Building a Validity Argument for a Listening Test of Academic Proficiency

Building a Validity Argument for a Listening Test of Academic Proficiency

By

Vahid Aryadoust

CAMBRIDGE SCHOLARS PUBLISHING

Building a Validity Argument for a Listening Test of Academic Proficiency,
by Vahid Aryadoust

This book first published 2013

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-4863-8, ISBN (13): 978-1-4438-4863-3

TABLE OF CONTENTS

Appendices

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This book was written based on the findings of my PhD studies at the National Institute of Education (NIE) of Nanyang Technological University (NTU), Singapore. It would have been impossible to write this book without the aid and support of the kind people around me, to whom I am truly indebted and thankful.

First, I am sincerely and heartily grateful to Professor Christine Goh at NIE for her invaluable support and guidance. I also wish to express my appreciation to Professors Lee Ong Kim, Hu Guangwei, Timothy Teo, Chew Lee Chin, Antonia Chandrasegaran, and Antony Kunnan who supported, taught, and helped me.

For their close cooperation on data collection, my thanks are due to Philippe Caron and Kirsteen Donaghy of the British Council (Malaysia center), Colin Wilson of the Ascend Language School in Singapore, Elyse Go from a language center in the Philippines, and the students in these centers; I also wish to thank Bita Beheshti and Bita Irani for collecting the Australian and Malaysian data; and Fred Mayer, and Kavishna Ranmali for their productive comments on the manuscript.

My thanks go to PGELT students at NIE and engineering students from Nanyang Technological University who voluntarily participated in the study. I would also like to thank my family and my classmates who boosted me morally.

Finally, I would like to acknowledge that portions of this book have been published in two articles, as follows:

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of IELTS listening test. *International Journal of Listening, 26*(1), 40-60.

Aryadoust, V., Goh, C., & Lee, O. K. (2012). Developing an academic listening self-assessment questionnaire: A study of modeling academic listening. *Psychological Test and Assessment Modeling, 54*(3), 227-256.

Permission has been granted by the publishers to reprint these articles.

SUMMARY

This book builds a validity argument for a listening test of academic proficiency. To date, various models of validation have emerged in psychological and educational assessment research, which can be classified into two major groups: traditional models which view validity as a multicomponential concept including, for example, content, construct, and predictive validity (Cronbach, 1969; Cronbach & Meehl, 1955) and modern models conceptualizing validity as a unitary construct which is evaluated through argumentation (Messick, 1989; Kane, 2002, 2004, 2013). I have adapted the recent validity concept in the present study due to the advantages that it confers: it characterizes validity as a unitary concept; it has a clear start and end point; and it involves both supporting and rebutting evidence of validity.

It is argued that the IELTS listening test is a while-listening-performance (abbreviated here as WLP) test; test takers read and answer test items *while* they are listening to the oral stimuli, and thus engage in the following simultaneous activities: (a) reading test items; (b) listening to the oral text; (c) supplying or choosing the correct answer; and (d) following the oral text to move to the next test item. The simultaneity of test performance raises a number of theoretical questions. For example, previous IELTS listening studies indicate: (a) a role for guessing in the test items that invite word-level comprehension, despite the conventional wisdom that guessing is a confounding factor only in multiple choice questions (MCQs); (b) that failure to answer a test item correctly does not necessarily imply a lack of comprehension; (c) differential item functioning (DIF) in multiple test items; (d) that the test items of the IELTS listening module primarily evaluate test takers' understanding of details and explicitly stated information and making paraphrases—the ability to make inferences, interpret illocutionary meaning, and draw conclusions are not tested; and (e) a weak association with external measurement criteria of listening comprehension.

To address these issues, this book proposes five main research questions addressing meaningfulness of scores and bias of the test. The questions are concerned with the listening sub-skills that the test engages, dimensionality of the test, variables that predict item difficulty parameters, bias across age, nationality, previous exposure to the test, and gender, as

well as predictive-referenced evidence of validity. Subsequently, it reviews relevant literature for each research question in Chapters One and Two and connects each question to the corresponding validity inferences along with their underlying warrants, assumptions, and backings.

