Building and Supporting a Case for Test Use

Lyle F. Bachman

University of California, Los Angeles

The fields of language testing and educational and psychological measurement have not, as yet, developed a set of principles and procedures for linking test scores and score-based inferences to test use and the consequences of test use. Although Messick (1989) discusses test use and consequences, his framework provides virtually no guidance on how to go about investigating these in the course of practical test development. Argument-based formulations of validity (e.g., Kane, 1992, 2000; Kane, Crooks, & Cohen, 1999; Mislevy, in press; Mislevy, Steinberg, & Almond, 2003) provide a logic and set of procedures for investigating and supporting claims about score-based inferences but do not address issues of test use and the consequences of test use. Recent formulations in language testing (e.g., Bachman & Palmer, 1996; Kunnan, 2003; Lynch, 2001) are essentially lists of more or less independent qualities and questions, with no clear mechanism for integrating these into a set of procedures for test developers and users to follow. What has been called “critical language testing” (e.g., Shohamy, 1999, 2001) has alerted us to the political uses and abuses of language tests and to the need for test developers and test users alike to be self-critical of the ways in which tests are used. However, this perspective treats consequences as essentially unrelated to the validity of inferences and provides little guidance about how to go about either anticipating and avoiding, or redressing, the problems with test use that it discusses.

In this article I describe how an argument for test use might be structured so as to provide a clear linkage from test performance to interpretations and from interpretations to uses. An assessment use argument is an overall logical framework for linking assessment performance to use (decisions). This assessment use argument includes two parts: an assessment utilization argument, linking an interpretation to a decision, and an assessment validity argument, which links assessment performance to an interpretation. I then discuss ways in which issues and questions that have been raised by language testers regarding uses, abuses, consequences, validity, and fairness in
language testing can provide a basis for articulating claims and counterclaims in an assessment use argument. In my view, an assessment use argument can guide the design and development of assessments and can also lead to a focused, efficient program for collecting the most critical evidence in support of the interpretations and uses for which the assessment is intended.

PRELUDE

For the better part of the past 25 years, I have been investigating the areas of language ability that seem to be engaged when test takers interact with language assessment tasks, and the nature of these tasks themselves.¹ I have also been concerned with the ways in which we can provide these individuals, as well as other stakeholders, with reasonable assurances that the inferences and decisions we make on the basis of their test scores are useful. Quite early on, I identified what I believed to be two sets of determinants of test performance that I felt needed to be accounted for in any test use. The first determinants were the various attributes that the test takers themselves bring to the assessment. These attributes include, of course, the attribute in which most language testers are interested—language ability. The attribute that we intend to test is what we have all come to know and love as the construct. The other determinants include the various aspects of the assessment tasks and the situation in which the assessment takes place. This set of determinants constitutes what many would refer to as the context of assessment. Investigating the relative effects of these two sets of determinants on test takers’ performance and on the scores they receive in language tests, which is essentially the process of validation, has consumed much of my own research and thinking for quite some time.

However, despite my work on validation, I have never lost sight of the fact that it is the ways in which we use tests that is at the heart of language assessment. In 1990, I wrote,

The single most important consideration in both the development of language tests and the interpretation of their results is the purpose or purposes which the particular tests are intended to serve. (Bachman, 1990, p. 55)

I also pointed out that

tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or society at large. (Bachman, 1990, p. 279)

¹The logic of an assessment use argument applies equally well to all types of assessments, whether these are measures, tests, or other types of assessments. Thus, for the purposes of this article, I use “assessment” and “test” more or less interchangeably. (See Bachman, 2004 for a detailed discussion of the terms “assessment,” “measurement,” and “test.”)
Finally, I urged language testers to consider both validity and test use, in the following charge:

As test developers and test users, ... it is our responsibility to provide as complete evidence as possible that the tests that are used are valid indicators of the abilities of interest and that these abilities are appropriate to the intended use, and then to insist that this evidence be used in the determination of test use. (Bachman, 1990, p. 285)

In the past decade or so there has been an increasing interest in and debate among language testers about issues of fair test use, consequences, and ethics, and this debate has invigorated our field. Until now, I have remained largely at the side of this debate for two reasons. First, I felt that we still had much to learn about linking test performance to interpretations. Second, I did not see a way to make what I felt was the critical link between score-based interpretations on the one hand, and the uses that we make of these on the other. However, I have recently found what I believe to be a way forward in argument-based approaches to validation in educational measurement. It is thus the link between validity and test use that I explore in the discussion to follow.

What I share with you in this article is very much a work in progress, some initial thoughts on how we might go about linking interpretations to uses. The following quote from Toulmin eloquently expresses the spirit in which I share my thoughts with you:

Nothing in what follows pretends to be final, and I shall have fulfilled my purpose if my results are found suggestive. If they are also found provoking, so much the better; in that case there is some hope that, out of the ensuing clash of opinions, the proper solutions of the problems here raised will become apparent. (Toulmin, 2003, p. 1)

It is thus my hope that these ideas will not only spur debate in the field, but will also lead to applications in the practice of our profession. I believe that it is ultimately in the forge of practical test design, development and use that these ideas will be either refined and extended, or rejected.

INTRODUCTION

Although the fields of language testing and educational and psychological measurement have been discussing test use and the consequences of test use for well over a quarter of a century, neither field has, as yet, developed a comprehensive set of principles and procedures for linking test scores and score-based interpretations to test use and the consequences of test use. The current “standard” validity theory, that is, Messick (1989), discusses test use and consequences, to be sure. However, this theory provides virtually no guidance on how to go about investigating these conse-
quences in the course of practical test development. More recent argument-based approaches to validation (e.g., Kane, 1992, 2001, 2002; Kane, Crooks, & Cohen, 1999; Mislevy, 2003; Mislevy, Steinberg, & Almond, 2002, 2003) provide a logic and set of procedures for articulating claims and for collecting evidence to support these claims. However, these argument-based approaches have focused primarily on claims about the interpretation of test scores, and have not, until very recently, begun to address issues of test use and the consequences of test use.

Historically, within the field of language testing, issues of test use, consequences, and ethics have, with few exceptions, generally been (a) ignored entirely, as in pre-Messick discussions of validity (e.g., Lado, 1961; Pilliner, 1968; Spolsky, 1968, 1973); (b) embedded within a unitary (Messickian) conceptualization of validity (e.g., Bachman, 1990; McNamara, 2001); or (c) discussed as issues that may build on, but that are essentially unrelated to, validity (e.g., Lynch, 1997; McNamara, 1998; Norton & Starfield, 1997; Parry, 1994; Shohamy, 2001a, 2001b; Spolsky, 1997). Formulations that separate out consequences or use from validity, such as Bachman and Palmer’s (1996) notion of test usefulness, Lynch’s (2001) discussion of ethics and validity, and Kunnan’s (2003) formulation of test fairness are essentially lists of more or less independent qualities, or aspects of usefulness, validity, or fairness, respectively. These discussions, though clearly enlarging our view of the social and educational milieu in which language testing is situated, do not, in my view, bring us any closer to a coherent theory of test use.

