Annual Review of Applied Linguistics (1999) **19**, 254–272. Printed in the USA.
Copyright © 1999 Cambridge University Press 0267-1905/99 $9.50

VALIDITY IN LANGUAGE ASSESSMENT

Carol A. Chapelle

INTRODUCTION

All previous papers on language assessment in the *Annual Review of Applied Linguistics* make explicit reference to *validity*. These reviews, like other work on language testing, use the term to refer to the quality or acceptability of a test. Beneath the apparent stability and clarity of the term, however, its meaning and scope have shifted over the past years. Given the significance of changes in the conception of validity, the time is ideal to probe its meaning for language assessment.

The definition of validity affects all language test users because accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation. In other words, assumptions about validity and the process of validation underlie assertions about the value of a particular type of test (e.g., "integrative," "discrete," or "performance"). Researchers in educational measurement (Linn, Baker and Dunbar 1991) have argued that some validation methods—particularly those relying on correlations among tests—are stacked against tests in which students are asked to display complex, integrated abilities (such as one might see in an oral interview) while favoring tests of discrete knowledge (such as what is called for on a multiple choice test of grammar). The Linn, *et al.* review, as well as other papers in educational measurement and language testing over the past decade, has stressed that if new test methods are to succeed, it is necessary to rewrite the rules for evaluating those tests (i.e., the methods of validation).

Exactly how validation should be recast is an ongoing debate, but it is possible to identify some directions. In describing them, one might discuss diverging philosophical bases in education, demographic changes in test takers, and advances in the statistical, analytic and technological methods for testing, all of which have provided some impetus for change. However, given the limitations of

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

space, this paper focuses most specifically on explaining the emerging view of validation that is likely to continue to impact research and practice in language assessment for the foreseeable future. An understanding of current work requires knowledge of earlier conceptions of validity, so a historical perspective is presented first along with a summary of contrasts between past and current views. Procedures for validation are then described and challenges facing this perspective are identified.

A HISTORY OF VALIDATION IN LANGUAGE TESTING

The term validity has been defined explicitly in texts on language testing and exemplified through language testing research. In Robert Lado's (1961) classic volume, *Language testing*, validity is defined as follows: "Does a test measure what it is supposed to measure? If it does, it is valid" (Lado, 1961:321). In other words, Lado portrayed validity as a characteristic of a language test—as an all-or-nothing attribute. Validity was seen as one of two important qualities of language tests; the other, reliability (i.e., consistency), was seen as distinct from validity, but most language testing researchers at that time agreed that reliability was a prerequisite for validity. In Oller's (1979) text, for example, validity is defined partly in terms of reliability: "...the ultimate criterion for the validity of language tests is the extent to which they *reliably* assess the ability of examinees to process discourse" (Oller 1979:406; emphasis added). Proponents of this view tended to equate validity with correlation. In other words, the typical empirical method for demonstrating validity of a test was to show "...that the test is valid in the sense of correlating with other [valid and reliable language tests] (Oller 1979:417–418). The language and methods of the papers in Palmer and Spolsky's (1975) volume on language testing reflect these perspectives.

In practice, correlational methods were seen as central to validation, and yet the "criterion-related validity" investigated through correlations was considered as only one type of validity. The other "validities" were defined as content-related validity, consisting of expert judgement about test content, and construct validity, showing results from empirical research consistent with theory-based expectations. In the 1970s, teachers and graduate students taking a course in educational measurement would learn about the three validities, but choosing and implementing validation methods was associated with large-scale research and development (e.g., proficiency testing for decisions about employment and academic admissions). This view is evident in Spolsky's (1975) paper pointing out that for classroom tests "the problem [of validation] is not serious, for the textbook or syllabus writer has already specified what should be tested" (Spolsky 1975:153). Large-scale research and development in language testing in the United States tended to stick to the notions of reliability as prerequisite for validity and validity through correlations. At the end of the 1970s, however, the tide began to turn when language testers started to probe questions about construct validation for tests of communicative competence (Palmer, Groot and Trosper 1981).

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

The language testing research in the 1980s continued the trend that began with the papers in the Palmer, *et al.* (1981) volume. Early issues of the journal, *Language Testing*, for example, reported a variety of methods for investigating score meaning, such as gathering data on strategies used during test taking (Cohen 1984), comparing test methods (Shohamy 1984), and identifying bias through item analysis (Chen and Henning 1985). Researchers were helping to clarify the hypothesis-testing process of validation through explicit prediction and testing based on construct theory (Bachman 1982, Klein-Braley 1985). At the same time, new performance tests were appearing which would challenge views about reliability and validity of the previous decade (Wesche 1987). The textbooks of the 1980's also expanded somewhat on the earlier trio of validities. Henning (1987) identified five types of validity by adding "response validity"—the extent to which examinees respond in an appropriate manner to test tasks—and by dividing criterion-related validity into concurrent and predictive (depending on the timing of the criterion measure). Henning also described several methods for investigating construct validity and stressed that "a test may be valid for some purposes but not for others" (1987:89). Madsen (1983) identified validity and reliability in traditional ways but added *affect*—the extent to which the test causes undue anxiety—as a third test quality of concern. Hughes (1989) introduced the three validities but added *washback*—the effect of the test on the process of teaching and learning—as an additional quality. Canale's (1987) review of language testing in the *Annual Review of Applied Linguistics* included discussion of issues typically related to validity (i.e., what to test, and how to test), but included with equal status discussion of the *ethics* of language testing (i.e., why to test).

