94

Ongoing Challenges in Language Assessment

Lyle F. Bachman
University of California, Los Angeles, USA

Introduction

In considering the challenges that face the field of language assessment in the early decades of the 21st century, it is clear that many of those that have persisted over the years continue to engage us. At the same time new challenges, engendered in part by our own accomplishments and the progress made over the past half century and in part by changes in the economic, social, and educational contexts of the 21st century, assure us that the field will continue to be a vibrant and exciting one. Thus, while the field has matured both in the breadth of research questions it addresses and in the range of research approaches at its disposal for addressing them, it still grapples with the questions that are fundamental to our enterprise. What is the nature of language ability? How can we assure that the interpretations about test takers' language ability at which we arrive on the basis of assessment results are meaningful to them and other stakeholders? How can we assure that these assessment-based interpretations generalize to language use situations beyond the assessment itself? To what extent can we justify the uses for which our assessments are intended? To what extent do our assessments, the uses to which they are put, and the consequences of these uses respect the individual rights and the societal and educational values of those who are affected by these uses and consequences?

Addressing these questions has led to a better understanding of their complexity, their persistence, and the importance of continuing to address them through research. We now have a broader and more inclusive view of language ability, along with a wide range of methodologies—both quantitative and qualitative—that we can employ in the development of practical language assessments and in basic research. We also have a clearer understanding of the social and political contexts within which the uses of language assessment are embedded and a firmer

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla128

grasp both of the ethical issues involved in using language assessments and of how these contexts and issues need to inform and shape what we do. In short, the field is in a much better position to deal with the issues of developing and using language assessments in the real world than ever before.

At the same time, it is important to realize that theoretical frameworks of language ability, sophisticated research methods, social theory, and moral, ethical, and philosophical explorations all provide, at best, general guidance for the craft or practice of language assessment, which is to develop language assessments whose intended uses can be justified to those whose lives will be affected by these uses. The richness and complexity of the theoretical and methodological frameworks that now underlie our practice heighten the single most important and continuing challenge for language testers: how to discharge our duties *responsibly*—both methodologically and ethically—as we develop and use language assessments in the real world.

New challenges for the field have also arisen from the increasing worldwide demand for individuals with high levels of language ability. These demands are twofold: (1) huge and growing numbers of students worldwide whose native language is not the language of instruction and who may need to learn the majority or "official" language of a country in order to become fully functioning members of the society; and (2) globalization and the increasing demands for employees who can function in multilingual work settings. Along with these growing demands for high-level users of languages has come an increasing demand for accountability in language teaching (see below). Governments, from nations to states to school districts to local schools, are increasingly requiring that educational institutions and teachers be held accountable for the levels of language ability attained by learners, given the resources—human as well as time, space, and money—that have been expended. Similarly, corporations and businesses are increasingly expecting educational institutions—schools, colleges, and universities—to produce potential employees whose language ability is sufficient for them to function in a multilingual workplace. These demands for accountability reinforce schools' and teachers' normal interest in providing instruction that is more effective and appropriate for enhancing their students' learning. In virtually all such situations, the tools for collecting information that will inform decisions—both accountability decisions and instructional ones—are language assessments.

Growing numbers of "young language learners" in schools pose challenges for classroom language assessment as well as for high stakes accountability assessment. For classroom language assessment, the challenge is how to apply the knowledge we have acquired (1) to develop assessments that will serve the purposes of learning and instruction; and (2) to provide training in language assessment for classroom language teachers. For accountability assessments, the challenge is how to apply the knowledge we have, as *language* testers, to inform the kinds of assessments that are made of students' achievement not only in the language of instruction, but also in other areas, such as math and science, where the language of the assessment may not be the native language of the test takers.

The displacement of huge numbers of individuals across countries and continents, whether voluntary or involuntary, due to political unrest, economic

hardship, or personal circumstances, presents another kind of challenge for the field. In many such situations governments require those seeking to immigrate to demonstrate proficiency in a particular language. In the case of individuals who are voluntarily intending to immigrate in order to seek employment, governments typically require them to demonstrate proficiency in the dominant or official language of the country. In cases where individuals are involuntarily seeking political asylum, governments may wish to determine what their native language or dialect is in order to make a decision about granting them asylum. Again, the instruments that are used to collect information to support these decisions are language tests.

In this chapter I will describe what I regard as issues of continuing concern, as well as the new challenges that the field of language testing is facing—or will face in the years to come. I will then briefly describe an "assessment use argument" as a conceptual framework for problematizing many of these issues and for providing a principled basis for bringing together the rich diversity of research approaches at our disposal in order to investigate them empirically. I will conclude by pointing out that these challenges also offer opportunities for those language testers in the 21st century who are willing to address them.

Issues of Continuing Concern

Several issues have concerned language testers for the past decade or so: (1) the validity of score-based interpretations and the nature of the construct we want to assess—language ability; (2) ethics and professionalism in the way we develop and use language assessments; (3) the role of language assessments in accountability decisions; and (4) the impact of assessments on instruction.