It was found that: (a) the cognitive sub-skills measured by the test are critically narrow, thereby underrepresenting the listening construct; (b) both construct-relevant and -irrelevant factors can predict item difficulty; (c) construct irrelevant factors seem to have contaminated the test structure and scores; (d) differential item functioning seems to have affected test items; and (e) test scores correlate moderately with a criterion listening test, and weakly with a self-assessment listening questionnaire developed in the study. By using these findings as backings for one or more validity inferences, the study builds a validity argument framework which organizes these thematically related findings into a coherent treatment of the validity of the listening test. This validity argument, however, is not well-supported and is attenuated in multiple cases by the findings of the studies reported in the monograph.

CHAPTER ONE

INTRODUCTION

> The argument-based framework [for test validity] is quite simple and involves two steps. First, specify the proposed interpretations and uses of the scores in some detail. Second, evaluate the overall plausibility of the proposed interpretations and uses. (Kane, 2012, p. 4)

1.1. Building a Validity Argument

The role of test developers and publishers is to furnish evidence that test scores are being used and interpreted reliably and validly (Chapelle, Enright, & Jamieson, 2008a, b, 2010). This evidence is especially important in high-stakes tests, tests whose results may meaningfully impact test takers' academic or career prospects (Allington & McGill-Franzen, 1992). Given the impact of high-stakes tests on test takers, families, schools, and professional institutions (Barksdale-Ladd & Thomas, 2000), many of these tests, including English proficiency tests, fund research programs to investigate the validity of inferences made from test scores(Bachman & Purpura, 2008).

A test's validity is the extent to which it measures what it is intended to measure. Messick (1994) distinguishes two prevailing approaches to validating test score uses and interpretations: competency- and task-based approaches. These two approaches have also been synthesized into the argument-based approach to validity.

Competency based approach. The competency-based approach takes observed scores to represent performance, and uses them to infer individuals' endowment of the underlying construct under assessment—that is, their ability level (see Chapelle et al., 2008b, p. 3). Messick (1994) described this approach as follows:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are

> otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (Messick, 1994, p. 17)

Competency-based tests may suffer from the presence of construct-irrelevant factors, which can confound measurement. For example, students taking a listening comprehension test may be unable to read the test items fast enough, confounding the interpretation of their scores(Chapelle et al., 2008a, b, 2010).

Task-based approach. The task-based approach to inference-making exposes test takers to a series of tasks designed to elicit the same performance that they would use in a real-world situation (Messick, 1994). Tests that use this approach simulate the intended context as faithfully as possible, so that observed scores can be used to infer likely performance (see Chapelle et al., 2008b, p. 4).

The task-based approach is more difficult to apply than the competency-based approach. Creating an environment that simulates the intended context entails additional complexity and requires controlling for multiple additional variables. Thus, Messick (1994) suggests that the administrators of task-based tests compromise on both the "comprehensiveness" of simulated "situational aspects" and their "fidelity" to real-world tasks. The scores awarded in a task-based test may also be less generalizable than those awarded in a competency-based test.

Argument-based approach. Because of the shortcomings of both approaches to test score interpretation, language researchers have begun to adopt an "argument-based" approach combining both perspectives, as well as multiple types of validity evidence, to support the inferences made from scores. (Notably, Chapelle et al. [2008a, b, 2010] adopted this approach in their work to assess the validity of the Test of English as a Foreign Language [TOEFL].)

Kane (2002, 2013a, b) describes two stages to the argument-based approach: a) developing an interpretive argument[1] (IA), which justifies the uses and interpretations of test scores; and b) developing a validity argument (VA), which justifies the theoretical framework underlying the assessment. The IA should have a clear structure; include both competency- and task-based approaches to score use and interpretation;

[1] Most recently, Kane (2013a, b) used the term interpretation/use argument (IUA) to replace IA, although its definition remained unchanged. I use the term IA throughout this volume.

and draw on multiple validity inferences as evidence (Chapelle et al., 2008a). This volume employs a six-stage IA framework originally developed by Chapelle et al. (2008a, b) to validate the IELTS listening test. Figure 1.1 presents this framework, and subsequent chapters explore it in greater detail.