In all, the 1980s saw language testers discussing qualities of tests with greater sophistication than in the previous decade and using a wider range of analytic tools for research. However, with the exception of a few papers arguing against equating "authenticity" with "validity" (e.g., Stevenson 1985), and one suggesting the use of methods from cognitive psychology for validation (Grotjahn 1986), little explicit discussion of validity itself appeared in the 1980s. In educational measurement, in contrast, the definition and scope of validity was certainly under discussion (e.g., Anastasi 1986, Angoff 1988, Cronbach 1988, Landy 1986). Three important developments resulted. First, the 1985 *AERA/APA/NCME standards for educational and psychological testing*[1] replaced the former definition of three validities with a single unified view of validity, one which portrays construct validity as central. Content and correlational analyses were presented as methods for investigating construct validity. Second, the philosophical underpinnings of the validation process began to be probed (Cherryholmes 1988) from perspectives that would expand through the next decade (Moss 1992; 1994, Wiggins 1993).

The third event was the publication of Messick's seminal paper, "Validity," in the third edition of the *Handbook of educational measurement* (Messick 1989). It underscored the previous two points and articulated a definition of validity which incorporated not only the types of research associated with

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

construct validity but also test consequences—for example, the concerns about affect raised by Madsen, washback as described by Hughes, and ethics brought up by Canale. The notion that validation should take into account the consequences of test use had historical roots in educational measurement (Shepard 1997), but the idea was taken seriously enough to cause widespread debate for the first time as a result of Messick's (1989) paper.[2]

Douglas' (1995) paper in *Annual Review of Applied Linguistics* refers to 1990 as a "watershed in language testing" because of the language testing conferences held, the movement toward establishing the International Language Testing Association, the formation of LTEST-L on internet, and the publication of several books on language testing. In addition to, and perhaps because of these developments, 1990 also marked the beginning of a decade of explicit discussion on the nature of validity in language assessment. Among the first items on the agenda for the International Language Testing Association was a project to identify international standards for language testing—a project that inevitably directed attention to validation (Davidson, Turner, and Huhta 1997). LTEST-L during the 1990s has regularly served as a forum for conversation about validity—a conversation which frequently points beyond the language testing literature into educational measurement, and therefore broadens the intellectual basis for redefining validity in language assessment.

The most influential mark of the 1990s was Bachman's (1990a) chapter on validity which he framed in terms of the *AERA/APA/NCME Standards* (1985) and Messick's (1989) paper. Bachman introduced validity as a unitary concept pertaining to test interpretation and use, emphasizing that the inferences made on the basis of test scores, and their uses are the object of validation rather than the tests themselves. Construct validity is the overarching validity concept, while content and criterion-related (correlational) investigations can be used to investigate construct validity. Following Messick, he included the consequences of test use rather than only "what the test measures" within the scope of validity. Bachman presented validation as a process through which a variety of evidence about test interpretation and use is produced; such evidence can include but is not limited to various forms of reliabilities and correlations with other tests.

Throughout the 1990s, other work in language testing has also adopted Messick's perspective on validity (Chapelle 1994; forthcoming a, Chapelle and Douglas 1993, Cumming 1996, Kunnan 1997; 1998, Lussier and Turner 1995). The consequential aspects of validity, including washback and social responsibility, have been discussed regularly in the language testing literature (e.g., Davies 1997). Recently, a "meta-analysis" was conducted to probe conceptions of validity more explicitly by analyzing the philosophical perspectives toward validation apparent in research reported throughout the history of the Language Testing Research Colloquium (Hamp-Lyons and Lynch 1998). In short, language testers are adopting, adapting, and contributing to validity perspectives in educational

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

measurement. Table 1 summarizes key changes in the way that validation was and is conceptualized.

Table 1. Summary of contrasts between past and current conceptions of validation

Past	Current
Validity was considered a *characteristic of a test*: the extent to which a test measures what it is supposed to measure.[3]	Validity is considered an *argument* concerning test interpretation and use: the extent to which test interpretations and uses can be justified.
Reliability was seen as distinct from and a necessary *condition for validity*.	Reliability can be seen as *one type of validity evidence*.
Validity was often established through *correlations* of a test with other tests.	Validity is argued on the basis of a number of types of *rationales and evidence*, including the consequences of testing.
Construct validity was seen as one of *three types of validity* (the three validities were content, criterion-related, and construct).	Validity is a *unitary concept* with construct validity as central (content and criterion-related evidence can be used as evidence about construct validity).
Establishing validity was considered within the purview of *testing researchers* responsible for developing large-scale, high-stakes tests.	Justifying the validity of test use is the responsibility of *all test users*.

