Chapter 2

VALIDITY ARGUMENT DESIGN

The *Standards* presents validation as an evidence-based process of developing and evaluating arguments about the interpretation and use of test scores. It states, "Decisions about what types of evidence are important for the validation argument in each instance can be clarified by developing a set of propositions or claims that support the proposed interpretation for the particular purpose of testing" (AERA et al., 2014, p. 12). In view of the multifaceted interpretations and uses that need to be taken into account in validation, a comprehensive, systematic approach to generating propositions is needed. In Kane's (1992) terms, validation efforts need to focus "attention on the details of the interpretation" (p. 527). This chapter introduces how the details of interpretation and use are taken into account in argument-based validity, as presented by Kane (e.g., 1992, 2001, 2006, 2013). It begins by introducing three tests that will serve as examples throughout the book. These tests provide concrete examples of how claims and inferences are used to express the meanings that make up their respective score interpretations and uses, how the claims serve as connectors in multipart arguments, as well as how argument-based validity provides the tools for specifying the evidence required to develop and evaluate a validity argument.

Expressing Interpretations and Uses: Three Example Tests

Three example tests were selected to represent three different contexts and test uses. Each was developed within the academic traditions of validation research outlined in Chapter 1, even though only one, the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT), has had a Kane-style validity argument developed to support its interpretation and use. First published in 2005, the TOEFL iBT is a test of academic English proficiency intended primarily to aid in university admissions decisions for applicants whose first language is not English. The TOEFL iBT is the most recent version in a tradition begun in the 1960s of developing and administering a test for use in admissions decisions at North American universities. Published by Educational Testing Service, an established research and development organization in the United States, the TOEFL iBT has been the subject of years of validation research. Although research continues to be published in journal articles and research reports by Educational Testing Service, in

the mid-2000s results were summarized in a book that presents the validity argument for the intended interpretations and uses of the TOEFL iBT (Chapelle, Enright, & Jamieson, 2008).

The second example is the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT), a test of emotional intelligence that was developed by psychologists and published by MHS Assessments for use in a range of settings in which intellectual capacities for perceiving emotions and reasoning about them are of interest to score users. Research on emotional intelligence dates back decades, leaving a trail of research articles in major academic journals in psychology, some of which are about the MSCEIT and its development. The published research provides the large majority of the publicly available information about the test (Mayer, Caruso, & Salovey, 2016; Mayer, Salovey, & Caruso, 2008).

The third example is the mathematics section of the Iowa Assessments (University of Iowa, 2015), which is intended to assess the mathematics skills of students taught in the public school curriculum from kindergarten through Grade 12 in the United States. The Iowa Assessments are a product of decades of research and development at the University of Iowa. The information presented about the Iowa Assessments comes from the *Research and Development Guide: Iowa Assessments, Forms E and F* (University of Iowa, 2015), which contains references to journal articles reporting additional research. Intended for scores users, the *Research and Development Guide* explains the development, basis for validity claims, and ongoing analysis of test results.

As tests developed by professionals, all three produce scores that have been shown to be reliable, and therefore various types of reliability claims have appeared in the reports about the tests. However, reliability claims do not encompass all of the meanings entailed in the score interpretations and uses. For each of the three tests, Table 2.1 shows an example of another claim about each of the test scores, the meaning it conveys about the test scores, and the inference required to attribute the meaning to the test scores. Argument-based validity provides for all claims made about test scores to express the detail of the test interpretation and use.