The framework includes five validity inferences, beginning with evaluation and ending in utilization. Validity inferences are conclusions about particular aspects of a test's validity, drawn on the basis of available data. Each inference is based on warrants, general validating statements (Toulmin, 1958/2003) based on underlying assumptions which render them plausible, appropriate, and accurate. These assumptions must in turn be supported by analytical and statistical evidence collectively known as backings (see Aberdein, 2005; Voss, 2005).

The five validity inferences are as follows:

a) The *evaluation inference* takes the test's scoring rubrics and procedures as evidence. It posits that the test's scoring processes have been standardized to a sufficient degree; that test taker behavior is consistently observed; and that scores are consistently awarded.
b) The *generalization inference* uses the conclusion of the evaluation inference as data, and establishes that observed test scores are generalizable to expected scores in the target test domain across parallel test forms and multiple test administrations.
c) The *explanation inference* is the most important constituent of a test's validity argument; it establishes that the test's developers have determined the meaning, key elements, limits, and boundaries of the construct targeted for assessment, and that the test indeed engages the target construct.
d) The *extrapolation inference* uses the conclusion of the explanation inference to link the construct to the target scores. It establishes that test takers are likely to obtain very similar scores on other indicators of ability.
e) The *utilization inference* is based on the explanation inference, and establishes that the observed scores represent the ability level of the test taker relative to the challenges that he or she will encounter in the target context.

Relatedly, the evaluation and explanation inferences often appear to be similar or highly related. Although the evaluation inference simply establishes that the test has been consistently administered, it has some

common elements with extrapolation inference such as the quality of the test's theoretical underpinnings or its degree of construct representation. The extrapolation inference is where test developers explore if the test indeed measures what they set out to measure (for example, construct representation).

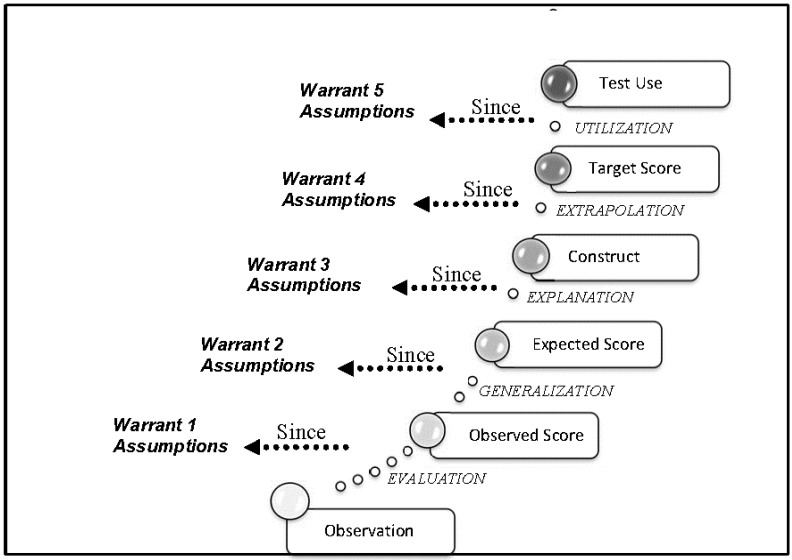

Figure 1.1. The interpretive argument structure used as this volume's conceptual framework (based on Chapelle et al., 2008b, p. 18).

In two recent commentaries, Newton (2013) and Sireci (2013) suggested that Kane's model should be simplified by eliminating the IA or merging the IA with the VA, since IAs are the basic forms of VAs. They argued that if test developers lay out the intended goals of their tests clearly, they would not need to develop separate IAs. Expressing his agreement, Kane (2013b, p. 117) also did not view the articulation of an IA as mandatory; however, Kane further argued that an explicit articulation of proposed uses and interpretations of test scores would help the stake holders preclude from "inconsistencies and ambiguities before they cause too much trouble." Accordingly, test developers should initially specify the interpretations and uses of the test scores (i.e., IA) and then critically evaluate them by using proper tools (i.e., VA).