CURRENT APPROACHES TO VALIDATION IN LANGUAGE TESTING

Messick's seminal paper explained validity and the process of validation through the use of what has become a widely cited "progressive matrix" (approximated in Figure 1) intended to portray validity as a unitary but multifaceted concept. The column labels (inferences and uses) represent the outcomes of testing. In other words, testing results in inferences being made about test-takers abilities, knowledge, or performance, for example, and in decisions being made such as whether to teach "apologies" again, whether to admit the test taker to college, or whether to hire the test taker for a job. The row labels (evidence and consequences) refer to the types of arguments that should be used to justify testing outcomes. The matrix is progressive because each of the cells contains "construct validity" but adds on an additional facet.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

	Inferences	Uses
Evidence	Construct validity	Construct validity + Relevance/utility
Consequences	Construct validity + Value implications	Construct validity + Value implications + Relevance/utility + Social consequences

Figure 1. Progressive matrix for defining the facets of validity (adapted from Messick 1989:20)

Building on this conceptual definition, Messick went on to identify particular types of evidence and consequences that can be used in a validity argument. In short, this work encompasses guidelines for how evidence can be produced—in other words, what constitutes methods for test validation. Validation begins with a hypothesis about the appropriateness of testing outcomes (i.e., inferences and uses). Data pertaining to the hypothesis are gathered and results are organized into an argument from which a "validity conclusion" (Shepard 1997:6) can be drawn about the validity of testing outcomes.

1. Hypotheses about testing outcomes

In educational measurement, construct validation has been framed in terms of hypothesis testing for some time (Cronbach and Meehl 1955, Kane 1992, Landy 1986). Hypotheses about language tests refer to assumptions about what a test measures (i.e., the inferences drawn from test scores) and what their scores can be used for (i.e., decisions based on test scores).

> Inferences and the validation of inferences is hypothesis testing. However, it is not hypothesis testing in isolation but, rather, theory testing more broadly because the source, meaning, and import of score based hypotheses derive from the interpretive theories of score meaning in which these hypotheses are rooted (Messick 1989:14).

For example, in her study of the IELTS, Clapham (1996) hypothesized that subject area knowledge would work together with language ability during test performance, and therefore test performance could be used to infer subject-specific language ability. What follows from this hypothesis is that students who take a version of the test requiring them to work with language about their own subject areas will score better than those who take a test with language from a different subject area. The inference was that test performance would reflect subject-specific language ability, which would provide an appropriate basis for decisions about examinees' readiness for academic study. This hypothesis about test performance is derived from a theory of what is involved in responding to the test

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

questions, which requires a construct theory of subject-specific language ability. Hypotheses might also be developed from anticipated testing consequences, such as the robustness of decisions made about admissions to universities, or satisfaction test takers might be expected to feel as a result of taking a subject specific language test.

2. Relevant evidence for testing the hypotheses

Messick identified several distinct types of evidence that can come into play in validation; in other words, he outlined the methods that can be undertaken to investigate hypotheses:

> We can look at the content of a test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to items or tasks. We can examine relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is the test's external structure. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions—such as instructional or therapeutic treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also the unintended side effects (Messick 1989:16).

Examples of each of these strategies or approaches to validity evidence can be found in the language testing research of the 1990s. Six approaches to validity evidence are also discussed below.

The first approach, **content analysis**, consists of experts' judgments of what they believe a test measures—judgements about the "content relevance, representativeness, and technical quality" of the test material (Messick 1995:6). In other words, content analysis provides evidence for the hypothesized match between test items or tasks and the construct that the test is intended to measure. This approach to validation has evolved from the "content validity" of the 1970s; use of the content analysis in support of a content validity argument, however, underscores the need for an explicit construct definition to guide analysis. A number of studies illustrate approaches to and problems with content analysis of language tests (e.g., Alderson, 1993, Bachman, Kunnan, Vanniarajan and Lynch 1988). The most interesting issue that this type of analysis raises for language testing is the question of what should be analyzed as "test content." The accepted approach has been for expert raters to make judgements about the cognitive knowledge and processes they believed would be required for test performance (e.g., Carroll 1976); however, such an approach assumes that the construct is

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

defined in terms of knowledge and processes—an assumption which does not always hold in performance tests (McNamara 1996).

Empirical item or task analysis, a second approach, supplies evidence for the "substantive aspect" of construct validity (Messick 1995:6) by revealing the extent to which hypothesized knowledge and processes appear to be responsible for learners' performance. The analysis in this case is not judgmental but instead relies on empirical analysis of learners' responses. Quantitative analyses can investigate the extent to which relevant factors affect item difficulty and discrimination (Carroll 1989). An example of this approach is Kirsch and Mosenthal's (1988; 1990) construct validation of tests of "document literacy"—the ability to read documents to be able to do something. On the basis of their construct definition, they hypothesized particular variables would be related to task difficulty. Construct validity of the test is supported to the extent that these variables are significant predictors of test difficulty.