The Validity of Assessment-Based Interpretations: The Nature of Language Ability

A major requirement of any language assessment is that the interpretations we make about test takers' language ability on the basis of assessment results be valid. What this requirement entails is that the assessment results can be interpreted as indicators of the areas of language ability we want to assess, and of very little else. In the past 30 years the conceptualization of validity has evolved considerably, but central to all of these conceptualizations is the notion that the test developer and/or test user have defined the construct or ability that is to be assessed. For language tests, this construct is language ability. Thus one major area of inquiry continues to be the nature of language ability. In the past 35 years the field has seen a move from viewing language ability/proficiency as a unitary or global ability (e.g., Oller, 1979) to a view that language ability is multicomponential (e.g., Canale, 1983; Oller, 1983; Bachman, 1990; Bachman & Palmer, 1996). The dominant view in the field continues to be that language ability consists of a number of interrelated areas such as grammatical knowledge, textual knowledge, and pragmatic knowledge and that these areas of language knowledge are managed by a set of metacognitive strategies that also determine how language ability is realized in language use or in the situated negotiation of meaning (Bachman, 1990; Bachman

& Palmer, 1996; Purpura, 1998; Chapelle, 1998, 2006; Phakiti, 2003, 2008). Researchers who focus more closely on the nature of the interactions in language use have argued that the view of language ability as solely a cognitive attribute of language users ignores the essentially social nature of the interactions that take place in discourse. These researchers argue that language ability resides in the contextualized interactions or discursive practices that characterize language use (e.g., Chalhoub-Deville, 1995, 2003; McNamara, 1997, 2003). More recent research with paired and group interviews—which are oral assessments in which two or more test takers speak with each other rather than with, or in addition to speaking to, an examiner—suggest that, while such assessments can engage test takers in interactive language use, actually measuring the interactional competence of individual test takers can be problematic for both methodological and ethical reasons.

In a critical review of this debate, Bachman (2007) identifies three different approaches to defining language ability: (1) ability-focused, (2) task-focused, and (3) interaction-focused. He concludes that the theoretical issues raised by these different approaches to defining the construct, language ability, are challenging both for empirical research in language testing and for practical test design, development, and use. For language-testing research, these issues imply the need for a much broader methodological approach, involving both quantitative and qualitative perspectives. For language-testing practice, they imply that focus on ability, task, or interaction, to the exclusion of the others, will lead to weaknesses in the assessment itself or to limitations on the uses for which the assessment is appropriate.

A closely related issue is that of the extent to which language ability includes topical knowledge. The effect of test takers' topical or content knowledge on language test performance is well documented in the language assessment literature (e.g., Alderson & Urquhart, 1985; Douglas, 1997), and the dominant view has been that this is a source of bias in language tests. That is, in designing a language test and in interpreting scores from such a test, it is either generally assumed or specifically stated that "language knowledge" or "language ability" is what we want to assess, and not test takers' content knowledge. An alternative, or perhaps complementary, view has been articulated in the area of language for specific purposes (LSP) assessment. According to this view, what we want to assess is what Douglas (2000) has called "specific purpose language ability," which is a combination of language ability and background knowledge. Davies (2001) has argued that LSP assessment has no theoretical basis but can be justified largely on pragmatic grounds. Bachman and Palmer (1996) have argued that whether one includes topical knowledge as part of the construct to be assessed in a language test is essentially a function of the specific purpose for which the test is intended and of the levels of topical knowledge that the test developer can assume test takers to have.

As John B. Carroll (1973) noted 40 years ago, questions about the nature of language ability and the validity of score-based interpretations will be a perpetual concern for language testers. In terms of ontology, there will always be debates about whether language ability actually exists in the "real world" and, if so, where, while in terms of epistemology researchers will undoubtedly debate approaches to understanding precisely what language ability is (see Bachman,

2006a for a discussion of these issues). Furthermore, as Bachman (2007) has pointed out, the field has seen numerous approaches to defining this construct, and we are not likely to see universal agreement on any particular "model" in the near future. Nevertheless, in terms of practical research and development aimed at providing language assessments that can be justified to stakeholders, language testers will be well advised, in my view, to use these philosophical and theoretical issues more as general guidelines for informing the way the construct of language ability is defined for any particular language assessment and less as scientific theories of language ability that can somehow be verified through research and development. The question of validity, then, is not whether, or to what extent, a given test score can be seen to be an indicator of some abstract theoretical model of language ability, but rather whether score-based interpretations are meaningful and can be justified to stakeholders.