Opponents of Kane's approach such as Borsboom and Markus (2013) have argued that Kane's model does not concern itself with the *truth*, as it

is based on justified realities. Yet Brennan (2013, p. 74) finds the model practical, pragmatic, and useful, stating that, "Kane's argument-based approach to validation offers by far the best current basis for optimism about improvements in validation".

In the present book, I build a validity argument for a listening test of academic proficiency: a version of the listening test, and the assertions made by the test's International English Language Testing System (IELTS) listening module. The next several sections of this chapter examine the IELTS developers regarding its structure and scope. The chapter reviews the history of IELTS; examines research on the IELTS listening test; elaborates the gaps that are evident in IELTS listening research; poses research questions for the studies conducted later in this book; reviews the concept of validity in educational and language assessment and the definitional issues surrounding it; and finally, delineates the key terms used in the book.

1.2. Development of IELTS

IELTS developed from the English Language Testing Service (ELTS), which was launched in 1980. ELTS was itself a replacement for the English Proficiency Test Battery (EPTB), a lengthy multiple choice test battery used by the British Council since the mid-1960s to evaluate the English proficiency of international applicants to tertiary institutions in the UK (see Davies, 1965).

In 1987, the University of Cambridge Local Examinations Syndicate (UCLES) and the British Council commissioned experts of Edinburgh University to undertake a validation study of the ELTS test (see Alderson & Clapham, 1992; Clapham & Alderson, 1997; Criper & Davies, 1988; Hughes, Porter, Weir, 1988). Criper and Davies (1988) found a substantial "mismatch" between test takers' ELTS scores and their English proficiency levels, likely because of the test's great length and the anxiety it occasioned in test takers. Accordingly, Criper and Davies (1988) suggested that ELTS be restructured. This suggestion, along with corroborating reports from overseas test centers, convinced ELTS's sponsors to launch the Revision Project in 1988 to reshape the micro- and macro-structures of the test.

Eventually, the International Development Program of Australian Universities and Colleges (IDP Education Australia) joined the British Council and UCLES to form the international coalition that developed the new test (Charge & Taylor 1997; Weir, 1993), called the International English Language Testing System (IELTS). The addition of

"international" to the test's title indicated its global scope. Because it had been developed by a multinational coalition, the test was presumed to be devoid of cultural bias; Charge and Taylor (1997, p. 374) wrote that "one important advantage of this the new test was that it helped to prevent any perception of Eurocentric bias and instead ensured a fully international perspective."

The updated IELTS exam has four sections, called "modules": the listening module (approximately 40 minutes), reading (60 minutes), writing (60 minutes), and speaking (10 to15 minutes). In a 1995 revision of the test, these four modules were made thematically independent (Charge & Taylor 1997, p. 375), and standardized to use the same nine-band score scale. Global IELTS candidature has grown rapidly, with over one million people taking the test in 2010.

1.3. Research on the IELTS Listening Test

IELTS listening test may be defined as a while-listening-performance (WLP) tests where test takers are demanded to read test items before the auditory experience; then, they hear the audio text and simultaneously reread test items and supply answers. The specific format of WLP tests (specifically IELTS) would create undesirable washback effects on language education (see Alderson & Hamp-Lyons, 1996) and can be a source of construct-irrelevant (CI) variance. For example, Hayes & Read (2004) found that some English instructors teaching IELTS preparation courses had limited their instructions to the skills demanded by the tests, and their choice of teaching materials were extremely restricted to test-specific textbooks.

