES.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- McDowell, C. (2000). *Monitoring IELTS examiner training effectiveness*. In R. Tulloh (Ed.), *IELTS Research Reports* (Vol. 3), (pp. 109-142). Canberra: IELTS Australia, Pty Ltd.
- McNamara, T. (1996). *Measuring second language performance*. Longman: New York.

- McNamara, T. (2006). Validity and values: Inferences and generalizability in language testing. In M. Chalhoub-Deville (Ed.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 27-46). Philadelphia, PA, USA: John Benjamins Publishing Company.
- Merrylees, B. (1999). An investigation of speaking test reliability. In R. Tulloh (Ed.), *IELTS Research Reports*, (Vol. 2) (pp. 1-52). Canberra: IELTS Australia Pty Limited.
- 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. In R. Tulloh (Ed.), *IELTS Research Reports*, (Vol. 4) (pp. 1-58). Canberra: IELTS Australia Pty Limited.
- Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, 35, 1012-27.
- Messick, S. (1988). *Meaning and values in test validation: The science and ethics of assessment*. Research Reports. New Jersey: Educational Testing Service, Princeton.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: ACE and Macmillan.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.
- O'Loughlin, K. (2000). The impact of gender in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS Research Reports*, 3, (pp. 1-28). Canberra: IELTS Australia, Pty Ltd.
- O'Sullivan, B., & Lu, Y. (2006). The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports*, (Vol. 6) (pp. 119-160). Canberra: IELTS Australia, Pty Ltd.
- Paul, A. (2007). IELTS as a predictor of academic language performance, Part 2. In P. McGovern & S. Walsh, *IELTS Research Reports*, (Vol. 6) (pp. 205-240). Canberra: IELTS Australia, Pty Ltd.
- Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes*, 20, 12-16.

- Read, J., & Nation, P. (2006). *An investigation of the lexical dimension of the IELTS Speaking Test*. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports* (Vol. 6) (pp. 207-231). Canberra: IELTS Australia, Pty Ltd & British Council.
- Rea-Dickins, P., Keily, R., & Yu, G. (2007). Student identity, learning and progression: The affective and academic impact of IELTS on 'successful' candidates. In P. McGovern & S. Walsh, *IELTS Research Reports*, (Vol. 6) (pp. 56-136). Canberra: IELTS Australia, Pty Ltd.
- Seedhouse, P., & Egbert, M. (2006). The interactional organization of the IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS Research Report*, (Vol.6) (pp. 161-206). Canberra: IELTS Australia, Pty Ltd & British Council.
- Stoynoff, S. J. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching*, 42(1), 1-40.
- Taylor, L. (2001). Revising the IELTS speaking test: retraining the IELTS examiners worldwide. *Research Notes*, 6, 9-11.
- Taylor, L., & Jones, N. (2001). Revising the IELTS speaking test: Retraining the IELTS examiners worldwide. *Research Notes*, 6, 9-11.
- Tonkyn, A. (1998). *Reading University/UCLES IELTS rating research project*. Unpublished Interim Report.
- Toulmin, S. E. (1958/2003). *The uses of argument*. Cambridge: Cambridge University Press.
- Upshur, J. A. (1979). Functional proficiency theory and a research role for language tests. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 75-100). Washington DC: TESOL.
- Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, 27, 1-21.
- Weir, C., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: An intra-task perspective. In P. McGovern & S. Walsh (Eds.), *IELTS Research Reports* (Vol.6) (pp. 119-160). Canberra: IELTS Australia, Pty Ltd & British Council.