Issues of Ethics and Professionalism in Language Assessment Use

Although validity and validation continue to form a major area of focus in language assessment research (e.g., Bachman, 2005), this is no longer the sole, or even the dominant, concern of the field. Language testers are investigating difficult questions about how and why language assessments are used, about the ethical responsibilities of test developers and users (e.g., Stansfield, 1993; McNamara, 1998, 2001), about fairness in language assessment (e.g., Elder, 1997; Kunnan, 2000a, 2004), about the impact and consequences of assessment use (e.g. Shohamy, 2001), particularly on instructional practice (e.g., Alderson & Wall, 1993, 1996; Cheng, 1997; Wall, 2005), and about the societal values that underlie such use and the larger sociocultural contexts in which language tests are used (e.g., McNamara & Roever, 2006).

Language testers are still debating issues of fairness and professionalism and will no doubt continue to do so for the foreseeable future. And while to some this ongoing debate may reflect a lack of progress and consensus in the field about these critical issues, I view it as healthy for a number of reasons. First, it reflects the intense commitment of language testers to assuring that language assessments are developed professionally and used fairly. Second, it engages language testers with other discourse communities—such as philosophers, who are grappling with ethics—and with other professions—such as medicine and law, which must also deal with issues of professional ethics. Finally, what I find extremely encouraging is that these two strains of research and concern are coming together in a growing body of research that investigates both the validity of score interpretations and the consequences of assessment use (e.g., papers in Kunnan, 2000b; Bachman, 2005; Bachman, 2006b).

Language Assessment for Accountability

The assessment demands of No Child Left Behind (NCLB) in the US (United States Congress, 2001) have greatly increased the pressure on states to develop more useful assessments, both for accountability and in the service of classroom language learning. In neither area, in my view, have language testers been adequately

involved. Of particular concern to language testers and other applied linguists should be issues of assessing the English language development and academic achievement of English language learners (ELLs).

Recent initiatives on the part of the US government to increase that nation's capacity in foreign languages are also placing great demands for useful assessments of foreign languages, particularly the less commonly taught languages (US Department of Education, US Department of State, US Department of Defense, & Office of the Director of National Intelligence, 2006). As increasingly larger amounts of government resources are likely to be going into foreign language instruction in the coming years, at all levels, from federal, state, and local authorities, there will most likely be a concomitant need for greater accountability. In K-12 education an accountability mechanism is already in place, through NCLB; for better or worse, one can expect that, as the federal government invests more heavily in language instruction at this level, an accountability mechanism will be required and that this will necessitate the development of assessments of foreign language proficiency that meet accepted professional standards for validity and impact.

Similar demands for the involvement of individuals with expertise in language assessment can be found in countries around the globe, where governments and institutions are applying increased pressure on language testers to develop language assessments whose results can be meaningfully interpreted on a common scale of language ability. In Europe, for example, governmental policy is driving massive efforts to develop language assessments in all 14 languages of the European Community, as well as requiring that high stakes language assessments be reported on a single scale: the Common European Frame of Reference (Council of Europe, 2001). Similar efforts are being implemented in many other countries, where the need for high stakes accountability assessments is being driven by the demand for individuals with higher levels of language ability (see, for example, the papers in Martyniuk, 2010).

The worldwide demand for high-level users of a wide range of languages is unlikely to diminish in the foreseeable future; this demand will continue to create a need for accountability; and this need, in turn, will inevitably sustain the ongoing need and demand for language assessments. In my view, in their rush to meet the political demands of governments and other institutions for language assessments, language testers in general have not adequately considered the issues of professionalism and fairness discussed above. For example, rather than asking governmental agencies or institutions questions like "*Why* do you want us to report our test results on a common international scale?" or "How can we *justify* doing this?," language testers are taking the easy way out and making claims about their assessments that may or may not be justifiable, merely in order to satisfy the political agendas of governments and institutions.

Impact on Instruction

The impact of language assessments on instruction (also referred to as "washback") was for many years considered to be relatively straightforward: "good" assessments would cause teachers to follow "good" instructional practice, while "bad" assessments would cause teachers to follow "bad" instructional practice. It

may be implicitly recognized that language tests can have a positive impact on instruction by promoting instructional practices that teachers and educators consider to be appropriate and effective for learning. Nevertheless, much of the discussion around washback (a process also referred to as "backwash") has focused on the negative effects of assessment on teaching, notably its leading to instructional practices that teachers and educators believe are detrimental to learning, such as the phenomenon of "teaching to the test" and the "narrowing of the curriculum."

It was not until language-testing researchers rigorously investigated washback empirically that the field began to realize the complexity of this phenomenon. Two large-scale studies, both of which investigated attempts planned by governments in two different countries to engineer changes in English teaching curricula and in instructional practice through changes in public English examinations, were instrumental in demonstrating to the field that washback is neither simple nor straightforward. The first study was Wall and Alderson's pioneering research into the impact of introducing a change in the English part of the secondary school-leaving examination in Sri Lanka (Alderson & Wall, 1993; Wall & Alderson, 1993). This study revealed that washback works in different ways and to varying degrees on different parts of an educational system—classroom teachers, curriculum developers, and textbook publishers. The second large-scale study of the impact of language assessment was Cheng's (1997) research into the impact of introducing a test of English language speaking into the secondary school-leaving examination in Hong Kong. Her results supported Wall and Alderson's findings in general and extended them to demonstrate that both classroom teachers and students differed in their perceptions of reactions to the new examination. Many of the issues raised by the Sri Lankan study were addressed in a special issue of *Language Testing*, guest-edited by Alderson and Wall (Alderson & Wall, 1996). Bailey (1999) provides a review of the research into and conceptualization of washback.