The net result of these discussions of test use and consequences has, in my view, been the same: We have neither a coherent theory of test use that links this to validity, nor a set of procedures for investigating test use and consequences for any particular assessment. That is, we do not, at present, have a set of manageable procedures for investigating test use and consequences that is grounded on a coherent theoretical model of test use. Rather, we have simply several different checklists of questions and considerations with which test developers and test users should concern themselves.

In this article, I provide a brief overview of discussions of test use and consequences in the language testing literature. I then discuss argument-based approaches to validation, focusing on Toulmin’s argument structure. Next, I discuss how this argument structure can be applied to formulating a validity argument, illustrating this with an example from language testing. I then describe how this argument structure can be extended to articulate an assessment utilization argument, also illustrating this with an example from language testing. I then briefly discuss how some of the qualities that language testers have proposed to address issues of test use and fairness can be included in an assessment use argument. I conclude by arguing that an assessment use argument, which comprises the validity and utilization arguments, can provide the logical framework both for guiding test design and development and for developing a substantively grounded set of procedures for justifying the use of a particular assessment.
DISCUSSIONS OF THE CONSEQUENCES OF LANGUAGE ASSESSMENT USE

Though there have been extensive theoretical discussions and empirical investigations of validity in the language assessment literature over the past 20 years, discussions of and research into the consequences of test use have been sparse indeed, with the majority coming within the past decade (see Kunnan, 2003, for an overview of discussions of ethics and fairness in language testing). In one of the earliest discussions of the consequences of test use in the language testing literature, Spolsky (1981) posed two questions that are still, in my view, at the heart of test use: “How sure are you of your decision?” and “How sure are you of the evidence that you’re using to make that decision?” (Spolsky, 1981, p. 19). The first question recognizes that the fundamental use of language tests is to make decisions. Though Spolsky was focusing primarily on the use of language tests for selecting individuals, it is quite clear that the use of tests to make decisions is fundamental and underlies virtually all uses of language tests. Spolsky’s second question pertains to the strength or credibility of the evidence on which we make these decisions. Spolsky’s concerns for the ethics of test use have been picked up and elaborated on in more recent discussions of ethical test use.2 The assessment use argument that I describe in this article provides, I believe, a general approach to answering these two questions by framing them in a slightly different way. The first question can be rephrased as, “How convincing is the argument for using the assessment in this way?” and the second can be rephrased as, “How credible is the evidence that supports this argument?” Building a convincing case that the decisions we make are defensible and supporting that case with credible evidence are the two components of the validation process.

A major consequence of test use that has been dealt with extensively in the language testing literature is that of the impact of tests on instruction, or “washback.”3 As a result of this extensive research, we now have a much better appreciation for the complexity of washback: the multitude of factors, in addition to the test itself, that may influence the extent to which, if at all, washback occurs; the different ways in which different individuals react to or ignore attempts to engineer “positive” washback with language tests; and the different levels—from government, to institution, to classroom, to individual teachers and students—at which washback may or may not occur. However rich these findings, washback has generally been discussed as an issue that is either separate from validity, or assumes that the score-based interpretations that may have an impact on instruction are themselves valid.

2A review of this literature is beyond the scope of this article. Interested readers are referred to the following: Davies (1997), Hamp-Lyons (1997a), and Kunnan (2000).

3A review of this literature is beyond the scope of this article. Interested readers are referred to the following: Alderson and Wall (1996), Wall (1997), and Cheng, Watanabe and Curtis (2004).
In a perspective on the consequences of test use that has come to be known as “critical language testing,” a term coined by Shohamy (1997), who is its chief exponent, language testers have been alerted to the political uses and potential abuses of language tests. Shohamy (2001b), has very eloquently articulated the need for test developers and test users alike to be self-critical of the ways in which tests are used. Nevertheless, this perspective also tends to treat consequences as unrelated to the validity of inferences. However, as I will point out below, many of the concerns articulated by critical language testing can be integrated into an assessment use argument.

**FORMULATIONS OF VALIDITY, IMPACT, ETHICS, AND FAIRNESS IN LANGUAGE TESTING**

Though the aforementioned discussions have tended to treat validity and test use as essentially unrelated issues in assessment, a number of language testers have addressed both validity and consequences or ethical considerations in language testing. Hamp-Lyons (1996, 1997b), for example, has consistently considered both validity and consequences in her discussions of ethics in language testing, while Elder (1997, 2000) has discussed fairness issues raised by the analysis of test bias, which has been traditionally considered a validity issue. In addition, there have been several attempts to bring the considerations of validity and test use together. Drawing on Messick (1989) and Bachman (1990), Bachman and Palmer (1996) defined an overall quality, “usefulness,” as comprising six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. In their formulation, both the construct validity of our score-based interpretations, and the impact, or consequences, of test use need to be considered from the very beginning of test design, with the test developer and test users working together to prioritize the relative importance of these qualities. Bachman and Palmer provide lists of illustrative questions about the potential construct validity and impact of the interpretations and uses derived from the test. The answers to these questions provide a framework for quality control during the test development phase, and for investigating overall usefulness throughout the life cycle of the operational test. This formulation clearly brings considerations of validity and impact under a unitary concept of test usefulness. However, what it does not do is indicate how, if at all, these qualities are related to each other, or how construct validity and test use are directly related. Specifically, it does not answer the questions of whether valid score-based interpretations can be misused or lead to undesirable consequences, or whether invalid score-based interpretations can nevertheless provide a basis for appropriate test use and desirable consequences.

Another attempt to link validity and consequences is that of Lynch (2001), who draws on postmodern critical theory to consider validity and ethics. Lynch de-
scribes what he calls a “validity framework,” which he claims “integrates validity with ethical considerations” (Lynch, 2001, p. 364). Lynch’s validity framework includes five categories: fairness, ontological authenticity, cross-referential authenticity, impact–consequential validity, and evolved power relations. Without discussing these qualities in detail here, it is clear that, with the exception of impact–consequential validity, these are quite different from those proposed by Bachman and Palmer. Lynch’s categories are essentially values drawn from the literature in critical theory and the qualitative research tradition, and push us even further than do Bachman and Palmer from the comfort zone of the traditional “psychometric” qualities of reliability and construct validity. Lynch’s categories thus clearly take us further down to the road toward consideration of test use. Nevertheless, this framework suffers, in my view, from the same limitation as Bachman and Palmer’s: It is a list of potentially important, but essentially unrelated qualities or categories, with no explicit linkages among them or between validity and test use.

Kunnan (2003) has drawn on the research in ethics in educational measurement and language testing, as well as in secular ethics, to link validity and consequences in what he calls a “test fairness” framework. This framework includes five qualities: validity, absence of bias, access, administration, and social consequences. Although this framework includes two qualities—validity and social consequences—that are included in Bachman and Palmer’s quality of usefulness, it also introduces three new qualities—absence of bias, access, and administration. As with Lynch’s validity framework, Kunnan’s fairness framework expands the list of qualities we need to consider, in the direction of more inclusiveness. But again, I would argue, this is another list of important but essentially unrelated qualities, with no logical mechanism for integrating these. Thus, we are again left with validity at one end and consequences at the other, with no link between.