Qualitative analyses attempt to document the strategies and language that learners use as they complete test tasks. The hypothesis in these studies would be that the test taker is engaging in construct-relevant processes during test taking. A number of studies have been conducted to evaluate this type of hypothesis on tests of listening and reading, as well as cloze tests and C-tests (Buck 1991, Cohen forthcoming, Feldmann and Stemmer 1987, Yi'an 1998). Results tend to indicate that test takers rely more heavily on metacognitive problem-solving strategies than on the communicative strategies that one would hope would affect performance in a language test—a finding which fails to provide evidence for validity of inferences about communicative language strategies. Studies of learners' processes during test taking can also focus on the language produced by the test taker. In such cases, discourse analysis is used to compare the linguistic and pragmatic characteristics of the language that learners produce in a test with what is implied from the construct definition (Lazerton 1996).

A third approach, **Dimensionality analysis**, investigates the internal structure of the test by assessing the extent to which observed dimensionality of response data is consistent with the hypothesized dimensionality of a construct. Observed dimensionality is tested by estimating the fit of the test response data to a psychometric model which must correspond to the construct theory. When the psychometric model is unidimensional (Henning, Hudson and Turner 1985), there are several ways to investigate the data fit including classical true-score reliability methods and certain item response theory (IRT) methods (Bachman 1990a, Blais and Laurier 1995, Choi and Bachman 1992). The problem, which has been the source of much debate, is that many language tests are developed on the basis of multidimensional construct definitions. To the extent that the test user wants reliable score information about each aspect of the construct (e.g., pragmatic competence vs. grammatical competence), a multidimensional model is needed. Although multidimensional psychometric models are a topic of research (Ackerman

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

1994, Embretson 1985, Mislevy 1993; 1994), work in this area remains somewhat tentative.

The fourth type of evidence comes from investigation of **relationships** of test scores with other tests and behaviors. The hypotheses investigated in these validity studies specify the anticipated relationships of the test under investigation with other tests or quantifiable performances. An important paradigm for systematizing theoretical predictions of correlations is the multitrait-multimethod (MTMM) research design which has been used for language testing research (e.g., Bachman and Palmer 1982, Stevenson 1981, Swain 1990). The MTMM design specifies that tests of several different constructs are chosen so that each construct is measured using several different methods, and then evidence for validity is found if the correlations among the tests of the same construct are stronger than correlations among tests of different constructs. Hypotheses about the strengths of relationships (e.g., divergent and convergent correlations) among tests can be made on the basis of other theoretical criteria as well, such as content analyses of tests (Chapelle and Abraham 1990).

The fifth source of evidence is drawn from results of research on **differences** in test performance. Hypotheses are based on a theory of the construct which includes how it should behave differently across groups of test-takers, time, instruction, or test task characteristics. The study of how differences in test task characteristics influence performance is framed in terms of generalizability (Bachman 1997)—the study of the extent to which performance on one test task can be assumed to generalize to other tasks. This type of evidence has been particularly important as test developers attempt to design tests with fewer, but more complex test tasks (McNamara 1996). Hypotheses about bias resulting from language test tasks delivered on the computer can also be tested by comparing scores of test-takers with varying degrees of prior experience with computers (Taylor, Kirsch, Jamieson and Eignor in press).

The final type of argument cited as pertaining to validity are those arguments based upon **testing consequences**. Consequences refer to the value implications of the interpretations made from test scores and the social consequences of test use. Testing consequences present a different dimension for a validity argument than the other forms because they involve hypotheses and research directed beyond the test inferences to the ways in which the test impacts people involved with it. A recent study investigating consequences of the TOEFL on teaching in an intensive English program, for example, found that consequences of the TOEFL could be identified, but that they were mediated by other factors in the language program (Alderson and Hamp-Lyons 1996). The problem of investigating consequences of language tests is an important, current issue (Alderson and Wall 1993, Bailey 1996, Wall 1997).

Messick's conception of validity and the types of validity evidence outlined above have served well in providing a coherent introduction to research on

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

validation (e.g., Chapelle and Douglas 1993, Cumming 1996, Kunnan 1998). Their real purpose, however, is to guide validation research which integrates evidence from these approaches into a validity conclusion about one test.

3. Developing a validity argument

A validity argument should present and integrate evidence and rationales from which a validity conclusion can be drawn pertaining to particular score-based inferences and uses of a test. A study of a reading comprehension test (Anderson, Bachman, Perkins and Cohen 1991) illustrated how data might be integrated from three sources: content analysis, investigation of strategies, and quantitative item performance data. The results showed how particular strategies were linked to success on items with particular characteristics, but the qualitative, item level report of results also shows the difficulty in integrating detailed data into a validity conclusion. A second effort to develop a validity argument is illustrated by an attempt to organize existing data about a test method (the C-test) in order to draw a conclusion about particular test inferences and uses (Chapelle 1994). In this case, the relevant rationales are presented in a table to show arguments both for and against the validity of specific inferences. These are only two examples that demonstrate the difficulty in developing a validity argument that is sufficiently pointed to draw a single conclusion.