One of the claims about the TOEFL iBT scores is that they are relevant to the quality of linguistic performance in English-medium universities. The test developers make this claim because they have designed the test to assess academic language proficiency rather than general language proficiency or language proficiency in another domain such as tourism, mechanics, or business. This claim is important for the TOEFL iBT users because they recognize that the language demands in higher education require certain types of performance: Students need to use English to learn new concepts, critically analyze what they read and hear, and express their

Table 2.1 Three Example Tests, Claims About Their Score Meaning, Types of Meanings, and Inferences Required to Attribute Meaning

Test	*Claim About Test Scores*	*Meaning Attributed to Test Scores*	*Inference Required to Attribute Meaning*
Academic English: TOEFL iBT	TOEFL iBT scores are relevant to the quality of linguistic performance in English-medium universities.	Real-world relevance	Extrapolation
Emotional intelligence: MSCEIT	Scores reflect the ability to recognize and reason about emotions.	Substantive sense	Explanation
Mathematics achievement test: Iowa Assessments	Iowa Assessments are useful for educational purposes requiring descriptive data on an individual student or groups of students.	Functional role	Utilization

knowledge and analyses. English-language tests are not all equally suited to assess academic language, and therefore they would not be relevant to performance at a university. By interpreting the test scores as relevant to language performance in English-medium academic contexts, test users make an extrapolation inference. This means that they are using the score on the test to extrapolate, or extend beyond, the known test score to the unknown judgments about test takers' language performance in an English-medium university.

One of the claims about the MSCEIT is that its scores reflect test takers' ability to recognize and reason about emotions. This is a claim about the construct, or substantive sense, of the scores because it gives the score meaning with respect to the knowledge, skills, and abilities that the test is intended to assess. The construct label *emotional intelligence* is a shorthand descriptive name for the substantive sense, but that label alone is not sufficiently precise to express the substantive meaning of the test score interpretation. A number of tests are referred to as tests of emotional intelligence,

but they do not all actually assess the cognitive aspects of recognition and reasoning, so making a specific claim about the intended score meaning in addition to the label of emotional intelligence is important. When test users interpret the test scores as having the substantive sense of recognizing and reasoning about emotions, they make an explanation inference. In other words, they are accepting that the ability to recognize and reason about emotions explains the test scores, and the explanation comes from a psychological definition of the capacity (Mislevy, 2006).

One of the claims about the Iowa Assessments mathematics achievement test scores is that they are useful for "a variety of important educational purposes that involve the collection and use of information describing either an individual student or groups of students" (University of Iowa, 2015, p. 3). The claim about the intended role of the scores gives them a functional meaning by indicating what they should be used for. A test of mathematics would be designed differently if results were intended to be used for college admissions or certified public accountant (CPA) licensure, for example. The *Research and Development Guide* (University of Iowa, 2015) explains the functional role in more detail by stating types of school-based decisions that the scores are intended to support. For a mathematics test designed for college admissions or licensure, statements about the functional role would refer to decision making for institutions and for society rather than decision making for students. The inference required for putting the test scores to use for a particular purpose is utilization.

Using Claims and Inferences to Express Interpretations and Uses

The claims shown for each of the example tests illustrate what Kane (1992) meant by "the details of the interpretation" (p. 527). Claims express the types of meanings intended when test scores are interpreted. The term *claim* refers to a statement that is made about the test scores, including various aspects of their qualities, meanings, and intended impacts. The term *claim* is used instead of *fact* because a claim is a statement that is open to dispute and, therefore, typically requires evidence supporting its credibility. For each of the example claims in Table 2.1, the research conducted on the respective test has offered some support for the claim. Because claims attribute meaning to test scores on the basis of evidence, they act as conclusions drawn by making inferences. "Inference" in argument-based validity refers to the process of drawing conclusions about score meaning.

To use argument-based validity, therefore, a tester needs to be able to render intended score meanings (i.e., interpretations and uses) as claims that serve as conclusions for certain inferences. Table 2.2 summarizes four

Table 2.2 Four Meanings Attributed to Test Scores, General Claims, and Inferences Leading to Their Respective Claims

Meaning Attributed to Test Scores	*General Claim*	*Inference Leading to the Claim: Definition*
Real-world relevance	Test scores are based on performance on test tasks relevant to the context of interest.	Extrapolation: The score user accepts that the score meaning extends to the context of interest.
Substantive sense	The test scores reflect the intended construct.	Explanation: The score user surmises that the score meaning is explained by the defined construct.
Functional role	The test scores are useful for their stated purpose.	Utilization: The score user trusts that the scores should be used for the stated purpose.
Degree of stability	The test produces reliable scores.	Generalization: The score user concludes that the test produces reliable scores.

aspects of test score meaning along with claims that are stated in general terms and, for each claim, the type of inference that would be made if the claim were accepted.