In summary, discussions of validity and test use in the language testing literature have generally failed to provide an explicit link between these two essential considerations. The extensive research on validity and validation has tended to ignore test use, on the one hand, while discussions of test use and consequences have tended to ignore validity, on the other. To their credit, those researchers who have attempted to link validity and test use have enlarged our perspective beyond Messick’s unitary validity model. In articulating the test qualities that they believe to be important, these researchers have opened up the lines of debate and expanded the dialogue about what should be the overarching concern in language assessment—the way language assessments get used and the consequences of these uses. Nevertheless, what we have at present, as a basis for justifying test use, are essentially lists of qualities and questions that test developers and test users need to consider, with no clear logical mechanism for integrating these into a set of procedures for test developers and users to follow. In the remainder of this article, I describe what I believe to be a means for providing this logical linkage between validity and test use.
ARGUMENT-BASED APPROACHES TO VALIDATION

Recent approaches to validation in the field of educational measurement have elaborated on the notion of a validity argument that has been part of validation research for a quarter of a century (e.g., Cronbach, 1980, 1988; House, 1980). In a seminal article, Kane (1992) developed the notion of an interpretive argument as providing framework for the gathering and disseminating of evidence supporting intended score interpretations. Drawing on the literature in practical argumentation, Kane describes what he calls an “interpretive argument” as consisting of inferences and assumptions, which need to be supported by relevant evidence. Kane points out that, unlike purely logical or mathematical arguments, the assumptions in an interpretive argument cannot be taken as givens, and the evidence in support of these assumptions is often incomplete or debatable. Thus, the conclusions of interpretive arguments are not proven, but can only be evaluated in terms of how convincing or plausible they are. He discusses three criteria for evaluating the inferences made on the basis of an interpretive argument: (a) clarity of argumentation, (b) coherence of argument, and (c) plausibility of assumptions. He then goes on to provide the broad outlines of the kinds of inferences, assumptions, and evidence that would typically be found in an inferential argument.

Kane, Crooks, and Cohen (1999) built on Kane’s 1992 discussion, explicating the details of an interpretive argument for linking observations to interpretations. This argument consists of four parts, each linked to the next by an inference. This interpretive argument is illustrated in Figure 1.

The first link, “scoring,” is an inference from an observation of performance to a score, and is based on assumptions about the appropriateness and consistency of the scoring procedures and the conditions under which the performance was obtained. The second link, “generalization,” is from an observed score on a particular measure to a universe score, or the score that might be obtained from performances on multiple tasks similar to those included in the assessment. This link is based on the assumptions of measurement theory (e.g., generalizability theory). The third link, “extrapolation” is from the universe score to what Kane et al. have called a “target score,” which is essentially an interpretation of what a test taker knows or can do, based on the universe score. This link is based on the claims in an interpretive argument and the evidence supporting these claims.
Beginning from a different starting point, that of test design and development, Mislevy et al. (2002, 2003) also draw on the literature in evidentiary reasoning to develop a set of steps and procedures for what they call “evidence-centered design” (ECD). At the heart of ECD is what they refer to as an “evidentiary argument,” which provides the logical chain of reasoning between the claims, or interpretations we want to make, on the one hand, and the kinds of evidence that we need to collect to support these claims. This evidentiary argument integrates construct definitions, characteristics of assessment tasks, and the psychometric models that are needed to deal with complex performance data. It also has practical benefits in terms of guiding the development of performance tasks and scoring rubrics for these, and laying the foundation for collecting evidence for validity and generalization (Mislevy et al., 2002, p. 479).

The Structure of Validity Arguments

Mislevy et al. (2003) have proposed that we base validity arguments on Toulmin’s (2003) argument structure. For Toulmin, an argument consists essentially of making claims on the basis of data and warrants (see Figure 2).

- A *claim* is “a conclusion whose merits we are seeking to establish” (Toulmin, 2003, p. 90). This is essentially the *interpretation* that we want to make, on the basis of the data, about what a test taker knows or can do.
- The *data* consist of “information on which the claim is based” (Toulmin, 2003, p. 90). In the case of assessment, these are the responses of the test taker to assessment tasks—what the test taker says or does.

![Figure 2](http://example.com/figure2.png)  
*FIGURE 2  Toulmin diagram of the structure of arguments (adapted from Mislevy, Steinberg, & Almond, 2003, p. 11. Used by permission.).*
The arrow from the data to the claim represents an inference, which is justified on the basis of a warrant.

- A warrant is a general statement that provides legitimacy of a particular step in the argument (Toulmin, 2003, p. 92). Warrants are thus propositions that we use to justify the inference from data to claim. This warrant is based on backing.
- The backing consists of “other assurances, without which the warrants themselves would possess neither authority nor currency” (Toulmin, 2003, p. 96). The backing for our warrants in language testing generally comes from theory, prior research or experience, or evidence collected specifically as part of the validation process.
- A rebuttal consists of “exceptional conditions which might be capable of defeating or rebutting the warranted conclusion” (Toulmin, 2003, p. 94). Rebuttals constitute alternative explanations, or counterclaims to the intended inference. In a validity argument, these counterclaims correspond to potential sources of invalidity that may result in what Messick (1989) referred to as “construct irrelevant variance” and “construct underrepresentation.”
- The rebuttal data consist of evidence that may support, weaken, or reject the alternative explanation.

Here’s an example that illustrates this argument structure:

According to the U.S. Constitution, all individuals who are born in the U.S. are U.S. citizens. Since Marc was born in the U.S., and unless he has renounced his U.S. citizenship, we can conclude that Marc is a U.S. citizen.

The structure of this argument is illustrated in Figure 3. The components of this argument are as follows:

- Data: Marc was born in the U.S.
- Claim: Marc is a U.S. citizen.
- Warrant: All individuals born in the U.S. are U.S. citizens.
- Backing: According to the U.S. Constitution, anyone born in the U.S. is a U.S. citizen.
- Rebuttal: Marc has renounced his U.S. citizenship.
- Rebuttal data: Marc’s affidavit renouncing his U.S. citizenship.

In this example, the rebuttal is supported by the rebuttal data, so we would reject the claim that Marc is a U.S. citizen, and conclude that Marc is not a U.S. citizen. If we were not able to produce any evidence that Marc had renounced his U.S. citizenship, then this would reject the rebuttal and we could conclude that Marc is a U.S. citizen.
Now let me provide an example of Toulmin argument structure applied to a validity argument for language testing. Here are descriptions of the use for which this assessment is intended, of the assessment task, and of the scoring procedure.

- **Purpose:** An international hotel company is hiring people to take room reservations over the phone. The hotel requires all employees who must work with customers to use English, and many of the applicants are not native speakers of English. Therefore, the company needs a screening test to determine if applicants have sufficient English ability to take room reservations over the phone. In addition to this language test, an assessment of professional knowledge and skills related to the specific job and to the hotel company’s administrative procedures in general is also given to applicants. Applicants’ scores on both assessments are considered in the employment decision.

- **Assessment task:**
  1. Interact in English with an examiner, playing the role of a customer, in a simulated phone conversation;
  2. Elicit information relevant to making a room reservation.
  3. Record this information in a hotel registration form.
- **Scoring:** The amount and accuracy of relevant information that is recorded.
The validity argument for a hypothetical test taker, “HH” in this example, might be as follows:

When interacting orally in English with customers on the phone, employees should to be able to collect and record information relevant to making a room reservation. HH participated in a simulated phone conversation with an examiner who played the role of a customer, and was not able to collect and record information accurately and completely. Therefore, we conclude that HH has a low level of ability to interpret specific topical information related to making hotel reservations.