CURRENT CHALLENGES IN LANGUAGE TEST VALIDATION

The changes of the past decade have helped to make validation of language assessment among the most interesting and important areas within applied linguistics. Language assessment is critical in many facets of the field; current perspectives make the applied linguists who use tests responsible for justifying the validity of their use. This responsibility invites all test users to share with language testing researchers the challenges of defining language constructs and developing validity arguments in order to apply validation theory to testing practice.

1. Defining the language construct to be measured

Each of the past reviews of language testing in *ARAL* has named as significant the issue of how best to define what a test is intended to measure (e.g., Bachman 1990b, Canale 1987, Douglas 1995). This problem is no less central to discussions of validation in 1999 than it was to each of the broader overviews in previous volumes. Construct validation, which is central to all validation, requires a construct theory upon which hypotheses can be developed and against which evidence can be evaluated. Progress has been made in recent years through clarification of different theoretical approaches toward construct definition (Chapelle forthcoming, Skehan 1998) and links between construct definition and language test use (Bachman and Palmer 1996). While work remains to be done on how approaches to construct definition might best be matched with test purposes, the

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

biggest problem—regardless of the approach to construct definition—is the level of detail to be included. Some of the validation research described above requires precise hypotheses and can yield detailed data about the specifics of test content and performance. For example, results from empirical task analysis can reveal very specific processes that learners use. And yet, a construct theory that is too detailed, or too oriented toward processing, risks losing its usefulness as a meaningful interpretation of performance (Chapelle forthcoming b).

2. Developing a validity argument for a particular test use

The challenge of developing a validity argument begins with the difficulties in settling on a construct definition, but additional complications arise in identifying the appropriate types and number of justifications as well as in integrating them to draw a validity conclusion. The process of validation costs time and money, so despite the fact that theoretically one can consider it an on-going process, practically speaking, a test user has to make a decision about the results that are essential to justify a particular test use. Davies (1990) introduced discussion of the relative strength of different approaches to validity and the need to combine validity evidence in order to support hypotheses, but it is not clear how generally these ideas can be applied given the context-specific nature of test use. Shepard (1993) suggests that test use serve as a guide to the selection and interpretation of validity evidence, making validity arguments vary from one situation to another. Despite these suggestions, in the end, a validity conclusion is an argument-based, context-specific judgement, rather than a proof-based, categorical result.

3. Applying validation theory to language testing practice

The largest challenge for validation in language testing is to adapt current understanding of validity from the measurement literature into practices in second language classes, programs, and research. The view of validity presented here may be clearer than it was in the past and particular aspects have been amplified, but the basic tenets (e.g., that validity refers to test interpretation and use rather than to tests) have been present in the educational measurement literature for decades. However, researchers in educational measurement are seldom the ones in the position to construct language tests for classrooms, analyze placement tests for language programs, or propose measures for SLA research. Validation theory stresses the responsibility of test users to justify validity for whatever their specific test uses might be, and therefore it underscores the need for comprehensible procedures and education for test users. Bachman and Palmer's (1996) book, *Language testing in practice,* illustrates one way in which this challenge is beginning to be addressed. They substitute "usefulness" for "validity of score-based inferences and uses" and outline how test developers can maximize usefulness through specific measures taken in test development.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

CONCLUSION

For those who have followed work in validation of language assessment, there is no question that real progress has been made, moving beyond Lado's conception that validity is whether or not a test measures what it is supposed to. This progress promises more thoughtfully designed and investigated language tests in addition to more thoughtful and investigative test users. Based on discussions in the educational measurement literature, one can expect the *AERA/APA/NCME Standards* currently under revision to define validity in a manner similar to what is explained here. Based on discussions in the language testing literature, language testing researchers can be expected to be more closely allied with these views than ever before. As a consequence, for applied linguists who think that "the validity of a language test" is its correlation with another language test, now is a good time to reconsider.

NOTES

1. The *AERA/APA/NCME standards for educational and psychological testing* is the official code of professional practice in the US. The acronyms stand for American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, respectively. A new edition of the code has appeared approximately each decade since the 1950s (1954, 1966, 1974, 1985). The next edition is in preparation.

2. The key issue now on the table is how validity should be portrayed in the next version of the *AERA/APA/NCME Standards* which will appear soon (Messick 1994, Moss 1992; 1994, Shepard 1993, *Educational measurement: Issues and practice* 1997).