The example claim for the TOEFL iBT academic English test in Table 2.1 illustrated one way of expressing a claim about the real-world relevance of the test score. Generally speaking, the claim is that test scores are based on test performance relevant to the context of interest. Such a claim attributes meaning to the scores in terms of the congruity of the test tasks with tasks that people do in the real world and, in particular, in the context of interest to score users. Such a claim gives the test scores a vivid meaning to many score users who can see the connection between what the test taker was required to do on the test and what they have to do in the real world. Kane, Crooks, and Cohen (1999) emphasized that accepting such a claim requires the score user to be able to extrapolate from the test score to performance in a particular context of interest, and the inference is therefore referred to as extrapolation.

The example claim about the substantive sense of the MSCEIT is that scores reflect the test takers' ability to recognize and reason about

emotions. The general claim is that the test scores reflect the intended construct, which is typically expressed as the knowledge, skills, and abilities required for performance. As Messick (1989) put it, constructs are meaningful interpretations of performance consistency. The ability to troubleshoot computer failures, proficiency in speaking French, and knowledge of multiplication tables are examples of constructs. Constructs are not observed directly. It must be surmised that the score meaning is explained by the defined construct. The inference is therefore called explanation.

The example claim about the functional role of the mathematics subtest of the Iowa Assessments is that the scores are useful for educational purposes requiring information that describes individuals or groups of students. Generally speaking, the claim is that the scores are useful for their stated purpose. Purposes can include the range of functions that tests are created to serve such as certification, placement, and diagnosis. As Cureton (1951) emphasized, the purpose also includes the test takers for whom the stated uses are intended. When score users trust that the scores should be used for their stated purpose, they are making a utilization inference.

Claims about reliability attribute the scores with a meaning about their degree of stability, or consistency. *Consistency* can refer to the stability of scores across different forms and occasions of testing. It can indicate consistency across tasks on the test, meaning that the score reflects multiple samples of performance that are justifiably combined into one score. Consistency can also refer to consistent judgments of multiple raters across their ratings or occasions of rating. In other words, reliability encompasses multiple different types of consistencies, each of which is estimated in a different way.

The following chapters examine these claims and inferences in more detail and introduce some additional ones. But for this chapter, these four types of claims provide a basis for understanding the tools required for structuring claims into arguments and identifying the evidence required to support them.

Structuring Claims in a Validity Argument: From Grounds to Conclusions

For any test, more than a single claim is made to express the score interpretation and use, so validation is never a single study for investigating one claim. Instead, multiple claims with their inferences are structured together into what Cronbach (1988) called a "validity argument": Based on the idea that validation is evaluation, Cronbach suggested that "what House (1977) has called 'the logic of evaluation argument' applies," and he invited testers

to "think of 'validity argument' rather than validation research" (Cronbach, 1988, p. 4).

Cronbach saw a validity argument as having a political dimension because it provides a means of integrating multiple meanings of test scores for diverse audiences. Kane (1992) developed argument-based validity from a more technical standpoint, as a practical argument supported by incomplete or even questionable evidence. Practical arguments are never proven; they are "at best, convincing or plausible" (p. 527). Kane structured claims and inferences into an argument by drawing upon Toulmin's (2003) argument structure that begins with a premise, or grounds, and ends with a conclusion. An inference makes the connection, or link, from the grounds to the conclusion. Applying this argument structure to testing, Figure 2.1 illustrates how the claim about the substantive meaning of the scores for the MSCEIT serves as a conclusion for the explanation inference. The premise is the test scores, and an explanation inference leads to the conclusion that the scores reflect the test takers' ability to recognize and reason about emotions. In this argument, the claim serves as the conclusion.

To develop Figure 2.1 into a more complete argument, the basic three-part structure needs to be expanded to accommodate additional claims and inferences. Figure 2.2 illustrates how this is done by adding a claim about reliability and a generalization inference. Figure 2.2 again shows the test scores as the premise, or grounds. The first inference, generalization, leads to the conclusion that the test produces reliable scores. This conclusion also serves as the premise for the explanation inference, which leads to the conclusion that the scores reflect the test takers' ability to recognize and reason about emotions. This illustrates how a validity argument structure strands together premise–inference–conclusion sequences in which the conclusion from one valid inference serves as the premise for the next. This example

Figure 2.1 Structure of an Argument About Test Scores for the MSCEIT Serving as a Premise for an Explanation Inference that Concludes the Scores Reflect the Test Takers' Ability to Recognize and Reason About Emotions

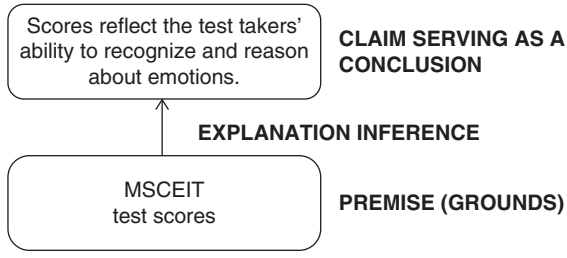
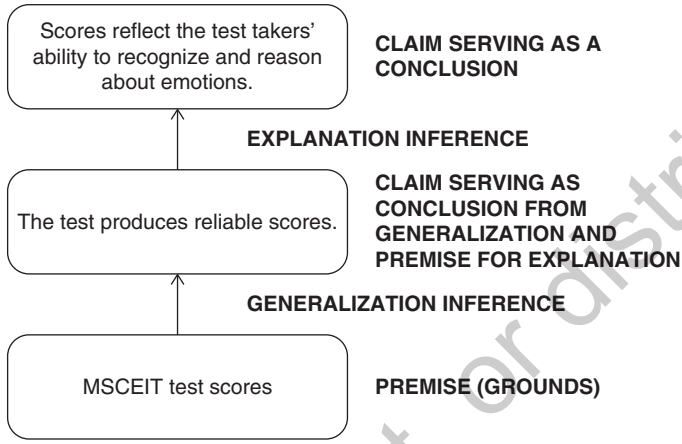


Figure 2.2 Structure of an Argument About Test Scores for the MSCEIT With a Premise (Grounds) and Inferences Leading to Two Logically Related Claims About Reliability and Constructs

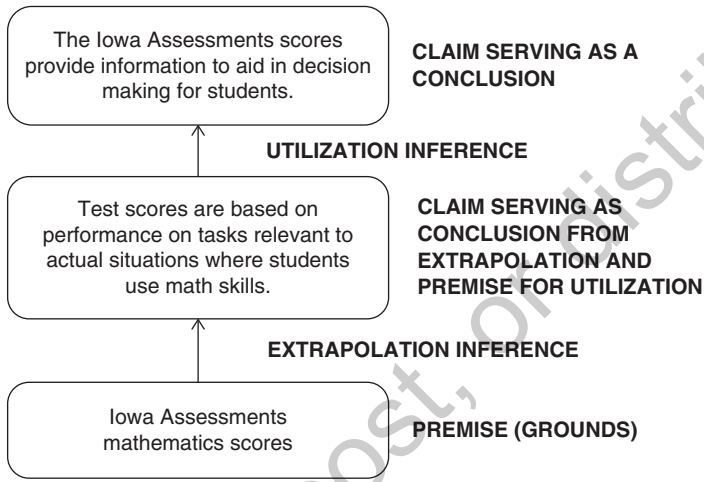


also shows a notation for expressing Cronbach's and Messick's view that a construct interpretation can be made only if scores are shown to be reliable: The reliable scores are the grounds for the explanation inference.

A second example of a chain of claims appears in Figure 2.3, which illustrates the logic behind the claim about the usefulness of the scores on the Iowa Assessments mathematics section. The test scores are the premise or grounds. The first inference leads to the conclusion that the test scores are based on performance on tasks relevant to actual situations in which students use math skills. This conclusion also serves as the premise for the utilization inference, and the utilization inference leads to the final conclusion, the claim that the Iowa Assessments mathematics scores provide information to aid in decision making for students.

Validity arguments typically have more claims and inferences than those shown in Figures 2.2 and 2.3, but these examples should suffice to demonstrate the validity argument figures in this book, which uses the metaphor of grounds to place the premise at the bottom of the argument diagrams. The figures in this book are consistent with many of the publications about validity argument, but one can also find ample examples of argument diagrams that place the premises to the left of their respective inferences and conclusions. The left-to-right reading of such diagrams has the same meaning as the corresponding bottom-to-top reading of the diagrams in this book.

Figure 2.3 Structure of an Argument About Test Scores for the Iowa Assessments Mathematics Achievement Test, Including the Test Scores as Premise (Grounds) and Inferences Leading to Two Logically Related Claims About Relevance and Test Use



Each of the four claims appearing in Figures 2.2 and 2.3 is about the specific score interpretation for one of the example tests. To summarize the structuring of claims and inferences in more general terms, Table 2.3 presents the premises, inferences, and general claims for the generic versions of these argument structures and the meaning that each inference adds to the scores.

When the scores serve as the premise or grounds for the first inference, they are typically stated as a phrase, for example, "Iowa Assessments mathematics scores." The grounds for the following inferences, however, are conclusions from the previous inferences (e.g., "Test scores are based on performance on tasks relevant to actual situations in which students use math skills."). Conclusions are often restated as claims, which are full sentences; although, when considered as grounds, such conclusions can be restated as phrases, too. For example, the conclusion "Test scores are based on performance on tasks relevant to actual situations in which students use math skills" can be restated as grounds with the expression "relevant math skills." The recognition of the statement/phrase form of claims in a validity argument is useful for reading academic articles on validity argument because they often use shorthand expressions in validity argument diagrams to represent conclusions and premises. Instead of the

relatively transparent expressions such as "reliable scores" and "relevant scores," they also use expressions such as "expected scores" and "target scores." The use of these technical terms in place of the complete sentences used to express claims is a useful shorthand device, but only if they are understood.

The four core meanings of test scores can be expressed with the claims and inferences exemplified in Table 2.2, but other inferences entailed in test score interpretation express variations of the core meanings, as well. These will be introduced in the following chapters to bring the total to seven inferences. But argument-based validity is not conceived as a fixed list of inferences. Instead, the intent is to provide the conceptual tools and language to help test developers and researchers to identify inferences that are important in their specific test interpretations and uses. A validity argument for a particular test "should reflect the proposed interpretation and use; it should not be constrained to fit some predefined structure" (Kane, 2013, p. 10).

Table 2.3 General Claims With Their Grounds and Inferences

Figure	*Meaning*	*Premise (Grounds)*	*Inference*	*General Claim*
Figure 2.2	Degree of stability	Test scores	Generalization	The test produces reliable scores.
	Substantive sense	The test produces reliable scores.	Explanation	The test scores reflect the intended construct.
Figure 2.3	Real-world relevance	Test scores	Extrapolation	Test scores are based on performance on test tasks relevant to the context of interest.
	Functional role	Test scores are based on performance on test tasks relevant to the context of interest.	Utilization	The scores are useful for their intended purpose.

Nevertheless, the four inferences introduced so far demonstrate how some of the basic constituents are combined to sketch the structure for an argument.

Identifying Evidence: Warrants, Assumptions, and Backing

The example arguments outlined above would be the beginning of what Kane (2013) calls an "interpretation/use argument." To develop interpretation/use arguments into validity arguments, evidence is needed to support each of the inferences leading to its respective claim. In order to identify the types of evidence that would serve in support, more detail is needed. In validity arguments, the detail is expressed in warrants and assumptions.

Kane (2013) defined a warrant as a statement that expresses a "rule for inferring claims of a certain kind from data of a certain kind" (p. 12). Warrants are thought of as allowing, authorizing, or licensing inferences. They do so by first adding precision to the meaning of inferences. For example, an explanation inference allows for an interpretation about the substantive construct meaning of the test score, but what does that imply for the validation research needed for the emotional intelligence test? Construct validation can entail a full range of research methodologies, as suggested by the five sources of evidence identified in the *Standards.* Warrants need to serve in formulating specific research goals whose results may support the inferences in the validity argument.

Table 2.4 provides examples of the types of warrants that could be used to license the inferences leading to the claims about scores on the tests. The extrapolation inference leading to the claim that the academic English TOEFL iBT scores are relevant to the quality of linguistic performance in English-medium universities has a warrant that adds to the inference: "The construct of academic language proficiency as assessed by the TOEFL iBT accounts for the quality of linguistic performance in English-medium institutions of higher education" (Chapelle et al., 2008, p. 348). Support for this warrant will require research that investigates the relationship between the TOEFL scores and other criterion scores that are indicators of aspects of academic language proficiency and linguistic performance in higher education.

Which criterion measures should be accepted as relevant and what kind of relationships should be expected require still another level of detail, which should be built into the validity argument by adding assumptions underlying each of the warrants. Assumptions underlying this warrant in the TOEFL validity argument name specific types of criterion measures, including test takers' self-assessments, professor's judgments, and scores

Table 2.4 Example Claims, Inferences, and Warrants in Validity Arguments

Claim	*Inference Leading to Claim*	*Example Warrant Licensing the Inference*
TOEFL iBT scores are relevant to the quality of linguistic performance in English-medium universities.	Extrapolation	The construct of academic language proficiency as assessed by the TOEFL iBT accounts for the quality of linguistic performance in English-medium institutions of higher education (Chapelle et al., 2008, p. 348).
MSCEIT scores reflect the ability to recognize and reason about emotions.	Explanation	Test scores support the theorized four-component model of emotional intelligence that includes managing emotions to attain specific goals, understanding emotions and emotional language and signals, using emotions to facilitate thinking, and perceiving emotions accurately in oneself and in others (Mayer et al., 2016).
The Iowa Assessments scores provide information to aid in decision making for students.	Utilization	The scores can "identify strengths and weaknesses in student performance—make relative comparisons of student performance from one content area to another" (University of Iowa, 2015, p. 3).

on another academic English test. Assumptions are examined for the extrapolation inference in Chapter 4, and for all seven inferences in the following chapters, but suffice it to say in this chapter that well-written warrants and assumptions can pinpoint the research required to support a particular inference by identifying the empirically testable hypotheses.

The explanation inference about the construct of the MSCEIT scores needs a warrant that specifies in greater detail the meaning of the construct. The warrant would be that test scores support the theorized four-component model of emotional intelligence that includes managing emotions to attain specific goals, understanding emotions and emotional language and

signals, using emotions to facilitate thinking, and perceiving emotions accurately in oneself and in others (Mayer et al., 2016). Chapter 4 shows how such a warrant about the construct is used to develop more specific assumptions about, for example, the hypothesized degree of relationship among the four components of the model and their place in a larger nomothetic network of subconstructs of intelligence. These assumptions, in turn, point to the types of research required to support the warrant. If research results are indeed supportive, the results serve as backing for the warrant, which authorizes the inference leading to its conclusion that the test assesses emotional intelligence.

The third example shown in Table 2.4 is the utilization inference leading to the claim that the Iowa Assessments mathematics scores provide information to aid in decision making for students. The warrants authorizing such an inference would state the intended uses of the test scores and the rules for specific score-based decisions. Assumptions would identify the findings needed to make the warrants credible.

Warrants such as those illustrated in Table 2.4 add detail to the meaning of inferences in validity arguments, but additional detail is needed to specify research questions. Assumptions provide detail that suggests types of evidence that need to be found through research. In the validity argument, such evidence is "backing" for assumptions because certain pieces of evidence serve as support for making particular assumptions. Assumptions and backing obviously have to get deeply into the detail of validation for specific tests and are, therefore, illustrated further in the following chapters.

Identifying Weaknesses and Limitations in Arguments: Rebuttals

The claims, inferences, warrants, and assumptions illustrated earlier are all used to state the intended interpretations and uses of test scores. These are the primary concern for test developers wanting to present a validity argument supporting their test's interpretation and use. But they do not fully serve Cronbach's (1971) view of validation as scientific hypothesis testing, which requires a means of expressing threats to the intended interpretations, or "rival hypotheses that may challenge the proposed interpretation" (AERA et al., 2014, p. 12). In some cases, threats to intended interpretations need to be identified and investigated for certain individuals or groups. In the *Standards* (Chapter 3), such threats to validity are treated as concerns about test fairness. In validity arguments, threats are expressed at the level of individual inferences, to isolate the specific source of hypothesized unfairness within the

complex chain of inferences entailed in the intended interpretation and use. Hypothesized threats to validity for all test takers, or for certain individuals and groups, are expressed in argument-based validity with rebuttals.

A rebuttal in a validity argument states the conditions under which a particular warrant would not be able to license its respective inference, as illustrated in Table 2.5. For example, a rebuttal added to the claim in

Table 2.5 General Claims With Examples of Corresponding Warrants and Potential Rebuttals

General Claim (Inference)	*Warrant*	*Potential Rebuttals*
Test scores are based on performance on test tasks relevant to the context of interest. (Extrapolation)	The test tasks elicit test takers' performance that reflects their performance in situations of interest to test score users.	• The important characteristics of tasks of interest to test users were not adequately analyzed. • The test takers have experience different from that of the group used to norm the test and are unfamiliar with the task content.
The test produces reliable scores. (Generalization)	A sufficient number of test tasks are given to test takers to produce reliable scores.	• Test administration is not carried out as specified in some locations. • The internal consistency reliability of the scores is different across different subgroups of the population.
The test measures the intended construct. (Explanation)	Test scores support the theorized internal structure of the construct the test is intended to measure.	• Some individuals are advantaged due to coaching on test content and format. • The test-taking processes are ineffective for some individuals whose first language is not English.
The scores are useful for their intended purpose. (Utilization)	The scores are appropriate for making decisions about mastery of the content covered in class.	• Test users do not obtain test results in a timely fashion. • Test content disproportionately favors students who have attended the same school for several years.

the second row about reliability would read, "It can be concluded that scores are reliable because support was found for the assumptions underlying the warrant that a sufficient number of test tasks are given to test takers to produce reliable scores, unless it is also found that test administration is not carried out as specified in some locations." Rebuttals invite research to investigate the extent to which evidence supports them.

The examples in Table 2.5 show that rebuttals can be used to express what may go wrong in certain situations to weaken a validity argument. They can include cultural aspects of the setting that are different from those assumed by the test developer, test takers whose background is different from what is needed, a school situation in which the results cannot be acted upon as intended by the tester, and any number of other situation-specific and person-specific factors that potentially make test interpretation and use invalid. Rebuttals provide a heuristic for specifying testing practices likely to disproportionately affect a certain group of test takers or an individual with particular characteristics. In this way, rebuttals provide a means of including some of the fairness issues of interpretations and uses of test scores for decision making for certain individuals and groups, including those defined by demographics such as gender, race, and cultural background (Camilli, 2006; Xi, 2012).

If supported, rebuttals undermine the inferential process that the validity argument builds with claims, inferences, warrants, and assumptions. They are, therefore, useful tools for critics conducting evaluations of validity arguments developed by others, prospective test users wanting to evaluate test use for a different context, and test developers needing to identify areas requiring attention during test development. Disadvantages can typically be identified for certain groups of individuals, such as those with hearing or sight impairment, limited proficiency in the language of the test, or lack of experience with use of technology for testing. In these cases, accommodations need to be created to provide access for those individuals (*Standards,* Chapter 3). Even though test developers tend to focus on seeking support for claims and inferences, credible claims require the absence of support for rebuttals as well. Test developers, therefore, can use rebuttals to be proactive by identifying potential limits to the inferences and taking action by stating the limits on test use for certain groups and providing accommodations for other groups.

The Language of Validity Argument

Working with validity arguments requires testers to learn some new terms and ways of framing interpretation and use. Even though the claims,

inferences, warrants, and assumptions express the same basic inferential processes that have been used for decades in testing research, at first glance, these terms seem puzzling to many testers. Many testers working on validation use the language of "types of evidence" from Messick's (1989) presentation of the faceted unitary validity, which is reflected in the *Standards.* However, neither Messick nor the *Standards* develop the specific language and logic for crafting the claims and specifying their roles in the interpretation and use of a particular test. The *Standards* refers to score interpretation and use, the associated propositions or claims to be supported, and the five types of evidence, but in practice, these three pieces are difficult for test developers and researchers to generate and stitch together. The result is reports of validity research with vague or unstated interpretations, incomplete or absent propositions, and research presented without an explanation of its contribution to a validity argument. What is missing in a "types of evidence" approach to validation is a systematic way of expressing the validity argument that Cronbach (1988) invited testers to think of.

Expressing the argument requires some additional terms beyond propositions, claims, and evidence. Table 2.6 shows the correspondence between the terms used in the *Standards* and those a validity argument framework provides to test developers and researchers. The terms are arranged in descending levels of generality from top to bottom, with the three levels of analysis in the *Standards* on the left, the seven levels used in validity arguments in the middle, and the definitions of the terms on the right. The argument-based approach prompts the test developer to elaborate the test's interpretation and use by analyzing the intended score meanings, specify the meanings with claims, identify the inference that leads to each claim, and use the claims and inferences to structure an argument. The tester then needs to provide additional detail using warrants that authorize the inferences and assumptions that provide still more specificity about the research to be conducted to make the warrants credible. Results from research motivated by specific assumptions in the validity argument can be interpreted with respect to the corresponding inference. These terms provide the detail required to express all aspects of score interpretation and use in a manner that motivates particular validation research and provides a context for its interpretation. The terms, therefore, allow testers not only to think of validity argument but also to express validity arguments to make clear the role of validity evidence.

Table 2.6 Terms Used for Expressing Validity Arguments, Arranged by Their Levels of Generality

	Standards	*Validity Argument*	*Definitions*
General ↑	Interpretation and Use	Interpretation and Use	Overall statement of test purpose
		Score Meanings	General expressions denoting aspects of meaning
	Propositions (Claims)	Claims	General statements about interpretation and use
		Inferences	General technical terms denoting the steps in reasoning
		Warrants*	Statements indicating an inference can be authorized in a particular context
		Assumptions	Statements clarifying what evidence is needed
↓ Specific	Evidence	Backing	Statements, paragraphs, tables, figures in extended descriptions of findings

*Note: Rebuttals are the statements corresponding to warrants that indicate conditions under which an inference cannot be authorized in a particular context.

Conclusion

The validity argument framework presented in this chapter provides a means for testers to state the details of test interpretation and use by analyzing intended score meanings. Four basic meanings of test scores can be expressed as claims, which serve as conclusions for particular inferences. Four terms—warrant, assumption, backing, and rebuttal—were introduced to show how the support for inferences is conceived and challenged in ways that point to specific validation research. In the following chapters, the four components of interpretation and use will be expanded into a more nuanced palette of meanings with additional claims, warrants, and assumptions. The next chapter shows that the functional role of a test can be expressed in

terms of claims about test uses (e.g., achievement, prediction, diagnosis) and about specific decisions to be taken based on certain scores. The functional role can also be expressed as claims about consequences of test use on test takers, academic fields of study, or society. In this way, each chapter helps to expand the vocabulary of testing professionals for developing their own arguments about the validity of test interpretation and use.