The components of this validity argument include the following:

- Data–response: HH participated in a simulated phone conversation with an examiner who played the role of a customer. HH was not able to collect and record information accurately and completely.
- Claim–interpretation: HH has a low level of ability to interpret and record specific topical information related to making hotel reservations.
- Warrant: When interacting orally in English with customers on the phone, employees should to be able to collect and accurately record information relevant to making a room reservation.
- Backing: A needs analysis of tasks in the TLU domain—telephone receptionists’ use of English on the job. This needs analysis indicated that a critical task they need to perform is interacting with customers orally in English on the phone and accurately and completely recording information relevant to making a room reservation.

In the logic of a validity argument, the warrant is the major premise, the data provide the minor premise, and the claim, or interpretation, is the conclusion.

From this example, we can see that an additional type of data needs to be included in the structure of the argument—the relevant characteristics of the assessment task. This is because there is an implicit claim that the task HH performed on the assessment corresponds in specific ways to the language use task specified in the warrant. In this example, the relevant task characteristics are as follows:

Data–task Characteristics:

- The characteristics of the oral interaction involved in the task correspond closely to those in the TLU domain.
- The form for recording the relevant information is the one that is actually used in the TLU domain.
- The scoring is based on the amount of accurate information relevant to the task that is included in the test taker’s response.
The structure of the validity argument for this example is illustrated in Figure 4.

**Interpretation:** HH has a low level of ability to interpret and record specific topical information related to making hotel reservations.

**Warrant:** When interacting orally in English with customers on the phone, employees should be able to collect and accurately record information relevant to making a room reservation.

**Backing:** A needs analysis of tasks in the TLU domain—telephone receptionists’ use of English on the job—indicated that a critical task they need to perform is interacting with customers orally in English on the phone and accurately and completely recording information relevant to making a room reservation.

**Data 1:** HH participated in a simulated phone conversation with an examiner who played the role of a customer and was not able to collect and record information accurately and completely.

**Data 2:**
- The characteristics of the oral interaction involved in the task correspond closely to those in the TLU domain;
- The form for recording the relevant information is the one that is actually used in the TLU domain; and
- The scoring is based on the amount of accurate information relevant to the task that is included in the test taker’s response.

Thus far, we have only discussed the structure of the validity argument that supports our claims or intended interpretations. However, in any assessment, there are a number of potential alternative explanations for performance that may constitute counterclaims, or rebuttals, to our intended interpretations. One set of alternative explanations, which consist of variations in performance that are either random or associated with particular aspects of the measurement procedure (e.g., tasks, occasions, raters, forms), have traditionally been dealt with as sources of measurement error, under the rubric of reliability or dependability. Viewing these sources of measurement error as alternative explanations, or rebuttals, in the validity argu-
ment, provides a mechanism for directly including reliability in validity. Variations in either the attributes of test takers or in the characteristics of the assessment tasks can also lead to alternative explanations of performance and hence lessen the validity of our intended interpretations. Some alternative explanations for HH’s performance are illustrated in Figure 5.

In this example, one alternative explanation for HH’s low performance might be that he was not able to hear the examiner due to technical difficulties in the administration of the test. If the examiner reported such an irregularity, this would support this alternative explanation, and weaken our interpretation about HH’s language ability. However, if a thorough investigation revealed no such problem, this would reject this alternative explanation. Another alternative explanation might be that HH was ill on the day of the exam. But if HH is unable to produce evidence, in the form of a doctor’s report, then this would weaken this counterclaim.

To summarize, Toulmin’s model provides a basis for structuring a validity argument for any given language assessment. This structure will include a claim, or interpretation, about what a test taker knows or can do, based on data, which comprise the test taker’s response to the test task and the characteristics of the test task. This interpretation is supported by a warrant, or general statement that provides a rationale, or legitimacy, for the interpretation. This warrant, in turn, is supported by backing, drawn from theory, prior research or experience, or of evidence collected specifically as part of a validation study.

ASSESSMENT USE ARGUMENTS

Articulating an argument for the validity of score-based interpretations and collecting evidence in support of that argument are necessary components for justifying the uses we make of language tests. However, these are not sufficient for justifying the uses we make of these score-based interpretations. First of all, there is no guarantee that even valid score-based interpretations will be relevant, useful, and sufficient for the intended uses, or decisions. According to Messick (1989), two types of evidence are needed to support test use: (a) evidence that the ability measured by the test is relevant to the intended use or decision, and (b) evidence that the test score is actually useful for making this decision. Questions about the sufficiency of a single indicator provide the rationale for the use of multiple indicators for making educational decisions. Any given indicator may be insufficient, in that it will be unreliable to some degree, and may underrepresent the construct (Henderson-Montero, Julian, & Yen, 2003). The insufficiency of any single measure for providing information that is reliable and that adequately covers the richness and breadth of abilities that are typical of content and performance standards, as well as representing the variety of ways in which teachers may interpret these standards is of particular concern in large-scale, high-stakes assessments (Baker,
**Interpretation:** HH has a low level of ability to interpret and record specific topical information related to making hotel reservations.

**Alternative Explanation 1:**
HH was not able to hear the examiner due to problems with the phone connection.

**Alternative Explanation 2:**
HH was not well on the day of the exam, and this affected his performance.

**Rebuttal Data 1:**
The examiner reported technical difficulties with the phone connection.

**Rebuttal Data 2:**
After a thorough investigation, no evidence of technical problems were found with HH’s test.

**Rebuttal Data 3:**
HH is unable to produce a doctor’s report confirming his illness on the day of the exam.

**FIGURE 5** Alternative explanations for the hotel employment language test.
Second, there is no guarantee that these interpretations will not be subverted for other uses, or be used for decisions other than those for which they were intended. Finally, the validity argument by itself provides no basis for either anticipating or investigating potential unintended consequences of the way the score-based interpretations are used.

Bachman (2003) has argued that the argument-based approach to validation of score interpretations should be broadened to include an argument for test use, suggesting that this might entail a two-part “assessment use argument.” One part of the assessment use argument would be an argument with claims, warrants, and backing in support of the intended uses, as well as rebuttals of potential unintended consequences of test use. The second part would comprise essentially the validity argument linking test performance to interpretations (Bachman, 2003, p. 64). In this section I will describe an assessment use argument as the overall argument linking assessment performance to use (decisions). This assessment use argument includes two parts, a utilization argument, linking an interpretation to a decision, and a validity argument, which links assessment performance to an interpretation. I will use the term validation to refer to the process of collecting evidence, or backing, to support the entire assessment use argument.

Messick (1994) has argued that “the primary measurement concern regarding adverse consequences is that any negative impact on individuals or groups … should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance” (Messick, 1994, p. 21). I would take issue with this and argue that it is quite possible for adverse consequences and inappropriate uses of tests to occur that are not a result simply of sources of invalidity. That is, it is possible for the results of assessments to be used inappropriately, even though these assessments are valid indicators of the abilities they are intended to measure. For this reason, I would argue that an articulation of sources of adverse consequences that are beyond invalidity must be incorporated into the argument for the use of any assessment. I would thus consider issues of test use to be clearly within the concerns of measurement.

Kane (2001, 2002, 2004) has begun to address the role of test use in validation by extending the interpretive argument described by (Kane et al., 1999). Kane (2001, 2002) indicates that this interpretive argument should be divided into two parts. One part, which he calls the “descriptive part,” pertains to the links between scores and descriptive statements about individuals, whereas the other part, the “prescriptive part,” pertains to the link between these descriptive statements and the decisions that are made (Kane, 2001, p. 337). He argues that these two parts are needed because there is not necessarily a one-to-one correspondence between interpretations and decisions. That is, as Kane points out, a given test might be useful for more than one decision, whereas multiple measures, each with its own interpretation, might be useful for a single decision. In a subsequent article, Kane (2002) further delineates the interpretive argument in terms of different kinds of interpretations (descriptive and decision-based) and the kinds of assumptions (semantic
and policy, respectively) on which these are based. In a recent article, Kane (2004) describes an argument-based approach to validation as being based on two arguments, an interpretive argument and a validity argument. The interpretive argument “spells out the proposed interpretations and uses in some detail and then evaluates the coherence of this argument and the plausibility of its inferences and assumptions” (Kane, 2004, p. 167). The validity argument “evaluates the interpretive argument [and] will generally involve extended analysis and may require a number of empirical studies” (Kane, 2004, p. 167). In these articles, Kane has thus extended the linkages in the interpretive argument from scores all the way to decisions. This chain of inferences from observation to decision can be illustrated as in Figure 6.4

Chapelle, Enright, and Jamieson (2004) have recently drawn on Kane’s approach to develop a validity argument for the Test of English as a Foreign Language. They organize their argument in terms of the assumptions and evidence associated with six types of inferences: evaluation, generalization, extrapolation, explanation, decision making, and representation. They then illustrate how these different inferences are organized in the stages of design, implementation, and operation in the process of test development. Though these authors begin with a discussion of how one might go about developing a validity argument based on the joint Standards of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), they quickly move to Kane’s argument-based approach. This study thus represents an important move in language testing away from the highly abstract unified model of validity that informs the Standards, to a more transparent and usable argument-based approach to articulating an assessment use argument.

The Structure of Assessment Utilization Arguments

An assessment utilization argument (or “utilization argument,” for short) will have essentially the same Toulmin structure as a validity argument, and will provide explicit links between score-based interpretations and the decisions that are made on the basis of these. A utilization argument can thus be articulated in terms of claims, warrants, backing, and rebuttals. However, these will be different from those articulated in the validity argument.

4I refer to the link or inference from a score interpretation to a decision as “utilization.”
The *claim* in the utilization argument is essentially the decision we intend to make, and the *data* upon which this decision is based consist of the claims, or score-based interpretations from the validity argument. Thus, the claims, or interpretations, from our validity argument become the data upon which we base the claims, or decisions, in our utilization argument. The structure of an assessment utilization argument is presented in Figure 7.

I would suggest that there are four types of warrants for a utilization argument, all of which are essential to the argument. (I propose these as a starting place, rather than as a definitive, exhaustive list. Practical application of and experience with this approach may well reveal additional types warrants.) The first three types of warrants come from the third and fourth cells of Messick’s (1989) progressive matrix: relevance and utility, and social consequences.

Type 1 warrant: *Relevance* (that the score-based interpretation is relevant to the decision to be made).

Relevance has to do with both (a) the extent to which the ability assessed is a requisite part of the competence needed or task to be performed in the target language use (TLU) domain, and (b) the degree of correspondence between the characteristics of the assessment task and those of the TLU task. In the aforementioned hotel example, using a score from an oral assessment might be considered highly relevant for selecting individuals for the job specified because it assesses speaking and the task is very similar to the relevant task in the TLU domain. Using scores from a standardized multiple-choice of general English proficiency, on the other hand, might be con-
sidered only marginally relevant for this selection decision because it does not in-volve speaking, and the tasks are not at all like the relevant task in the TLU domain.

Type 2 warrant: Utility (that the score-based interpretation is actually useful for making the intended decision).

Utility has to do with the extent to which the scored-based interpretation provides information that increases the probability of making the appropriate decisions, or of avoiding decision errors. Thus, even though a multiple-choice test of general English proficiency might be considered only marginally relevant for selecting employees in the hotel example, it might turn out to predict quite accurately which test takers will be productive employees. If this were the case, then backing for the warrant of the relative utility of this assessment would need to be weighed against the backing for the relative utility of the oral assessment.

Type 3 warrant: Intended consequences (that the consequences of using the assessment and making intended decisions will be beneficial to individuals, the program, company, institution, or system, or to society at large).

Intended consequences pertain to the beneficial outcomes that the test user expects to achieve by using the assessment. In the hotel example, the intended consequence of not offering HH the position is the warrant that this will avoid unreasonable costs to the company—costs associated with employing a person who is not likely to be a productive employee, thus potentially affecting both the image and the profitability of the company adversely. In addition, if the warrant states that a nonproductive employee is not likely to be happy, then the warrant would also imply that HH will be better off working somewhere else.

A fourth type of warrant provides a basis for using a particular assessment or set of assessments as a basis for making decisions. Even though a given assessment might be interpreted as a valid indicator of an ability that is relevant to the decision, and that is useful to making the decision, it may be that this assessment, by itself, does not adequately cover the full breadth of the construct that is of interest for purposes of making the decision. This is, in essence, the argument in favor of multiple indicators, discussed earlier. There is always the possibility, in theory, at least, that adding one more assessment will provide more complete information about the ability of interest. This warrant provides a basis for claiming that the particular assessment(s) used are sufficient for making the decision.

Type 4 warrant: Sufficiency (that the assessment[s] provide sufficient information for making the decision).

Sufficiency will be a function of how the construct that is relevant to the decision is defined in any specific situation. Sufficiency is also related to the issue of what all gets included in the definition of the construct to be assessed. This issue has been dis-
cussed in the language assessment literature regarding whether, or to what extent, topical knowledge or skills, in addition to language ability, are included in the construct. Douglas (2000), for example, argues that in language for specific purposes contexts, the construct of interest includes both language ability and topical knowledge in the area of specialization. McNamara (1997), approaching this issue from the perspective of performance assessment, distinguishes what he calls a “strong performance hypothesis” from a “weak performance hypothesis.” In an assessment use argument, a Type 4 warrant stating that an assessment of language ability, by itself, is sufficient for making decisions about selection or certification corresponds to McNamara’s “strong” hypothesis. A Type 4 warrant stating that language ability is necessary but not sufficient by itself for such decisions corresponds to McNamara’s “weak” hypothesis. In the hotel example, the oral assessment of language ability is clearly not considered to be sufficient for making the selection decision, because the company also administers an assessment of the applicants’ professional knowledge and skills related to the specific job in question.

Each of these warrants will, of course, require appropriate and adequate backing, such as prior research, evidence specifically collected for this purpose, accepted community social practice and values, government regulations, laws, and legal precedents.

We can illustrate an assessment utilization argument by continuing from our previous example. Assuming that the score-based interpretation about HH’s ability is valid, the utilization argument for this assessment might be as follows:

Data (interpretation): HH has a low level of ability to interpret and record specific topical information related to making hotel reservations.

Claim (decision to be made): HH will not be offered the position.

Warrant 1 (relevance): The ability to elicit, interpret, and record (in English) specific topical information related to making hotel reservations is a critical part of the job description for the position advertised.

Backing 1: Needs analyses indicate that this task is regularly performed by employees in this job position. This task is listed as a critical task in the department’s job description for this position.

Note that the first warrant and its backing has also provided support for the validity argument, thus reinforcing the link between validity and use.

Warrant 2 (utility): Scores on this test are good predictors of the language ability needed for successful future job performance.

Backing 2: Past experience indicates that individuals who achieve high scores on this test are productive and perform their job efficiently. (Note that this backing is unidirectional; there is no evidence that individuals who receive low scores are not productive and efficient employees. Additional evidence would need to be collected to provide this backing.)
Warrant 3 (intended consequences): Individuals who have a high level of ability to interpret and record specific topical information related to making hotel reservations, and who are knowledgeable and skilled professionally, contribute positively to the overall performance of the department, increase profitability, and minimize customer complaints.

Backing 3: Past experience indicates that individuals who have a high level of ability to interpret and record specific topical information related to making hotel reservations, and who are knowledgeable and skilled professionally, contribute positively to the overall performance of the department, increase profitability, and minimize customer complaints.

Warrant 4 (sufficiency): Additional sources of information are not useful for assessing applicants’ English ability.

Backing 4: Previous experience indicates that other indicators of English ability, such as academic transcripts, grades from EFL courses, self-reports, and letters of reference are often poor predictors and can be misleading in that they either are unrelated to the specific job responsibilities or tend to inflate the applicant’s capabilities. Experience indicates that scores from this test, in conjunction with scores from the professional knowledge exam, provide the most reliable, valid, and efficient predictor of future job performance.

The structure of this argument is illustrated in Figure 8.

Just as with the validity argument, there are likely to be rebuttals in the utilization argument. Rebuttals to each of the four types of warrants discussed earlier could be articulated, and backing could be provided that either rejects or supports these. In addition to rebuttals of four types of warrants discussed earlier, a utilization argument will include two general types rebuttals. (Again, practical experience with this approach may reveal additional types of rebuttals.)

Type 1 rebuttal: Reasons for not making the intended decision, or for making a different decision.
Type 2 rebuttal: Unintended consequences of using the assessment and/or making the decision.

Continuing with our saga of HH and the phone company, the rebuttals might be as follows:

Rebuttal 1 (alternate reasons): HH could be someone’s friend or relative.
  Rebuttal data 1a: HH is the son of the hotel company’s owner (supports rebuttal).
  Rebuttal data 1b: HH and the HR Manager went to the same school (supports rebuttal).
  Rebuttal data 1c: HH has no connections with anyone in the hotel company (rejects rebuttal).
Rebuttal 2 (unintended consequences): Not hiring HH might endanger the hotel company.

Rebuttal data 2: HH has friends in the cement and explosives businesses and he knows where the hotel company owner lives (supports rebuttal).

The structure of these rebuttals and rebuttal data are illustrated in Figure 9.

In summary, an assessment utilization argument for a given assessment can be articulated in terms of claims (decisions to be made), warrants justifying making these decisions on the basis of an interpretation from a language assessment, and backing, or evidence supporting these warrants. The warrants justify making the
FIGURE 9  Rebuttals for use of interpretations from hotel employment language test.
decision in terms of (a) the relevance of the score-based interpretation to the decision, (b) the utility of the interpretation for making the decision, (c) the intended consequences of using the assessment and making the decision, and (d) the sufficiency of the assessment for making the intended decision. The backing can consist of evidence collected from a variety of sources, such as prior research, specific validation research to support the warrant that is conducted as part of the test development process, prevailing institutional and societal norms, government regulations, and laws. The utilization argument may also include rebuttals to specific warrants, as well as rebuttals articulating (a) reasons for not making the intended decision or for making a different decision and (b) potential unintended consequences of using the assessment and making the intended decisions. Backing to reject these rebuttals can consist of evidence collected from a variety of sources.

An assessment use argument is the overarching logical structure that provides a basis both for test design development and for score interpretation and use. In test development, we would typically begin by developing the assessment utilization argument, articulating the intended uses of the assessment. Then, given the nature of the decisions to be made, and the type of information needed to make those decisions, the assessment validity argument would be articulated. Even though the assessment use argument includes both the utilization and validity arguments, these two arguments are logically distinct because there is not a one-to-one relationship between interpretations and decisions. That is, we will often find it useful, particularly in high-stakes situations, to base a single decision on multiple assessments, and hence potentially multiple interpretations, that combine to provide a much broader representation of the construct than could be obtained from a single assessment. Similarly, a valid interpretation, based on a single assessment, might turn out to be useful and justifiable for making several different kinds of decisions. Lest this be misunderstood, a strong word of caution is in order. For either multiple assessments for a single decision or for multiple uses of a single assessment, the assessment utilization and validity arguments would need to be articulated, and convincing evidence would need to be collected to support the warrants and reject any potential rebuttals in these arguments. The structure of an assessment use argument, including the relationship between the assessment utilization and validity arguments, from the perspectives of test design and development and interpretation and use, is presented schematically in Figure 10.

QUALITIES OF USEFULNESS, VALIDITY, AND FAIRNESS
AND CONCERNS OF CRITICAL LANGUAGE TESTING
IN AN ASSESSMENT USE ARGUMENT

As indicated earlier, in the past few years a number of language testers have written extensively about issues of language use, and have expanded the list of qualities that we need to consider in designing, developing, and using language tests. In this
section I will discuss the extent to which these qualities and considerations can be addressed in an overall assessment use argument. Specifically, I will discuss ways in which these qualities and considerations might constitute warrants or rebuttals in either an assessment validity or an assessment utilization argument, or in both.

Test Usefulness

Bachman and Palmer (1996) argue that the overall usefulness of a test is a function of six qualities, as discussed earlier. They argue that the first five of these qualities—reliability, construct validity, authenticity, interactiveness, and impact—are qualities that the test developer needs to design into the test, and that the test user expects to see in the results. Practicality, on the other hand, is essentially a function of how the test developer prioritizes the first five qualities, and the implications this has for the kinds and amounts of resources that will be required. Practicality is thus not a quality of the test itself, but rather of the test development process; if the designer attempts to maximize all five qualities and has limited resources, the test simply will not get developed. If it is too costly to administer and score, it will not get used.
Looking at the first five qualities of usefulness, reliability and construct validity consist of warrants that would support the assessment validity argument, whereas the investigation of these qualities would provide the backing to support these warrants. The quality of impact, which for Bachman and Palmer includes washback, or impact on instruction, as well as the ways in which test use affects individuals, institutions, and society at large, could quite clearly be broken down into Type 3 Warrants in an assessment utilization argument about the intended consequences of the particular assessment. The qualities of authenticity and interactiveness would seem to be less clear cut. On the one hand, these could be seen as warrants in support of the validity argument. Authenticity constitutes a warrant that the assessment tasks correspond to language use tasks outside the test itself, and that we can extrapolate from universe scores to target domain scores or interpretations. Similarly, interactiveness could be seen as a warrant that the processes that are engaged by the assessment task are essentially the same as those defined in the construct. On the other hand, authenticity, as a perception, will also provide a Type 1 warrant in a utilization argument that the assessment and the score-based inference is relevant to the decision to be made. That is, if test takers and test users perceive the kinds of tasks that are included in the assessment, and the ability that is being assessed, to be relevant to the decision to be made, this will provide support for the intended use. This is essentially the warrant that underlies what has been called “test appeal,” or “public relations” (sometimes still referred to as “face validity”). Authenticity is, to a large extent, the warrant that underlies much of the appeal of so-called “performance assessments,” in which the characteristics of assessment tasks correspond closely to those of tasks in the target domains to which inferences are intended to generalize. Interactiveness likewise can function as a Type 1 warrant in a utilization argument to support the relevance of the interpretation to the decision. If the assessment task engages the same areas of ability as are engaged by target language use tasks, then this supports the relevance of the interpretation to the intended use.

Ethics and Validity

Lynch (2001) discusses a “validity framework” that includes five categories, as mentioned earlier. Four of these categories in can, in my view, be considered as Type 3 warrants in a utilization argument about the intended consequences of a given test use. If, for any given assessment, it is desirable for the participants in the assessment process to establish meaningful identities of themselves (ontological authenticity), to gain improved understanding of perspectives outside their own group (cross-referential authenticity), and to change the way in which they relate to each other and to themselves (evolved power relations; Lynch, 2001, pp. 365–366), then these could easily be included as Type 3 warrants about the intended consequences of the assessment. Similarly, the category of impact/conse-
quential validity could be covered by Type 3 warrants about intended consequences. Lynch’s first category, “fairness,” however, would seem to be a much broader consideration, not specifically tied to any particular test use. Lynch defines fairness in terms of the extent to which the perspectives of all affected participants in the assessment process are taken into account, and the extent to which the assessment is structured to maximize ethical behavior (Lynch, 2001, pp. 364–365). Since fairness, for Lynch, would appear to pertain to the broader enterprise and process of designing, developing, and using assessments, it would seem to be more of a standard for assessment practice in general, rather than a consideration of the use of a particular assessment.

**Fairness**

Kunnan (2003) describes a “test fairness” framework that includes five qualities, as mentioned earlier. Under “validity” Kunnan includes the standard kinds of validity-related evidence, along with reliability, and these would thus constitute warrants in support of the validity argument. Under “administration” Kunnan includes procedures about physical conditions and uniformity of administration that would normally be implemented to ensure consistency and that are intended to enhance the reliability of scores and validity of inferences. Warrants related to consistency in administrative conditions and procedures would thus also support the validity argument. The remaining three of Kunnan’s qualities—absence of bias, access, and social consequences—are, in my view, essentially aspects of impact, relating to assessment use. Bias consists of systematic unequal impact on specific subgroups, in terms of whether they are fully included in the assessment, how they are treated as part of the assessment process, how they perform, how their performance is scored, and how their test scores are used. One could conceivably state a Type 3 warrant in a utilization argument about absence of bias being an intended consequence of the assessment. However, as bias analysis is currently practiced, it would seem that the potential sources of bias for a particular assessment constitute Type 2 rebuttals about unintended consequences, and the kinds of analyses, both judgmental and statistical, that are conducted, are intended to provide backing to reject these rebuttals. Access is also an aspect of impact because, if particular individuals who should or want to take part in the assessment are not able to do so for various reasons (e.g., educational, financial, geographic, personal, equipment, and conditions for administration), this will bar these individuals from being considered in whatever decisions the assessment is intended to support. Warrants about access are thus Type 3 warrants about intended consequences of the assessment. Finally, social consequences are clearly an aspect of impact, which could be stated as either Type 3 warrants about the intended social consequences of the assessment, or as Type 2 rebuttals, about potential unintended social consequences.
Critical Language Testing

Shohamy (2001b) discusses a number of principles and desiderata that, for her, constitute the main features of critical language testing. Without going through these one by one, I believe that many of these could be articulated as warrants or rebuttals in validity or utilization arguments, whereas others appear to be more germane to a code of ethics or code of practice. Some of Shohamy’s principles—such as questioning the values, agendas, and needs that inform the test; questioning the goals that the test is intended to fulfill; and questioning the purposes and actual uses of the test—could be articulated as Type 3 warrants in a utilization argument about intended consequences or as potential rebuttals to the intended use. Questions about whose values, agendas, and needs are, or should be, considered in the assessment, could be incorporated in the backing that supports warrants about intended consequences. That is, the backing could specify where the evidence supporting the warrant comes from, and thus identify the individuals or institutions that are making the warrant. Questions about the knowledge on which the test is based could be articulated as both warrants or rebuttals in a validity argument and as Type 1 warrants or rebuttals in a utilization argument about the relevance of the ability measured to the intended use. Questions about whose knowledge this is, or should be, could be incorporated in the backing that supports warrants about validity and relevance. Considerations about the need for additional sources of information other than a test could be articulated as Type 4 warrants in a utilization argument about the sufficiency of the measures for the decision to be made. Other principles—such as the claim that “language testing is not neutral,” the need for language testers to develop a critical view of tests and to critique the value of these, and the view of test takers as “political subjects in a political context” (Shohamy, 2001b, p. 131)—are, it seems to me, premises that could well be incorporated into a code of ethics or a code of fair testing practice, rather than being articulated in an assessment use argument for a particular test.

In summary, it would appear that the validity and utilization arguments for any given language assessment could incorporate as warrants or rebuttals many of the qualities that have been discussed previously in frameworks of usefulness, validity, or fairness, or as principles of critical language testing. I would thus maintain that these arguments can enable test developers and users to more clearly articulate, for a particular test, in ways that are operational, many of the qualities and concerns that language testers have discussed regarding test use.

SOME PRACTICAL ISSUES

Though I have outlined the structure of an assessment use argument and described how the two parts of this—the utilization and validity arguments—are linked, there are some practical issues that also need to be addressed. One issue is that of
how one goes about implementing an assessment use argument in practical test design, development, and use. Another set of issues pertain to the feasibility of doing this in real-world test development settings in which there are often insufficient time and resources to conduct even the bare minimum of validation research. Because the focus of this article is on the logical structure of assessment use arguments, I will only touch on some of the practical issues. Hopefully, if language testers find this approach useful, experience with the approach will grow, and studies reporting on practical issues in actual test development settings will be forthcoming.

Implementation

Though I have discussed an assessment use argument as a two-part procedure, we could well ask which comes first and why, or if it matters where we begin. If, as many have argued, the most important consideration in the design and development of a test is the use or uses for which it is intended, then it would make sense to begin by clearly articulating these intended uses or decisions as claims in the utilization argument. We would then articulate the warrants in support of these intended uses. We would most likely start with Type 3 warrants about the intended consequences of the test, along with potential rebuttals about unintended consequences. Then we might think about the kinds of information needed to support the intended decisions, which would lead to consideration of the attributes of the individuals that are relevant to the decision (Type 1 warrants) and the different ways in which we might want to obtain this information (Type 4 warrants). If multiple indicators are considered necessary or desirable, then warrants need to be articulated in support of the specific kinds of indicators that will be used. At this stage, as possible constructs and information gathering procedures start to emerge, we would start constructing the validity argument, including the interpretations we want to be able to make, the kinds of data we would need to obtain from test takers to make these, and the warrants and types of backing needed to support the link between data and interpretations. Though this may appear to be a fairly linear process, in practice, it is likely to be much more interactive and cyclical, with modifications and additions being made to both the utilization and validity arguments as test design and development proceed and as evidence in support of these arguments accumulates.

The design, development, and use of a language test is a complex and technical undertaking, particularly in high-stakes situations, and requires access to professional expertise in both applied linguistics and measurement. In addition, test development and use always takes place in a value-laden sociopolitical context, where values, needs, and pressures from different stakeholders—individuals, groups, and institutions—may compete or even conflict with one another. “For example, conflicting values about the uses of scores from literacy tests could be formed into quite different arguments by students, teachers, school administrators,
minority-background parents, majority-background parents, knowledge-based industries, technical industries, politicians, and government-funding agencies” (anonymous reviewer, personal communication, 2004). In many settings, particularly those involving high-stakes uses of tests, it may be extremely difficult to develop a procedure or rationale for considering and prioritizing these competing values in a principled way. The articulation of an assessment use argument cannot ultimately solve the problem of considering, balancing, and perhaps prioritizing the values of different stakeholders. However, as I have argued earlier, an assessment use argument may provide a framework or procedure for at least clearly articulating what the competing values are and which stakeholders hold these values. Articulating the values of the relevant stakeholders openly, and getting these on the table would, one would hope, facilitate the negotiation of how these get included or prioritized in the test development process.

Feasibility

A number of researchers (e.g., Haertel, 1999; Kane, 2001) have pointed out that the demands of a strong program of validation, based on explicitly stated hypotheses and assumptions, can be quite heavy, both in terms of the theoretical base needed and the resources required. In addition, test developers seldom have the inclination, time, or funding to carry out a strong program of validation research in support of a given test use. Ironically, this appears to be particularly endemic to many high-stakes assessments, where the need for validation is the greatest. As Haertel put it, “undertaking studies that no one else is doing, that might call our work into question, and that might be used against us in court is, to say the least, unappealing” (Haertel, 1999, p. 6). As a result, test developers in practice have typically adopted a weak program of construct validation that may consist of whatever evidence is relatively easy to collect that provides support for the intended interpretation, or what Cronbach has called “sheer exploratory empiricism” (Cronbach, 1988, p. 12). In the face of this, we might well ask how we can expect test developers to engage in a program of assessment use research, which would extend the typical validation research to include research into the uses and consequences of the assessment.

I will not attempt to address issues of feasibility here. Rather, I would argue, as have others, that it is exactly because of limited time and resources for test development, along with the frequent lack of strong theoretical assumptions, that the articulation of an assessment use argument is so important and useful. Without a coherent use argument, test developers may expend valuable resources collecting evidence to support warrants that may not be questioned, rather than focusing on evidence that is most critical to the assessment use argument. In this regard, it should be noted that the assessment use argument is only as strong as its weakest link, which means that test developers should focus their resources on collecting evidence to reject the most likely or most serious rebuttals to their intended interpretations and uses. These can only be identified by articulating an assessment use
argument. I would thus argue that the time and effort expended in articulating the assessment use argument are resources well spent, as this will increase the feasibility of collecting evidence, because it will guide the test developer and enable her to focus on that evidence that will provide the strongest, most convincing support for the argument as a whole.

CONCLUSION

Discussions in language testing about the validity of score-based inferences and the consequences of test use have historically followed parallel trajectories, often with little more than superficial linkage between the two. Attempts to bring the two together into a single framework have provided little guidance for practice, other than lists of qualities and considerations to which test developers and test users should attend. I have argued that considerations of validity and consequences can be brought together by articulating a two-part assessment use argument. One part, the assessment validity argument, links assessment performance with interpretations, whereas the other part, the assessment utilization argument, links score-based interpretations with intended uses or decisions. The two arguments—validity and utilization—are integrated in an overall assessment use argument that provides explicit links from test performance to interpretations to test use.

Returning to Spolsky’s two questions, “How sure are you of your decision?” and “How sure are you of the evidence that you’re using to make that decision?” we can see that an assessment use argument provides a logical framework for addressing both. The first question is addressed in the assessment utilization argument, which explicitly articulates the warrants that justify the use of the assessment for making the decision, the possible rebuttals, or reasons for not using the assessment for the decision, and that links the decision directly to an assessment result—an interpretation. The second question is addressed in the validity argument, which articulates the warrants that justify the intended interpretation of the assessment result, the possible rebuttals or alternative interpretations, and which links the performance, or the evidence, directly to the interpretation.

Assessment qualities such as validity, usefulness, and fairness do not exist in the abstract; they pertain to specific assessments, indeed, to the way we interpret and use the test performance of each individual or group of test takers. Similarly, the intended uses and potential consequences of this use are essentially specific, local concerns in any assessment that need to be addressed regarding the stakeholders in that specific assessment—the various individuals who will be affected in one way or another by the assessment and by the way we use the it. We need to be able to justify the interpretations and uses we make of any assessment so that we can be accountable to the stakeholders. Therefore, the primary reason for articulating an assessment use argument, and for collecting evidence in support of this, is so that we can be accountable to stakeholders. The warrants in the assessment use
argument guide us in determining what kinds and how much evidence we need to collect to convince relevant stakeholders about the usefulness of our assessment. By demonstrating, through logical argumentation and the collection of supporting evidence, that our assessment is useful for its intended purpose, we provide the justification we need to be accountable to the individuals who are affected by the assessment and the way it is used.

In practice, arguments for intended interpretations and uses are not adequately articulated for many language assessments. This is particularly true for many large-scale assessments that are used for high-stakes accountability purposes (Koenig & Bachman, 2004). Efforts by test developers to collect evidence relevant to the validity of score-based interpretations are frequently shopping lists of correlations, content analyses, and other evidence collected more or less as time and resources permit. By articulating an assessment use argument, test developers and test user alike can arrive at a much clearer and more transparent conceptualization of interpretations and decisions they intend to make on the basis of an assessment, of the ways in which these interpretations and uses need to be justified, and the kinds of evidence they need to collect to support these intended interpretations and uses. An assessment use argument can thus not only guide the design and development of assessments, it can also lead to a much more focused, efficient program for collecting the most critical evidence (backing) in support of the interpretations and uses for which the assessment is intended.

ACKNOWLEDGMENTS

This is an expanded and revised version of the International Language Testing Association Lifetime Achievement Awardee Lecture given at the 26th Annual Language Testing Research Colloquium, Temecula, CA, March 28, 2004. I thank Carol Chapelle, Antony Kunnan, Robert Mislevy, Adrian Palmer, and Jim Purpura for their very helpful comments on earlier drafts of this article. My thanks, as well, to several anonymous reviewers for their valuable comments and suggestions.

REFERENCES