As a result of this research and theorizing, the complexity of washback is much better understood, and the field has a much better conceptual base upon which to continue empirical research into this vital area of language assessment. What language testers might want to consider, in my view, is finding ways in which this understanding can be used to inform policy about the use of language assessments in instruction, particularly about its use to engineer educational reform.

New and Recent Challenges

Several new and recent challenges face the field of language assessment: (1) the role of assessment in language classrooms, (2) training classroom teachers in language assessment, and (3) language assessment for citizenship and naturalization.

Classroom Assessment

If we consider the numbers of individuals around the world who are studying languages in classrooms—between about 1 and 2 billion people

are studying English alone worldwide (Graddol, 1997, 2006)—in conjunction with the finding that teachers spend significant amounts of time assessing their students—subject matter teachers in schools up to 40% and ESL (English as a second language) teachers about 25%—we quickly realize what a huge enterprise and undertaking classroom assessment is.

Nevertheless, language testers have been only marginally involved in issues of classroom assessment in schools and adult education, and this is still not considered "mainstream language testing" by many. In the past decade, however, classroom language assessment has emerged as one of the most exciting and challenging areas in our field. In this short time the field has seen a move from virtually no interest in school-based or classroom assessment to a growing interest and body of research and practice in this area.

Language testers have also become increasingly involved in two areas of classroom assessment: the assessment of young language learners; and the role and function of assessment in the language classroom. Seminal research in the assessment of young learners can be found in two special issues of the journal *Language Testing*, both edited by Rea-Dickins (2000, 2004), and in a special issue of the journal *Language Assessment Quarterly* edited by Brindley (2007).

The role and function of assessment in the language classroom have been discussed from two perspectives: that of formative assessment and that of so-called "dynamic assessment." *Formative assessment* can be defined broadly as assessment that takes place during instruction and learning and is intended to provide feedback for the improvement of both. It contrasts with *summative assessment*, which typically takes place at the end of instruction and learning and is intended to provide feedback for making decisions about advancement, progress, or certification. Drawing on work on formative assessment in the field of educational measurement (e.g., Black & Wiliam, 1998), a number of language-testing researchers have discussed the tension between high stakes accountability summative assessments on the one hand and teacher-based classroom assessments on the other; and they argue for increased emphasis on teacher-based formative assessment in the language classroom (e.g., Brindley, 1998; Leung, 2004; Leung & Mohan, 2004; Leung & Rea-Dickins, 2007).

Drawing on research in second language acquisition and on Vygotskian psychology, some researchers have discussed what is called "dynamic assessment," arguing that this form of assessment incorporates what is known about learning in general and language learning in particular and should therefore be the preferred mode of assessment in language classrooms (e.g., Lantolf & Poehner, 2004, 2011). Lantolf and Poehner (2004) further suggest that formative assessment might be reconceptualized within the principles of dynamic assessment.

A slightly different approach to the roles and functions of assessment in the language classroom is proposed by Bachman and Palmer (2010), who describe classroom assessment in terms of features, mode, characteristics, and purpose. They distinguish two modes. The *implicit mode* of assessment is fully integrated with teaching, being characterized as continuous, instantaneous, and cyclical; it is a mode in which the teacher and students are essentially unaware that assessment is taking place. This mode corresponds closely to "dynamic

Table 94.1 Modes of assessment (Bachman & Palmer, 2010, p. 29). © Oxford University Press

Mode	*Characteristics*	*Purpose*
Implicit	Continuous Instantaneous Cyclical Both teacher and students may be unaware that assessment is taking place	Formative decisions, e.g.: Correct or not correct student's response Change form of questioning Call on another student Produce a model utterance Request a group response Summative decisions, e.g.: Pass/fail decision based partly on classroom participation or performance
Explicit	Clearly distinct from teaching Both teacher and learners aware that assessment is taking place	Summative decisions, e.g.: Decide who passes the course Certify level of ability Formative decisions, e.g.: *Teacher*: move on to next lesson or review current lesson *Teacher*: focus more on a specific area of content *Student*: spend more time on a particular area of language ability *Student*: use a different learning strategy

assessment." The *explicit mode* of assessment is clearly distinct from teaching; both the teacher and the students being aware that assessment is taking place. The authors argue that both modes of assessment can serve the purposes of both formative and summative decisions. They illustrate these distinctions in Table 94.1.