The IELTS listening test remains underevaluated relative to the other IELTS test modules. A number of researchers have studied the relationships between IELTS listening test scores and academic performance (Cotton & Conrow, 1998; Ingram & Bayliss, 2007; Kerstjen & Nery, 2000; Merrylees, 2003), but only recently have published studies investigated the constituent structure of the test and its interaction with test takers' cognitive processes and strategies (Badger & Yan, 2009; Coleman & Heap, 1998; Field, 2009). The results of these studies have sometimes called into question the test's validity. For example, Field (2009) found evidence that test- and examinee-related factors—in particular, the simultaneous listen-read-write format of the lecture comprehension section—seriously challenge the test's cognitive validity. These findings underscore the need for further rigorous evaluation of the validity of the IELTS listening test.

Focus on predictive validity. IELTS research has mainly examined the test's *predictive validity*: the relationship between test scores and subsequent academic performance, as reflected in test takers' grade point average (GPA) at the institutions they go on to attend. This validity criterion can be problematic for several reasons. As Sawaki and Nissan (2009, p. 4) note, GPA a) lacks variability, especially in graduate programs, which attenuates the correlations; b) differs with the different standards of dissimilar disciplines and institutions; and c) reflects the "differential values" which lecturers and academicians assign to various kinds of academic achievement. Sawaki and Nissan (2009) add that the "summative" criteria of conventional predictive validity studies—which monitor students' overall achievement in specific courses—obscure the actual language performance of ESL students, since academic performance is a function of much more than language proficiency. In all, this method of predictive validity research seems to confound the value implications of test scores, which may help explain why some IELTS studies have reported weak or even negative correlations between scores on the listening module of the test and subsequent GPA.

1.4. IELTS Test Developers' Assertions

Validation studies typically begin by examining test developers' assertions about a test (Bachman & Palmer, 2010), and then attempt to evaluate these assertions. In the case of the IELTS listening test, this task is substantially complicated by the excessive broadness of IELTS test developers' assertions as to the appropriate uses and interpretations of listening test scores, and the lack of assessment research that would corroborate these claims.

Construct-irrelevant variables. IELTS partners stress their efforts to ensure that the test engages the intended cognitive processes and that the input simulates real-life situations (IELTS website, n.d.a,b). More specifically, to develop each version of the IELTS listening test, IELTS test developers state that they consider five primary factors:

a) the difficulty of complete test versions and the range of difficulty of individual items;
b) the balance of topic and genre;
c) the balance of gender and accent in the listening sections;
d) the balance of item formats (i.e., the relative number of multiple choice and other item types across versions); and
e) the range of listening and reading skills tested.

(IELTS website, n.d.a.)

These assertions must be supported by a thorough evaluation for potential construct-relevant and -irrelevant factors. This evaluation should identify the skills engaged by the test; assess the impact, if any, of test method effects (changes in assessed performance occasioned by changes in test format, such as between multiple choice and open-ended test items) (Bachman & Palmer, 2010); and check for bias by examining the test's functionality across various subpopulations (Kunnan, 1995, 2004; Roussos & Stout, 1996). Although some general research has addressed the effect of language test methods and bias in language assessment (for example, Aryadoust, Goh, & Lee, 2011), few research studies have directly evaluated the IELTS listening test for the skills it taps and the degree of intrusion of construct-irrelevant factors (see Coleman & Heap, 1998; Field, 2009). In practice, the available literature falls short of providing a solid framework that could determine the effect of test methods, test taker demographics, and other construct-irrelevant factors on IELTS listening test performance (see Buck, 2001; Roussos & Stout, 2004).

Target domain. The founders of IELTS state that "IELTS assesses the language ability of people who need to study or work where English is the language used in communication" (IELTS website, n.d.b), and that "IELTS ensures the right level of English to succeed in study, training or life in a new country, so you can count on the long-term satisfaction of your customers" (IELTS website, n.d.c). These statements suggest that success in English educational programs, training, and life in general result from English language proficiency, and that IELTS is the proper instrument to gauge it. These assertions may reflect a primarily promotional (rather than academic) purpose, but they should still be scrutinized closely, as they encourage professional and academic institutions and governments to rely on IELTS scores for making high-stakes decisions regarding test takers' academic and professional lives.