3. The idea that validity is a characteristic of a test has not been held by orthodox educational measurement researchers for some time, if ever. Cronbach and Meehl's (1955) paper, intended to amplify and explain some of the ideas presented in the first edition of the *Standards*, clearly stated "One does not validate a test, but only a principle for making inferences" (1955:297). Somehow the expression "test validity" (which is short for "validity of inferences and uses of a test") came to denote that tests themselves can be valid or invalid.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

ANNOTATED BIBLIOGRAPHY

Bachman, L. F. and A. S. Palmer. 1996. *Language testing in practice.* Oxford: Oxford University Press.

This book takes readers through an in-depth discussion of test development and formative evaluation—detailing each step of the way in view of the theoretical and practical concerns that should inform decisions. The book contributes substantively to current discussions of validity by proposing a means for evaluating language tests which incorporates current validation theory but which is framed in a manner that is sufficiently comprehensible and appropriately slanted toward language testing. This "framework for test usefulness" acts as the centerpiece of the book, which builds the concepts and procedures intended to help readers develop language tests that are useful for particular situations. The authors' choice of "usefulness" rather than "validity" succeeds in keeping in the forefront the critical idea that tests must be evaluated in view of the contexts for which they are intended.

Chapelle, C. A. Forthcoming a. Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.) *Second language acquisition and language testing interfaces.* Cambridge: Cambridge University Press.

Focusing on the significance of construct definition in the process of validation, this paper outlines three ways of defining a construct and explains the implication of one of these perspectives for framing validation studies. The three perspectives on constructs—trait, behaviorist, and interactionalist—are illustrated through definitions of vocabulary ability. Validation is discussed in terms of implications of the interactionalist definition for construct validity, relevance and utility, value implications, and social consequences.

Clapham, C. and D. Corson (eds.) 1997. *Encyclopedia of language and education. Volume 7. Language testing and assessment.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

This volume is a well-planned collection of brief papers from experts in various areas of language testing. Although it does not include a chapter on validation as a concept, it contains good introductions to construct and consequential forms of validation arguments. Relevant chapters include topics such as advances in quantitative test analysis, latent trait models, generalizability theory, qualitative approaches, washback, standards, accountability, and ethics.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

Cumming, A. 1996. Introduction: The concept of validation in language testing. In A. Cumming and R. Berwick (eds.) *Validation in language testing*. Clevedon, Avon: Multilingual Matters. 1–14.

This paper introduces the published papers of the Fourteenth Annual Language Testing Research Colloquium (1992) by reviewing approaches that have been taken toward validity and placing each paper in the volume into Messick's framework. In other words, it points out papers that the author sees as illustrations of both evidential and consequential approaches to justifying validity of test inference and use.

Educational Measurement: Issues and Practice. 1997. 16.2. [Special issue on validity.]

The first four articles in this issue provide a succinct, up-to-date sample of current debates about the ideal scope for validity. Two papers, those by Lorrie Shepard and Robert Linn, argue that social consequences should be considered within a validity framework, and that this perspective represents an evolution and clarification of prior statements about validity. James Popham and William Mehrens each portray the inclusion of social consequences as a threat to the clarity of the notion of validity as a characteristic of score-based inferences.

Hamp-Lyons, L. and B. Lynch. 1998. Perspectives on validity: A historical analysis of language testing conferences. In A. Kunnen (ed.) *Validation in language assessment*. Mahwah, NJ: L. Erlbaum. 253–277.

Unique in the language testing literature, this paper discusses philosophical approaches associated with perspectives on validity, distinguishing broadly between those working within a "positivistic-psychometric" paradigm from those who work in a "naturalistic-alternative" paradigm. They associate the work of Messick (as described in this paper) with the former and Moss (e.g., Moss 1992; 1994) with the latter. The authors attempt to classify the paradigms within which papers at the Language Testing Research Colloquium appear to have conducted their research, and they identify language in the abstracts for papers that signal the authors' perspectives on validity. They conclude that, while some shifts in treatment of validity have occurred, the dominant paradigm at LTRC remains positivistic-psychometric.

Kunnan, A. J. 1998. Approaches to validation in language assessment. In A. Kunnan (ed.) *Validation in language assessment*. Mahwah, NJ: L. Erlbaum. 1–16.

This paper introduces the published papers of the Seventeenth Annual Language Testing Research Colloquium (1995) with a brief historical view

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

of validity, an explanation of Messick's framework, and extensive examples of research that the author sees as illustrating evidential and consequential approaches to justifying validity of test inference and use. Papers in the volume are also placed within Messick's progressive matrix to show their orientation.

Messick, S. 1989. Validity. In R. L. Linn (ed.) *Educational measurement.* 3rd ed. New York: Macmillan. 13–103.

This is the seminal paper on validity. It presents the author's definition of validity as a multifaceted concept and describes the implications of the definition for the study of validation. Grounded in the history of educational measurement and philosophy of science, this presentation has had an impact on work in educational and psychological measurement as well as in language testing.

UNANNOTATED BIBLIOGRAPHY

Ackerman, T. 1994. Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement.* 18.257–275.

AERA/APA/NCME. 1985. *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Alderson, J. C. 1993. Judgements in language testing. In D. Douglas and C. Chapelle (eds.) *A new decade of language testing research.* Alexandria, VA: TESOL. 46–57.

_____ and L. Hamp-Lyons. 1996. TOEFL preparation courses: A study of washback. *Language Testing.* 13.280–297.

_____ and D. Wall. 1993. Does washback exist? *Applied Linguistics.* 14.115–129.

Anastasi, A. 1986. Evolving concepts of test validation. *Annual Review of Psychology.* 37.1–15.

Anderson, N. J., L. Bachman, K. Perkins and A. Cohen. 1991. An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing.* 8.41-66.

Angoff, W. H. 1988. Validity: An evolving concept. In H. Wainer and H. Braun (eds.) *Test validity.* Hillsdale, NJ: L. Erlbaum. 19–32.

Bachman, L. F. 1982. The trait structure of cloze test scores. *TESOL Quarterly.* 16.61–70.

_____ 1990a. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

_____ 1990b. Assessment and evaluation. In R. B. Kaplan, *et al.* (eds.) *Annual Review of Applied Linguistics, 10.* New York: Cambridge University Press. 210–226.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

Bachman, L. F. 1997. Generalizability theory. In C. Clapham and D. Corson (eds.) *Encyclopedia of language and education. Volume 7. Language testing and assessment.* Dordrecht, The Netherlands: Kluwer Academic Publishers. 255–262.

_____, A. Kunnan, S. Vanniarajan and B. Lynch. 1988. Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency tests. *Language Testing.* 5.128–159.

_____ and A. S. Palmer. 1982. The construct validation of some components of communicative competence. *TESOL Quarterly.* 16.449-465.

Bailey, K. 1996. Working for washback: A review of the washback concept in language testing. *Language Testing.* 13.257–279.

Blais, J-G. and M. D. Laurier. 1995. The dimensionality of a placement test from several analytical perspectives. *Language Testing.* 12.72–98.

Buck, G. 1991. The testing of listening comprehension: An introspective study. *Language Testing.* 8.67-91.

Canale, M. 1987. The measurement of communicative competence. In R. B. Kaplan, *et al.* (eds.) *Annual Review of Applied Linguistics, 8.* New York: Cambridge University Press. 67-84.

Carroll. J. B. 1976. Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (ed.) *The nature of intelligence.* Hillsdale, NJ: L. Erlbaum. 27–56.

_____ 1989. Intellectual abilities and aptitudes. In A. Lesgold and R. Glaser (eds.) *Foundations for a psychology of education.* Hillsdale, NJ: L. Erlbaum. 137-197.

Chapelle, C. A. 1994. Is a C-test valid for L2 vocabulary research? *Second Language Research.* 10.157–187.

_____ Forthcoming b. From reading theory to testing practice. In M. Chalhoub-Deville (ed.) *Development and research in computer adaptive language testing.* Cambridge: Cambridge University Press. 145–161.

_____ and R. G. Abraham. 1990. Cloze method: What difference does it make? *Language Testing.* 7.121–146.

_____ and D. Douglas. 1993. Foundations and directions for a new decade of language testing research. In D. Douglas and C. Chapelle (eds.) *A new decade of language testing research.* Alexandria VA: TESOL. 1-22.

Chen, Z. and G. Henning. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing.* 2.155–163.

Cheryholmes, C. 1988. *Power and criticism: Poststructural investigations in education.* New York: Teachers College Press.

Choi, I-C. and L. F. Bachman. 1992. An investigation into the adequacy of three IRT models for data from two EFL reading rests. *Language Testing.* 9.51–78.

Clapham, C. 1996. *The development of the IELTS: A study of the effect of background knowledge on reading comprehension.* Cambridge: Cambridge University Press.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

Cohen, A. 1984. On taking language tests: What the students report. *Language Testing*. 1.70–81.

_____ Forthcoming. Strategies and processes in test-taking and SLA. In L. Bachman and A. Cohen (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Cronbach, L. J. 1988. Five perspectives on validation argument. In H. Wainer and H. Braun (eds.) *Test validity*. Hillsdale, NJ: L. Erlbaum. 3–17.

_____ and P. E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*. 52.281–302.

Davidson, F., C. E. Turner and A. Huhta. 1997. Language testing standards. In C. Clapham and D. Corson (eds.) *Encyclopedia of language and education. Volume 7. Language testing and assessment*. Dordrecht, The Netherlands: Kluwer Academic Publishers. 301–311

Davies, A. 1990. *Principles of language testing*. Oxford: Basil Blackwell.

_____ (ed.) 1997. *Ethics in language testing*. [Special issue of *Language Testing*. 14.3]

Douglas, D. 1995. Developments in language testing. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistic, 15. Survey of applied linguistics*. New York: Cambridge University Press. 167–187.

Embretson, S. (ed.) 1985. *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.