Training Classroom Teachers in Language Assessment

Although there are dozens of textbooks in both language assessment and educational measurement that claim to be "practical" and written for teachers, it is widely recognized that teachers are generally neither knowledgable about nor well trained in assessment. And, while courses in language assessment are offered at many colleges and universities around the world, nevertheless, as Brown and Bailey (2008) conclude at the end of their article reporting the results of two surveys of individuals who teach such courses, "there is still much we do not know about how language testing is being taught in language teacher training programs around the world, and how it should be taught" (Brown & Bailey, 2008, p. 373). Furthermore, Leung (2004) points out that assessment is not generally part of the preservice training of language teachers.

What language teachers believe about assessment and what they actually do when they assess in the language classroom have been extensively researched. To date, however, there have been very few studies, in the language-testing literature, about how language teachers are trained in assessment. This is so despite numerous calls, in the language assessment literature, for the need to build teachers' capacity in language assessment. Thus, as Brown and Bailey (2008) note in their

review, most of what is known about teacher knowledge of and training in assessment comes from the field of educational measurement.

Two areas of research and discussion in the literature on training teachers in educational measurement and language assessment are (1) determining what constitutes teachers' assessment knowledge and the degree to which teachers have it and (2) developing and evaluating training programs aimed at helping teachers acquire knowledge of assessment.

Despite the huge demand for teachers who are competent in assessment, and despite calls from the field itself for the need to train teachers in language assessment, the field of language assessment clearly lags far behind its sibling discipline, educational measurement, not only in terms of understanding what language teachers know and need to know about assessment, but also in terms of developing appropriate programs for training classroom teachers in it. Virtually every article in the field that addresses these issues concludes that little is known and more research is needed. Given the huge numbers of language teachers worldwide, addressing these issues will indeed be a daunting challenge for the field.

Language Assessment for Immigration, Citizenship, and Asylum

As Shohamy and McNamara (2009a) point out in their editorial introduction to a special issue of the journal *Language Assessment Quarterly* on the use of language tests for immigration, citizenship, and asylum (hereafter ICA), this relatively recent area of concern and interest among language testing researchers is an outgrowth of the more general concern, discussed above, in professionalism and ethics in language testing. And, while a number of language-testing researchers (e.g., McNamara, 2001; Shohamy, 2001) have been writing about this issue for quite some time, it has only come to the forefront of language testing research in the past half decade. Most of the papers that have been written on this area of concern appeared in 2009, when two volumes, one edited by Hogan-Brun, Mar-Molinero, and Stevenson (2009), and another by Extra, Spotti, and Van Avermaet (2009), along with a special issue of the journal *Language Assessment Quarterly* (Shohamy & McNamara, 2009b), appeared. An excellent review of these three collections can be found in Lee (2011).

Two very general sets of issues have been discussed in the rapidly emerging literature: (1) language ideologies and ideologies of national identity; and (2) the qualities of and justification for specific assessments. A number of researchers have critically analyzed the ICA policies of governments, questioning the language ideologies and ideolo ies of national identity that underlie them (e.g., Blackledge, 2009) as well as the use of language as a requirement for ICA (e.g., Shohamy, 2009). Others have criticized specific language tests, which are used for ICA, either from the perspective of fairness issues (e.g., Eades, 2009) or from that of the technical qualities of the assessment or both (e.g., Kunnan, 2009). Yet others have argued strongly that both the specific assessment that is being used for ICA and the rationale for it are justified (e.g., de Jong, Lennig, Kerkhoff, & Poelmans, 2009), while others have taken a more neutral, proactive approach.

The consideration of the issues to be faced in developing and using language assessments for ICA raises many of the same now familiar ethical questions about

the role and position of language testers in developing assessment that could be used for purposes they themselves may either question or disagree with. Many of these questions are raised by Shohamy and McNamara (2009a, 2009b). To what extent should, or can, language testers themselves become involved in the setting and implementation of public policy? To what extent, and how, can language testers best apply their knowledge and skills to developing language assessments that can be justified for ICA? Addressing these issues will clearly be a challenge for the field.

Justifying the Uses of Language Assessments

Given all these different uses of language assessment, and given the fact that many of them are high stakes—that is, involve making decisions that have major, life-affecting consequences for test takers and other groups of stakeholders—the critical question faced by language testers is: To what extent can we justify the uses for which our assessments are intended? To what extent do our assessments, the uses to which they are put, and the consequences of these uses respect the individual rights and the societal and educational values of those who are affected by these uses and their consequences? As the demands for language assessments have increased and have become even more diverse, there is a growing demand for language testers themselves to be accountable to stakeholders—those who are affected by the uses of language assessments and by the decisions made on the basis of these assessments.