The broadness of IELTS developers' assertions corresponds to the wide range of uses to which IELTS test scores are put. Many universities whose medium of instruction is English have adopted and relied on IELTS as an entry requirement for non-native English speakers, often requiring a composite score of at least six out of nine as a threshold of minimum proficiency. (Applicants assessed to be "linguistically unprepared" are either denied admission, or placed into supplementary English language courses.) IELTS scores are also used to match job-seekers to jobs that require a high degree of English proficiency, and to screen potential immigrants to English-speaking countries such as Canada, the UK, Australia, and New Zealand (McNamara, 2006).

In view of the great diversity of IELTS test score uses, substantiating the assertion that "IELTS ensures the right level of English" across all of these uses would require extensive research. Although the test's developers have made praiseworthy attempts to make test score uses and interpretations transparent, appropriate, and adequate to their users' needs, unfortunately these attempts remain critically narrow; and stakeholders are evidently being encouraged to use IELTS scores for academic and professional purposes for which they have not yet been validated (see McNamara, 2006). As discussed in the previous subsection, the validity of score uses is particularly under-researched for the listening module.

1.5. Listening Comprehension

The IELTS listening module is also difficult to validate because of the lack of a commonly adhered-to model of listening comprehension in the field of language assessment.[2] Despite substantial research into the structure of listening comprehension (Bae & Bachman, 1998; Buck, 1990, 1991, 1992, 1994, 2001; Eom, 2008; Richards, 1983; Shin, 2008) and a number of proposed listening comprehension models (Buck, 2001; Dunkel, Henning, & Chaudron, 1993; Wagner, 2004), no consensus exists in the field as to the definition of the construct. Such a model should clearly specify a related theory, and be able to accurately predict both test item difficulty and test taker performance along a unidimensional latent trait line (Wilson, 2005).

1.6. Research Objectives

This study constructs and evaluates a validity argument for the IELTS listening test. This requires an analysis of the construct-related issues surrounding the test, and an examination of its underpinning structure, as well as the specification of a listening comprehension model to compensate for the lack of a broadly agreed upon model in the existing literature.

This study differs from others in the type of criterion-related evidence of validity it considers. Rather than using an aggregated validity criterion such as GPA, the study concentrates on academic listening ability itself, and examines the relationship between IELTS listening test scores and an academic listening criterion measure developed partly through test takers' self-assessment of their own English listening skills, as well as an

[2] Indeed, this lack adversely influences every listening study (Buck, 2001; Flowerdew, 1994).

academic listening test developed by Educational Testing Service's (ETS) researchers.

The following sections describe the theoretical framework of the book and enumerate its main research questions.

1.7. Theoretical Framework of the Book: Gaps to Bridge

As previously discussed, existing research on the IELTS listening test is constrained in scope; furthermore, the results of this research have tended to challenge the test's claim to validity. This volume addresses five major gaps in research on the IELTS listening test, as well as in the current literature on listening assessment.

First, Field (2009) found that the answer keys provided by the IELTS listening test developers disqualify some answers that fit the context perfectly and answer the questions. This test method issue carries obvious implications for the validity of test takers' scores and the decisions informed by them.

Second, although previous studies have investigated test item format as a source of variance in IELTS scores (Field, 2009; see also Wu, 1998), no study has assessed the full extent of this impact on listening test performance. Multiple choice questions (MCQs) and gap-filling and table completion test items have been employed successfully in a number of test environments (see Wu, 1998, for a discussion), but the high cognitive demands of these item formats, when delivered concurrently with challenging L2 listening comprehension tasks, appear to cause construct-irrelevant variance in listening test scores (see Kavaliauskiene, 2008).

