1 *Introduction to Validity Argument in Language Testing and Assessment*

Carol A. Chapelle and Erik Voss

Introduction

Professionals in applied linguistics and language teaching agree that language tests should be valid if they are to be used for making decisions about test takers and developing knowledge about language learning. Tests are used in education and business for important decisions such as what grade to award a student for a course, whether or not to certify a candidate's language proficiency at a certain level for general employment eligibility, and whether or not applicants' language performance is adequate to meet criteria for particular work, education, and immigration requirements. Tests are also used by teachers and researchers for assessing achievement of course outcomes, investigating effects of language instruction, and developing theories of language learning. Because of the important uses of tests, the central concern of researchers in language testing is how to investigate whether or not tests are appropriate for their intended purposes. Such validation research encompasses a range of quantitative and qualitative methods, but exactly how a validation research program is designed, carried out, and interpreted varies based on factors such as the test, its purpose, the developer, the researcher, and the intended audience for the results.

In language testing, a variety of validation research methods have been presented (e.g., Bachman, 1990; Bachman & Palmer, 1996; Weir, 2005), and more recent treatments of validation in scholarly books introduce argument-based validity (e.g., Bachman & Palmer, 2010; Chapelle, 2021; Fulcher, 2015; Kunnan, 2018). Argument-based validity is an analytic approach for conceptualizing, conducting, and interpreting validation research for all types of educational and psychological tests and assessments, which was not developed specifically for language testing. Its principal architect, Michael Kane (1992, 2001, 2006, 2013), introduced argument-based validity with examples

of tests for purposes such as placement and certification in a variety of academic and professional areas. Researchers in language testing and assessment have recognized the utility of argument-based validity, which has been used for presenting validation research for the Test of English as a Foreign Language iBT (Chapelle, Enright, & Jamieson, 2008) as well as for planning, appraising, and reporting research on other language tests, as described in the following two chapters.

Despite the discernible trend toward presenting and using argument-based validity in language testing, this approach has proven challenging for some language testers to adopt. One reason is that argument-based validity demands testers to state the basis for test score interpretations at a greater level of detail than they may be accustomed to doing. The detailed specification requires the use of terms and concepts that are commensurate with the level of technicality of the concepts of validation. For example, testers and researchers who have come to think of "reliability" as referring to a single concept or "authenticity" as having a vague referent find the new terms to be a challenge. The guidance on argument-based validity in the field in the past consisted primarily of chapter-length introductions, which did not provide sufficient examples of validation research for readers wanting to understand how to develop their own validity arguments. There are now more lengthy presentations of argument-based validity and numerous published examples of argument-based validity used as a research framework. However, in language testing, both introductions and examples of research vary in their use of the terminology and concepts for developing validity arguments.

This volume was curated to add clarity to research in language testing and assessment by providing an introduction to argument-based validation along with examples illustrating the use of the framework, terms, and concepts to investigate arguments for language tests and assessments. Throughout the volume, the terms *test* and *assessment* have essentially the same meaning because both refer to systematic procedures for gathering data from test takers, from which interpretations are made to assign scores that are used for making decisions. Both the terms test and assessment carry additional meaning in various contexts, but the basic functional meaning of both, as stated above, indicates that both require validation.

Argument-based validation is an approach for creating a research program to investigate the validity of test score interpretation and use. Such programs of research are required by professional standards of testing such as the *Standards for Educational and Psychological Testing*, which were developed and are periodically revised by associations directly concerned with the quality of tests: the American

Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014). The *Standards*, whose influence extends well beyond these organizations and the American context, indicates that testers need a framework for score interpretation and use to guide the validation process. The *International Language Testing Association Guidelines for Practice* also includes the guidance that the test developer should publish a test manual with "evidence of the reliability and validity of the test for the purpose for which it was designed."

Such standards and guidelines state the need for validation, but they are not intended to explain the specifics of how to plan and carry out validation research. Argument-based validity has proven to be an effective framework to fill this role for some language testers. In the interest of making this approach to language testing research more accessible to all language testers, applied linguists, and educators in other fields, this volume is intended to expand the diaspora of professionals conversant in argument-based validity. It does so by first presenting the basic terms and concepts in argument-based validity and its use in language testing. The details of research investigating validity are then described by chapters that focus on various aspects of a validity argument for English language tests. A final chapter summarizes the methods that serve in validity arguments and identifies challenges in the use of argument-based validity.

Evolving Validation Needs in Language Testing

Argument-based validity is being adopted in language testing because of its capacity to address current needs. The papers published in the major journals in language testing, *Language Testing* and *Language Assessment Quarterly*, reveal the range of issues, methods, and practices that fall within the purview of language testing today. Readers are likely to be struck by the many research approaches used in validation studies, which include an expanding set of statistical methods and research designs in addition to methods in natural language processing, conversation analysis, corpus linguistics, discourse analysis, language-focused ethnography, and critical inquiry into policy issues. These approaches to analysis stemming from a variety of research traditions and epistemologies need to be put to service in validation.

Three important changes in language assessment have prompted the expanded scope of inquiries that appear in the research. First, the roles for language assessment use have increased in number and diversity. The broadened roles of assessments in language learning, education,

workplace, and government demand context-sensitive research approaches for investigating the validity of each test for its purpose. For example, an assessment that effectively serves in providing feedback on learning to teachers and students in a particular language course should not be held to exactly the same criteria as one that is used for designation of an overall proficiency level to be used in decision-making for job readiness. Assessments used in research on language learning need to be evaluated taking into account the specific research objectives. The use of assessments for such a range of purposes demands research that investigates how fit the test is for its intended purpose. Argument-based validity accommodates the differences among language test purposes with a framework that requires the tester to identify the claims that are made on the basis of test scores. In other words, the teacher would want to be able to claim that the classroom assessment tests what was taught whereas a publisher selling a job readiness test would need to be able to claim that the test would be useful in making such employment decisions. In both cases validity is the concern, but validity means something different in each case.

Second, testing demands in many contexts have called for advances in test methods requiring the use of technology for test task development, delivery, and scoring, as well as for integration with online learning materials and learning management systems, data gathering and synthesis, and score and profile reporting. New technologies create the imperative to reexamine the language constructs measured by language tests because test tasks presented and scored by computer have implications for the meaning of the test scores. For example, tests requiring students to compose text or manipulate objects on the screen require some different abilities than their counterparts presented on paper. Scores based on tasks requiring test takers to speak to an interactive computer program can be different than those requiring speaking in a monologic or face-to-face format. Performance data captured in digital form or even gathered online during test taking create novel opportunities for studying test taking performance. As a result, the types of data that can be gathered and analyzed in validation research, once limited to test takers' responses on paper or voices stored in analogue on audio or video, require better methods of analysis and interpretation.

Third, the number and variety of participants involved in language testing has grown in pace with not only new technologies but also changing demographics, mobility, economic systems, and political trends. In response to the mobility of individuals across multilingual contexts, an expanding group of test users are the managers of language policies set for institutions and organizations both nationally

and internationally. In many countries, a culture of accountability in schools has placed testing at center stage, creating requirements for tests to perform functions such as identifying students in need of English support and certifying their readiness to exit such programs. Where language requirements exist, so does the need for language tests. Where a need for tests exists, companies that develop tests are there, as well. As a result, a broader number and range of people are involved in test development and marketing new products for a growing number of purposes. The expanding market for language tests has not yet prompted a response from the profession of setting up a formal licensing or accreditation process. However, in view of the high-stakes decisions that are made on the basis of test scores, the situation demands an improved consensus among professionals about what validation should consist of, how it should be reported, and how its adequacy should be appraised.

In this context of energized language assessment development and use, argument-based validity has attracted the attention of many language testers as a framework for planning, organizing, and interpreting validity evidence.

Argument-Based Validity

The *Standards for Educational and Psychological Testing* define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test scores" (AERA, APA, & NCME, 2014, p. 1). This definition emphasizes that validation is carried out in view of the specific test uses and that test uses are validated by showing support for interpretations of test scores. This definition reflects the definition of validity widely recognized in language testing: "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1989, p. 13). In this definition, "actions based on test scores" refers to test score uses. To put this definition into practice in language testing, Bachman and Palmer have explicitly emphasized test use in their evaluation frameworks, "test usefulness" (Bachman & Palmer, 1996) and "assessment use argument" (Bachman & Palmer, 2010).

The fact that language tests are evaluated in view of their particular uses hints at the complexity inherent in validation. A test used for certification of English proficiency in a professional domain, such as health care, would not be the same as an end-of-semester test for a general-purpose intensive English course. A test used for admissions to an English medium university would be different than either of these.

In these and many other language testing scenarios, test developers, researches, and users are concerned with the validity of the test for a particular use. Carrying out validation research therefore requires a means of first stating the intended test interpretations and uses and then conducting research required to assess the degree of support for making such statements.

Argument-based validity provides the tester the concepts of *claims* and *inferences* for beginning to tackle the task. Claims are the statements about, for example, the utility of the test for making certain decisions about test takers and the positive consequences that will ensue as a result of test use. An inference is the logical step that would allow one to conclude that such a claim is justified. For example, from an argument-based validity perspective, a tester would make a claim such as that a particular speaking test is useful for placement into levels in an intensive English program. Such a claim would follow as a logical conclusion from an inference about the suitability of the test for that purpose, which, in turn, would require certain types of support, which also need to be stated. In argument-based validity acceptance of claims about test use requires prior justification for claims about test interpretation. Such claims about interpretations include those about the quality of the test development methods; the accuracy, appropriateness, and security of test delivery and scoring; the findings about relevant aspects and levels of reliability; and evidence about measurement of the intended construct.

Parts II and III contain chapters that focus on investigating *interpretations* and *uses*, respectively. In each of these chapters, the author has sketched the overall validity argument for the test interpretation and use, but in order to show the detail of the research conducted to investigate one part of the validity argument, the chapters in Part II detail research investigating the claims and inferences underlying interpretations; chapters in Part III describe research investigating inferences underlying test use. To provide the necessary background for understanding the validity argument framework that has been tailored to each chapter, Part I of the book introduces the basic tools used to conceptualize, conduct, and interpret research results in an argument-based validity framework.

Chapter 2 introduces the basic concepts and uses of validity argument in language testing and assessment. Based on an analysis of recent documents in language testing that explain or use argument-based validity, Carol Chapelle and Hye-won Lee describe how language testers have operationalized the definition of validity presented by Messick (1989). The analysis demonstrates how argument-based validity takes into account the concepts that are important in language

testing and serves the multiple functions that language testers demand of their validation tools. It distinguishes between the two formulations of argument-based validity that appear in language testing to introduce the conventions used throughout the papers in the volume.

Chapter 3 examines the claims and inferences of the validity arguments that have appeared from 2006 to 2016 in a systematically gathered sample of forty-five journal articles and twenty-five doctoral dissertations. Ahmet Dursun and Zhi Li map the findings chronologically to reveal trends in using the argument-based approach. Based on the results, they make suggestions about constructing interpretation and use arguments as well as evaluating the coherence and plausibility of validity arguments in various testing contexts. These two chapters provide technical and contextual background for the following two sections focusing on test interpretation and test use, respectively.

Test Score Interpretation

Part II contains chapters reporting studies that use argument-based validity to investigate *score interpretation* for six different English language tests. Score interpretation is multifaceted because of the multiple inferences that are made by test users when they interpret a test score. Each time test users rely on a score for making a decision, they implicitly infer that the test design was appropriate and the development was rigorously done so as to result in test tasks that elicit relevant performance from test takers. Test users also infer that the test takers' performance on the test was evaluated accurately and summarized appropriately to produce test scores that are relevant to the test use. A third inference is that test scores are consistent across test tasks, forms of the test, testing occasions, and raters. In other words, test users treat scores as if they are reliable. A fourth inference of interpretation is that the test score should be taken to indicate the level of ability of the attribute stated by the test developer, which is typically assumed to be reflected in the name of the test. A fifth type of interpretation is made when test users infer that test scores should be understood to reflect future performance in a specific context, which may also appear in the name of the test.

All of these inferences are typically made by test users, at least implicitly, when test scores are interpreted, and therefore the tester's responsibility is to demonstrate that such inferences are warranted. Doing so entails expertise to formulate the precise language associated with making such inferences, to carry out the research and development work required to justify the inferences, to interpret the findings from the research and development work judiciously, and to state the case for

and against making each inference underlying the interpretation. Each inference itself entails a complex research project and therefore each chapter in this section focuses on only one or two aspects of the interpretation. In other words, each chapter reports on only one component of a larger test development and validation project, which the author refers readers to for more information. In addition, even if the larger project did not investigate all aspects of interpretation, each chapter begins by presenting all of the intended interpretations and uses of the test. Specifying the complete argument is the recommended practice in argument-based validation because the complete argument is needed to understand the scope of the research and evaluate the suitability of the inferences that are investigated. In this way, the validation research in each chapter is situated within a complete argument, which is essential for understanding the role of the research results reported in the chapter in view of all of the intended interpretations of the test scores. The chapters are organized logically beginning with the foundational inference about test design and development.

In Chapter 4, Moonyoung Park explains how the test development research he undertook was guided by the need to provide support for the inference referred to as *domain definition*, which means that the test design is inferred to be appropriate and that the development was rigorously done so as to result in test tasks that elicit relevant performance from test takers. The test was intended for making decisions about placement and providing diagnosis of language needs for air traffic controllers who need to use aviation English for their work in military aviation in an Asian country. Park illustrates how evidence-centered design (ECD) was used to carry out the research required to provide a strong foundation for test development (Mislevy, Steinberg, & Almond, 2003). The research included eliciting experts' views of language needs, analysis of documents used in aviation English courses currently in place, a needs analysis survey, and documentation of a systematic process of task modeling and design. Park makes observations about how argument-based validity aided in the test development by providing specific goals for construct definition and task design through the research required for the domain definition inference.

A second foundational aspect of test score interpretation is taken up in Chapter 5, where Hyejin Yang reports her investigation of raters' use of a new web-based rating platform that allows them to enter their ratings and diagnostic evaluations of test takers' performance on the Oral English Certification Test (OECT) at a Midwestern University in the USA. The web-based Rater-Platform (R-PLAT) changed the rating process from past practices in which raters

reported their ratings by writing on paper. The change required investigation because test score users infer that test takers' performance on the test is evaluated accurately and summarized appropriately to produce test scores that are relevant to the test use. Yang explains the critical role of the rating conditions in the interpretation of scores for deciding whether prospective international teaching assistants' (ITAs) spoken English is adequate to allow them to teach in their respective content areas at the university. The research reported in this chapter used interviews and a survey to investigate the perceptions of both experienced and new raters toward their use of the new online rating system. These findings are reported along with those from other research that supports the inference that test performance is evaluated accurately and summarized appropriately. As Yang explains, this inference is called *evaluation* in the interpretation/use argument, which demonstrates the foundational role of inferences about rating processes in the overall validity of test interpretations. The chapter illustrates research needed when integrating technology into this aspect of a speaking assessment.

Another inference that is often implicitly made by test users is that test scores are consistent across test tasks, forms of a test, testing occasions, and raters. These aspects of consistency can collectively be considered to be an absence of error in the test scores. Test score use implicitly presumes an absence of error, that is, that the score represents scores that the test taker would obtain on different tasks, test forms, occasions, and from different raters. This presumption is called a *generalization* inference. Because of the multiple potential sources of error (e.g., test tasks, raters), this aspect of test score interpretation can be investigated in a variety of ways. Among the most vexing issues in language assessment is the error present in students' extended spoken or written responses. Tasks eliciting such samples of language performance are both essential for many test uses and prone to all types of error. In Chapter 6, YunDeok Choi reports on her investigation of the inference about score consistency for interpretation of scores on a computer-mediated graphic-prompt writing test intended to measure test takers' source-based academic writing ability in English. Choi developed and piloted the test. Then, using Generalizability (G) Theory and Multi-Faceted Rasch Measurement (MFRM), she investigated the extent to which the test scores were dependable, the numbers of tasks and raters were sufficient to achieve the desired level of reliability for placement purposes, and the interrater reliability was sufficient. Results from these analyses are placed within the interpretation/use argument, in which conclusions are drawn about the potential for the graphic-prompt writing test as a reliable measure of the construct, and about the use

of the assumptions underlying the generalization inference to frame multiple aspects of reliability.

Also investigating reliability, Chapter 7 presents research framed from the perspective of an external evaluator of a speaking test, for which users are concerned about the error that may be introduced by the different forms of the test students take on different occasions. Acting as an evaluator with a particular interest, Rie Koizumi formulated a *rebuttal* for the *generalization* inference in the validation framework. The study examined the Telephone Standard Speaking Test (TSST), a telephone-based test of second language (L2) English-speaking proficiency whose scores were used to assess improvement in speaking proficiency over time. The validity of the test use required investigation of the consistency of scores across forms and occasions, which had not been provided by the publisher. Analysis of TSST scores from undergraduate students at two Japanese universities using a paired *t*-test, Levene's test for equality of variances, and a correlation did not support the rebuttal, that is, very little error was found in the test scores. The chapter shows how an external evaluator's perspective can pose rebuttals in the argument-based framework to reflect the standpoint of outsiders.

In Chapter 8, Erik Voss reports the investigation of an inference about the construct assessed by a computer-delivered test of collocational ability. The research and development project encompassed the inferences discussed in the previous chapters in addition to the *explanation* inference, which is the part of the interpretation that test users make when they accept the score as an indicator of the level of ability for the attribute stated by the test developer, in this case collocational ability. Because language ability constructs are theoretical entities and not observable, Voss had to specify the construct theory precisely enough to make predictions about expected empirically observable outcomes. He developed the construct definition based on applied linguistics theory and research. He then investigated construct-based predictions about the difficulty of each of the test items, relationships between collocational ability and other constructs (reading and vocabulary), and strategies used by test takers for completing the test. Data were gathered as item responses, test scores, screen capture of test taking, interviews, and a survey. The chapter provides insights into how a construct definition was developed to serve as a basis for the explanation inference, how scoring methods intersected with the construct definition, and how argument-based validity was used to state testable hypotheses about the construct.

In Chapter 9, Jooyoung Lee reports on her investigation of the interpretation of classroom assessments as one part of a larger project