Assessment Justification

Starting from the premise that test developers and decision makers need to be accountable to stakeholders—those individuals who, or those programs or institutions that, will be affected by the uses of the tests—Bachman and Palmer (2010) describe *assessment justification* as the process of providing a rationale and evidence to justify the use of a particular assessment.

Assessment justification includes both a rationale for the assessment and evidence to support this rationale. At the heart of assessment justification is what they call an "assessment use argument" (AUA). Drawing on argument-based approaches to validity in educational measurement (e.g., Kane, 2001; Mislevy, Steinberg, & Almond, 2002), Bachman and Palmer (2010) describe an AUA as a conceptual framework for linking inferences from assessment performance to interpretation and use. An AUA explicitly states the interpretations and decisions that are to be based on assessment performance, as well as the consequences of using an assessment and of the decisions that are made. Bachman and Palmer argue that an AUA provides an overarching inferential framework to guide the design and development of language assessments and the interpretation and use of language assessment results. An AUA consists of a series of claims that can be illustrated as in Figure 94.1.

The arrows between the rectangles go both ways to illustrate that the claims, which may also be stated as questions, serve as a guide both for test development

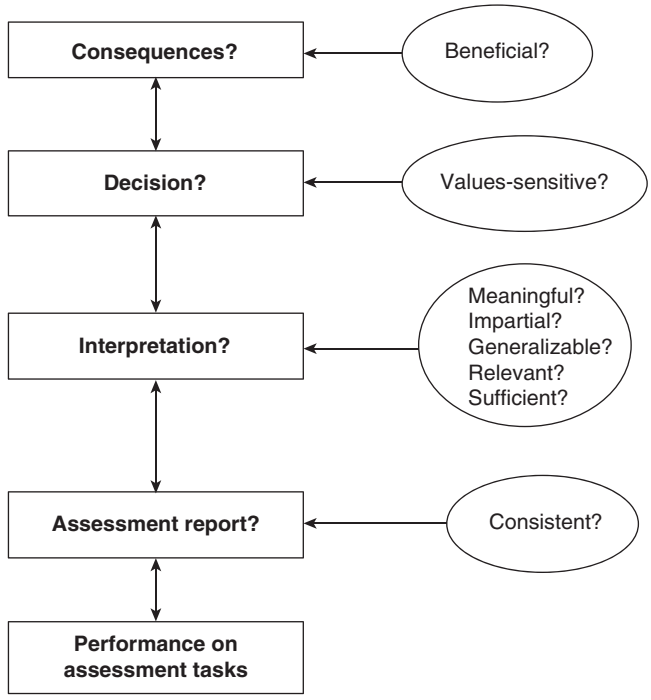

Figure 94.1 Assessment use argument (after Bachman & Palmer, 2010, p. 104) © Oxford University Press

and for the interpretation and use of assessment results. In using an AUA for designing and developing an assessment, the developer would first ask what the consequences of using the assessment might be and to what extent they would be *beneficial* to stakeholders. Then she would consider the decisions to be made and whether these are *sensitive* to existing societal values[1] and *equitable* toward different groups of stakeholders. Then she would consider the interpretations that are needed to make the intended decisions and the extent to which these interpretations will be *meaningful* with respect to a general theory of language ability, a needs or task analysis of a language use setting, or a particular learning syllabus: *impartial* to all groups of test takers; *generalizable* to the intended target language use domain; *relevant* to the decision to be made; and *sufficient* for the decision to be made. Finally the test developer would consider how to assure stakeholders that the assessment results (i.e., scores or descriptions) are *consistent* across different aspects of the measurement procedure (e.g., items, tasks, raters, forms).

In interpreting test takers' performance on an assessment, the assessment user would consider the inferences that are based on this performance. She would consider the consistency of the assessment report, the meaningfulness, impartiality, generalizability, relevance, and sufficiency of the interpretation, the values-sensitivity and impartiality of the decisions, and the beneficence of the consequences.

While the claims of an AUA constitute the conceptualization that is needed either to design an assessment or to interpret and use the results of an assessment, these claims need to be supported in order to justify using the assessment for a particular purpose. This support is provided in the form of warrants, which are propositions we use to justify the inference from one claim to the next (Bachman, 2005, p. 10). A warrant to support an inference from a score to an interpretation, for example, might be that the ratings derived from observing test takers' performance are consistent both across different raters and across multiple ratings by the same rater. Warrants supporting an inference from an interpretation to a decision might consist, for example, of the following:

- relevant legal requirements and existing community values are carefully considered in the decisions that are made (values-sensitivity warrant);
- stakeholders who are at equivalent levels on the construct to be assessed, as indicated by the interpretations of their assessment reports, have equivalent chances of being classified in the same group (equitability warrant).

Warrants, in turn, must be supported by backing, which comprises evidence from empirical research, documentation, regulations, laws, and community or societal values. The backing for the consistency of ratings, for example, might include classical inter- and intra-rater reliability estimates or variance components and dependability estimates from a generalizability study. The backing for the warrants of values and equitability, for example, might consist of:

- laws, regulations, policy, surveys of and focus group meetings with stakeholders;
- decision rules described in the assessment specifications; standard-setting procedures for setting cut scores; studies of the relationship between assessment performance and classification decisions.