The third gap in knowledge concerns test bias and is a function of IELTS's global administration (Clapham, 1996). The great diversity in candidates' backgrounds requires that the test be examined continuously for bias and differential item functioning (DIF), systematic differences in performance on one or more test items across subpopulations (Ackerman, 1996). This necessity has been recognized and voiced by the developers of other English tests. For example, Geranpayeh and Kunnan (2007, p. 193) argue that the tests developed by the University of Cambridge ESOL Examination Syndicate are likely to display DIF for such variables as nationality, age, gender, as well as previous exposure to the test ("test-wiseness") since "there has been a shift in the traditional test population, where test takers of many age [and nationality and gender] groups are taking these cognitively challenging tests." According to *Standards* (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/

NCME], 1999), without conducting DIF studies, test users and researchers are left to presume that the tests are fair and do not favor any group of test takers. The available literature shows that the IELTS listening test has never been subjected to DIF analysis.

Fourth, the IELTS listening test has not been examined for construct-irrelevant causes of test difficulty. The only study that addresses this concern is Field's (2009) qualitative study, in which test takers reported a number of construct-irrelevant cognitive processes which appeared to jeopardize the cognitive validity of the listening test. However, Field's study only examined the lecture comprehension section of the test, and so was limited in terms of item formats and scope. It is necessary to examine whether Field's findings are replicable in other contexts; whether they are sustained when statistical rather than qualitative approaches are used; and whether other item formats are contaminated by the same or different construct-irrelevant factors.

Finally, the IELTS listening test's predictive validity is an open question. Several studies have shown that the test does not correlate well with candidates' future academic performance; but an equal number report the opposite (see, for example, Ingram & Bayliss, 2007; Kerstjen & Nery, 2000; Merrylees, 2003). It is not yet known which listening sub-skills the IELTS listening test taps, and anecdotal evidence suggests that the test focuses on understanding explicitly stated information (Geranpayeh & Taylor, 2008); if this is true, then it appears to operationalize the listening construct rather narrowly (Shohamy & Inbar, 1991; Wagner, 2004) and to make disproportionate claims concerning the uses and interpretations of scores. Further investigation is needed into the test's correlation with external criterion measures, and particularly with measures which specifically assess academic listening.

1.8. Research Questions and Structure of the Book

In view of the concerns raised and gaps highlighted in Sections 1.3, 1.4, and 1.7, this volume aims to answer five main research questions (Table 1.1). Each question informs an inference of the argument-based validity framework that this book develops for the IELTS listening test. Together, the collected evidence examines the main claim that implicitly underlies IELTS developers' other claims: that the uses and interpretations of IELTS listening test scores are valid in academic contexts.

Table 1.1 summarizes the structure and content of this book. It presents the book's five main research questions; relevant literature comprising the theoretical background for each research question; the corresponding

validity inferences along with their underlying warrants, assumptions, and backings (see Sections 1.4 and 2.1 for further information); and the methods of data analysis implied by the backings. Each research question aims to address one of the major gaps, identified above, in literature on the IELTS listening test.

Following Kane's (1992) validity argument framework, the book uses the findings from investigation into each research question as backings for one or more validity inferences, organizing the findings into a coherent treatment of the validity of the IELTS listening test.

Table 1.1 ***Interpretive Argument of the IELTS Listening Test: Main Research Questions, Relevant Literature Search, Relevant Validity Inference, Warrants, Assumptions, and Methods of Data Analysis***

Research question	Relevant findings from IELTS literature survey	Corresponding validity inference and warrant	Likely claims made by test developers	Likely assumptions of warrant	Data analysis methods to yield backings
1. What listening sub-skills does the IELTS listening test assess? Is there evidence indicating that test items are tainted by construct-irrelevant factors?	The answer keys provided by the test developers disqualify some answers that fit the context perfectly and answer the questions (Field, 2009).	Evaluation inference (Warrant: Test performance is consistently observed and evaluated.)	Evaluation inference: Test developers have identified the important aspects of listening that provide evidence of listening comprehension skills and incorporated them into the IELTS listening test.	a) The test content has been clearly specified; b) Test development and administration processes have been standardized; and c) Rubrics for marking are clear and suitable for furnishing evidence of targeted listening comprehension.	Content and answer key review
	IELTS stresses the importance of ensuring that the test engages intended cognitive processes and that the input simulates real-life situations (website).	Generalization inference (Warrant: Test scores achieved over the test are generalizable to expected scores in the target test domain.)	Generalization inference: Observed test scores are generalizable to and reflect expected scores in the target test domain (i.e., across parallel test forms).	a) There are sufficient items on the test to provide a precise and reliable estimate of candidates' listening ability; b) The test stratifies test takers into several distinguishable person levels.	Analysis of internal consistency and reliability