Feldmann, U. and B. Stemmer. 1987. Thin___ aloud a___ retrospective da___ in C-te___ taking: Diffe___ languages—diff___ learners—sa___ approaches? In C. Færch and G. Kasper (eds.) *Introspection in second language research*. Philadelphia, PA: Multilingual Matters. 251–267

Grotjahn, R. 1986. Test validation and cognitive psychology: Some methodological considerations. *Language Testing*. 3.159-185.

Henning, G. 1987. *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.

_____, T. Hudson and J. Turner. 1985. Item Response Theory and the assumption of unidimensionality. *Language Testing*. 2.141–154.

Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

Kane, M. T. 1992. An argument-based approach to validity. *Psychological Bulletin*. 112.527–535.

Kirsch, I. S. and P. B. Mosenthal. 1988. *Understanding document literacy: Variables underlying the performance of young adults*. Princeton, NJ: Educational Testing Service. [Report no. ETS RR–88–62.]

_____ 1990. Exploring document literacy: Variables underlying performance of young adults. *Reading Research Quarterly*. 25.5–30.

Klein-Braley, C. 1985. A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing*. 2.76–104.

Kunnan, A. J. 1997. Connecting fairness with validation in language assessment. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (eds.) *Current*

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

developments and alternatives in language assessment. Proceedings of LTRC 96. Jyväskylä, Finland: University of Jyväskylä. 85–105.

Lado, R. 1961. *Language testing: The construction and use of foreign language tests.* New York: McGraw-Hill.

Landy, F. J. 1986. Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist.* 41.1183–1192.

Lazerton, A. 1996. Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing.* 13.151–172.

Linn, R. L., E. L. Baker and S. B. Dunbar. 1991. Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher.* 20.2.15-21.

Lussier, D. and C. E. Turner. 1995. *Le point sur...L'évaluation en didactique des langues.* [*Focus on evaluation in language teaching.*] Anjou, Québec: Centre Éducatif et Culturel.

Madsen, H. S. 1983. *Techniques in testing.* Oxford: Oxford University Press.

McNamara, T. 1996. *Measuring second language performance.* London: Longman.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher.* 23.8.13–23.

_____ 1995. Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice.* 14.5–8.

Mislevy, R. J. 1993. Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy and I. I. Bejar (eds.) *Test theory for a new generation of tests.* Hillsdale, NJ: L. Erlbaum. 19–39.

_____ 1994. Evidence and inference in educational assessment. *Psychometrika.* 59.439–483.

Moss, P. A. 1992. Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research.* 62.229-258.

_____ 1994. Can there be validity without reliability? *Educational Researcher.* 23.8.5–12.

Oller, J. 1979. *Language tests at school.* London: Longman.

Palmer, A. S., P. J. M. Groot and G. A. Trosper (eds.) 1981. *The construct validation of tests of communicative competence.* Washington, DC: TESOL.

Palmer, L. and B. Spolsky (eds.) 1975. *Papers on language testing. (1967–1974).* Washington, DC: TESOL.

Shepard, L. 1993. Evaluating test validity. *Review of Research in Education.* 19.405–450.

_____ 1997. The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice.* 16.2.5–8, 13, 24.

Shohamy, E. 1984. Does the testing method make a difference? The case of reading comprehension. *Language Testing.* 1.147–170.

Skehan, P. 1998. *A cognitive approach to language learning.* Oxford: Oxford University Press.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press

Spolsky, B. 1975. Language testing—The problem of validation. In L. Palmer and B. Spolsky (eds.) *Papers on language testing (1967–1974)*. Washington, DC: TESOL. 146–153.

Stevenson, D. K. 1981. Beyond faith and face validity: The multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests. In A. S. Palmer, P. J. M. Groot and G. A. Trosper (eds.) *The construct validation of tests of communicative competence*. Washington, DC: TESOL. 37-61.

Stevenson, D. K. 1985. Authenticity, validity, and a tea party. *Language Testing*. 2.41–47.

Swain, M. 1990. Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? In J. Alatis (ed.) *Linguistics, language teaching and language acquisition*. Georgetown University Round Table. Washington, DC: Georgetown University Press. 401–412.

Taylor, C., I. Kirsch, J. Jamieson and D. Eignor. In press. Estimating the effects of computer familiarity on computer-based TOEFL tasks. *Language Learning*.

Wall, D. 1997. Impact and washback in language testing. In C. Clapham and D. Corson (eds.) *Encyclopedia of language and education. Volume 7. Language testing and assessment*. Dordrecht, The Netherlands: Kluwer Academic Publishers. 291–302.

Wesche, M. 1987. Second language performance testing: The Ontario test of ESL as an example. *Language Testing*. 4.28–47.

Wiggins, G. P. 1993. *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass Publishers.

Yi'an, W. 1998. What do tests of listening comprehension test?—Retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*. 15. 21–44.

https://doi.org/10.1017/S0267190599190135 Published online by Cambridge University Press