Since it is the use of a *specific* assessment that needs to be justified, assessment justification is inherently a local process. Thus the AUA for a particular assessment provides a "local theory" that makes explicit claims about the roles of consequences, decisions, interpretations, and assessment reports in the assessment and identifies the evidence that needs to be collected to support these claims. The purpose of an AUA is thus *not* to falsify some general theory of language ability or a particular approach to designing language tests. Rather the purpose is to provide for, and to support empirically, a coherent argument capable of convincing the stakeholders that using the assessment will help promote the intended beneficial consequences.

The AUA also identifies appropriate methodologies for collecting evidence and thus embraces a multiplicity of methodological approaches, both "quantitative" and "qualitative."

Bachman and Palmer argue that the process of assessment justification, including the articulation of an assessment use argument, offers a conceptual framework for guiding both the development and the use of language assessments. It is this process that enables test developers and decision makers to be held accountable for the uses for which the assessments are intended. The authors further argue

that the process is applicable to a wide range of situations, from large-scale standardized tests to classroom assessments, and to a wide variety of purposes, from high stakes summative decisions about certification, entrance, and selection to low stakes formative decisions about improving teaching and learning.

Conclusion

The immediate and the long-term prospects for language testing (considered as a field) are filled with opportunities and challenges. I believe that the greatest challenges language assessment as a field faces are *not* in the cerebral spheres of validity theory, sociopsychological theory, postmodern critical social theory, or moral philosophy. Nor are they to be found in sophisticated statistical and measurement models or in ever refined approaches to naturalistic observation. Rather the challenges that we, as language testers, face are in the "real-world" arenas where language tests are being used to make decisions about individuals and institutions.

Turning these challenges into accomplishments will depend upon the willingness and capability of language testers to apply the knowledge and skills acquired over the past half century to the urgent practical assessment needs of our education system—from kindergarten to university and adult school—and of our society. It will also depend upon our willingness to leave the comfortable confines of the academy and join our colleagues in education and measurement to toil in the fields of practice. I believe that language testers have a unique combination of knowledge and skills, as well as a growing understanding of the issues involved in addressing the validity of interpretations and the consequences of test use. If we can but apply this expertise to the practical problems of assessment in our education systems and in our society, we are in a position to provide leadership and to contribute greatly to making our meritocracy fair and equitable.

SEE ALSO: Chapter 22, Language Testing for Immigration to Europe; Chapter 23, Language Testing for Immigration and Citizenship in the Netherlands; Chapter 41, Dynamic Assessment in the Classroom; Chapter 68, Consequences, Impact, and Washback; Chapter 89, Classroom-Based Assessment Issues for Language Teacher Education

Note

1 One of the thrusts of critical applied linguistics, as well as so-called "critical language testing" is that existing community values may themselves be inequitable and hence need to be constantly scrutinized, particularly by those who will be affected by the decisions that are made.

References

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, *2*(2), 192–204.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*, 115–29.

Alderson, J. C., & Wall, D. (Eds.). (1996). *Washback* (Special issue). *Language Testing*, *13*(3).

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34.

Bachman, L. F. (2006a). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165–207). Dordrecht, Netherlands: John Benjamins.

Bachman, L. F. (2006b, April). *Linking interpretation and use in educational assessments*. Paper presented at the National Council for Measurement in Education, San Francisco.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and context in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayless (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41–72). Ottawa, Canada: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Bailey, K. M. (1999). *Washback in language testing* (*TOEFL monograph series*). Princeton, NJ: Educational Testing Service.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Blackledge, A. (2009). "As a country we do expect": The further expansion of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, *6*(1), 6–16.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes. *Language Testing*, *15*, 45–85.

Brindley, G. (Ed.). (2007). Special issue on language assessment in schools. *Language Assessment Quarterly*, *4*(1).

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, *25*(3), 349–83.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research*. Rowley, MA: Newbury House.

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning*, *45*(2), 251–81.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*(4), 369–83.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York, NY: Cambridge University Press.

Chapelle, C. A. (2006). L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment. In M. Chalhoub-Deville, C. A. Chapelle, & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 47–64). Dordrecht, Netherlands: John Benjamins.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education*, *11*(1), 38–54.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, *18*(2), 133–48.

Davison, C. (2004). The contradictory culture of teacher-based assessment of written work of recently arrived immigrant ESL students. *Language Testing*, *21*(3), 305–34.

Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, *4*(4), 37–68.

de Jong, J. H. A. L., Lennig, M., Kerkhoff, A., & Poelmans, P. (2009). Development of a test of spoken Dutch for prospective immigrants. *Language Assessment Quarterly*, *6*(1), 41–60.

Douglas, D. (1997). Language for specific purpose testing. In C. Clapham & D. Cordon (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (pp. 111–19). Dordrecht, Netherlands: Kluwer Academic Publishers.

Douglas, D. (2000). *Assessing language for specific purposes: Theory and practice*. Cambridge, England: Cambridge University Press.

Eades, D. (2009). Testing the claims of asylum seekers: The role of language analysis. *Language Assessment Quarterly*, *6*(1), 30–40.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, *14*(3), 261–77.

Extra, G., Spotti, M., & Avermaet, P. V. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. London, England: Continuum International Publishing.

Hogan-Brun, G., Mar-Molinero, C., & Stevenson, P. (Eds.). (2009). *Discourse on langauge and integration: Critical perspectives on language testing regimes in Europe*. Amsterdam, Netherlands: John Benjamins.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–42.

Kunnan, A. J. (2000a). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, England: Cambridge University Press.

Kunnan, A. J. (Ed.). (2000b). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando*. Cambridge, England: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, England: Cambridge University Press.

Kunnan, A. J. (2009). Testing for citizenship: The US naturalization test. *Language Assessment Quarterly*, *6*(1), 89–97.

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, *1*(1), 49–72.

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research*, *15*(1), 11–33.

Lee, M. (2011). Is multiculturalism a poison to national identity? Looking behind the facade of language testing regimes. *Language Assessment Quarterly*, *8*(1), 92–8.

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice and change. *Language Assessment Quarterly*, *1*(1), 5–18.

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing*, *21*(3), 335–59.

Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Testing*, *4*(1), 6–36.

Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (Vol. 33). Cambridge, England: University of Cambridge ESOL Examinations/Cambridge University Press.

McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–66.

McNamara, T. F. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics*, *18*, 304–19.

McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, *18*(4), 333–49.

McNamara, T. F. (2003). Looking back, looking forward: Rethinking Bachman. *Language Testing*, *20*(4), 466–73.

McNamara, T. F., & Roever, K. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–96.

Oller, J. W., Jr. (1979). *Language tests at school*. London, England: Longman.

Oller, J. W., Jr. (1983). A consensus for the eighties? In J. W. Oller (Ed.), *Issues in language testing research* (pp. 351–6.). Rowley, MA: Newbury House.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26–56.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, *25*(2), 237–72.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability groups: A structural equation modelling approach. *Language Testing*, *15*(3), 333–79.

Rea-Dickins, P. (Ed.). (2000). *Assessing young learners* (Special issue). *Language Testing*, *1*.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, *18*(4), 393–407.

Rea-Dickins, P. (Ed.). (2004). *Exploring diversity in teacher assessment* (Special issue). *Language Testing*, *21*(3).

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London, England: Pearson.

Shohamy, E. (2009). Language tests for immigrants: Why language? Why tests? Why citizenship? In G. Hogan-Brun, C. Mar-Molinero & P. Stevenson (Eds.), *Discourses on langauge and integration: Critical perspectives on language testing regimes in Europe* (pp. 61–82). Amsterdam, Netherlands: John Benjamins.

Shohamy, E., & McNamara, T. (Eds.). (2009a). *Language tests for citizenship, immigration, and asylum* (Special issue). *Language Assessment Quarterly*, *6*(1).

Shohamy, E., & McNamara, T. (2009b). Editorial. In E. Shohamy & T. McNamara (Eds.), *Language tests for citizenship, immigration, and asylum* (Special issue). *Language Assessment Quarterly*, *6*(1), 1–5.

Stansfield, C. W. (1993). Ethics, standards and professionalism in language testing. *Issues in Applied Linguistics*, *4*(2), 15–30.

United States Congress. (2001). *H.R. 1, No Child Left Behind Act of 2001*.

United States Congress. (2002). *Public Law 107–110, No Child Left Behind Act of 2001*.

US Department of Education, US Department of State, US Department of Defense, & Office of the Director of National Intelligence. (2006). National security language initiative. Retrieved March 17, 2013 from http://www.ed.gov/about/inits/ed/competitiveness/nsli/nsli.pdf

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge, England: University of Cambridge ESOL Examinations and Cambridge University Press.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, *10*(1), 41–69.

Suggested Readings

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1–42.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, *13*(3), 257–79.

Cheng, L. (2004). *Changing language teaching through language testng: A washback study*. Cambridge, England: University of Cambridge, ESOL Examinations/Cambridge University Press.

Davies, A. (Ed.). (2004). *The ethics of language assessment* (Special issue). *Language Assessment Quarterly*, *1*(2–3).

Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, *4*(4), 37–68.

Llosa, L. (2008). Validating a standards-based classroom assessment of English proficiency: A multitrait–multimethod approach. *Language Testing*, *24*(4), 489–515.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–56.

Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, *9*(3), 233–65.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? *Language Testing*, *14*, 340–9.