Research question	Relevant findings from IELTS literature survey	Corresponding validity inference and warrant	Likely claims made by test developers	Likely assumptions of warrant	Data analysis methods to yield backings
	The range of listening skills tested is determined and verified (website).	Explanation inference (Warrant: Observed scores are attributable to the listening latent trait under assessment.)	Explanation inference: The meaning, limits, and boundaries of the listening comprehension latent trait have been accurately determined.	Explanation: a) Listening skills and processes required to understand the text and answer the items are consistent with listening theory; b) The IELTS listening test is not contaminated with test method effects (sources of construct-irrelevant variance).	Confirmatory factor analysis (CFA); Rasch fit statistics; principal component analysis of Rasch residuals(PCAR)
2. Does the test method (in particular, the item format) affect test performance?	Test developers have balanced the item format (i.e., the relative number of multiple choice and other item types) across test versions (website).	Explanation inference (Warrant: same as above)	Explanation inference (same as above)	Same as above	Multitrait-multimethod matrix (MTMM); PCAR

Research question	Relevant findings from IELTS literature survey	Corresponding validity inference and warrant	Likely claims made by test developers	Likely assumptions of warrant	Data analysis methods to yield backings
3. Is the IELTS listening construct represented similarly across different gender, nationality, and other subgroups?	Diversity in backgrounds of candidates would indicate that the test needs to be examined continuously for bias and differential item functioning (DIF) (Geranpayeh & Kunnan, 2007; Geranpayeh & Taylor, 2008)	Explanation inference (Warrant: same as above)	Explanation inference (same as above)	Same as above	Rasch-based differential item functioning (DIF)
4. What are the construct-relevant and -irrelevant factors that determine test item difficulty in the IELTS listening test?	The difficulty of complete test versions and the range of difficulty of individual items are evaluated (website).	Explanation inference	Explanation inference (same as above)	Same as above	Path modeling

1.9. Significance of the Study

The results of this study will contribute to the theory and practice of second language (L2) listening comprehension, language assessment, and argument-based validation. It should help to map some of the ways in which interpersonal variation in test takers' features (such as their background, level of education, and demographics) may affect their listening performance or interact with listening test items and tasks, and where in a listening theory these sorts of individual differences can fit. The study should also help uncover the constituent structure of the elements of listening comprehension that influence listening test performance. Finally, the book should highlight the well- and poorly-researched aspects of validity in the IELTS listening test; and its major objective is to construct and evaluate the test's validity argument.

The study also makes a methodological contribution: while Kane (1999, 2002) proposed a general model for constructing a validity argument, there has as yet been no systematic attempt to identify the relevant forms of data analysis for evaluating each of the model's inferences (Schilling, 2004). I propose a data analysis framework for Kane's model and evaluate the framework's efficacy in the context of the IELTS listening test.

1.10. Overview of the Book

This book consists of seven chapters, including this introductory chapter. Chapter Two introduces an expanded evidence-based validity argument framework, and reviews literature relevant to the theory of (academic) listening comprehension and assessment. Chapter Three describes both the IELTS listening test study and a study assessing the validity of the academic listening self-assessment questionnaire (SALQ). Chapter Four presents the results of both studies. Chapter Five summarizes the findings of the two studies, and develops a validity argument for the IELTS listening test. Chapter Six presents conclusions, highlighting assessment and educational implications. Finally, Chapter Seven reiterates the research questions; highlights the implications of the research conducted here for language assessment; and enumerates several lines of research that can inform the field of language assessment in the future.

The book seeks to lay out and examine the claims articulated about the IELTS listening test by test developers and answer the following research question: