مجموعه کتابهای جامع تخصصی آزمون دکتری

آمـوزش زبان انگلیسـی

ISSUES IN LANGUAGE TESTING

Masoume Ahmadi
Naser Sabourian Zadeh

In The Name Of God

ISSUES IN LANGUAGE TESTING

Compiled by:

Masoume Ahmadi
Naser Sabourian Zadeh

Contents

chapter **2**

Uses of Language Tests

Introduction

The two major uses of language tests are:

1. as sources of information for making decisions within the context of educational programs; and
2. as indicators of abilities or attributes that are of interest in research on language, language acquisition, and language teaching.

In educational settings the major uses of test scores are related to evaluation, or making decisions about people or programs. In order to justify the use of tests for educational evaluation, certain assumptions and considerations must be made regarding the usefulness and quality of the information which tests provide. An understanding of these assumptions and considerations, as well as of the different types of language tests and their roles in evaluation is therefore essential to the appropriate use of language tests.

Uses of language tests in educational programs

The fundamental use of testing in an educational program is to provide information for making decisions, that is, for evaluation. Evaluation comprises essentially two components: (1) *information*, and (2) *value judgments, or decisions*. The information relevant to evaluation can be either qualitative (non-measurement) or quantitative (measurement). Qualitative information can be obtained from observations in a wide variety of ways, including performance checklists and observation schedules, as well as from narrative accounts of class performance of student self-reports. Quantitative information can include measures such as class rank, teacher ratings, and self-ratings as well as tests.

Assumptions and considerations

The use of tests as a source of evaluation information requires three assumptions.

1. We must assume that information regarding educational outcomes is essential to effective formal education. That is, we must consider *accountability* and *feedback* as essential mechanisms for the continued effectiveness of any educational program. Accountability is described as being able to demonstrate

the extent to which we have effectively and efficiently discharged responsibility. Without accountability in language teaching, students can pass several semesters of language courses with high grades and still be unable to use the language for reading or for conversing with speakers of that language. *Feedback* simply refers to information that is provided to teachers, students, and other interested persons about the results or effects of the educational program.

2. A second assumption is that it is possible to improve learning and teaching through appropriate changes in the program, based on feedback. Without these assumptions there is no reason to test, since there are no decisions to be made and therefore no information required.

3. We must assume that the educational outcomes of the given program are measurable. This, of course, is a highly debatable assumption. *At* one extreme there is the view that any outcome that is learnable is measurable, while at the other extreme is the view that virtually no worthwhile educational objective can be adequately measured.

In addition to these assumptions, we must also consider how much and what kind of testing is needed, as well as the quality of information provided by our tests. The amount and type of testing, if any, which is done depends upon the decisions that are to be made and the type of information that is needed to make the correct decisions. A second consideration in using tests **is** the quality of the information that they must provide. In educational programs the decisions made are generally about people, and have some effect on their lives. It is therefore essential that the information upon which we base these decisions be as reliable and as valid as possible. Very few of us, for example, would consider the opinions of our students' friends and relatives to be reliable information for evaluating their classroom performance. Likewise, the students' performance in other classes, such as mathematics or geography, would not be a valid basis for evaluating their performance in a language class.

Test usefulness

An overriding consideration in designing, developing and using language tests is that of test usefulness, which Bachman and Painter (1996) define as comprising several qualities: *reliability, construct validity, authenticity,*

interactiveness, impact and practicality. The usefulness of a given rest depends to a great extent on how test, takers perform on the test. This implies that the evaluation of lest usefulness must include the empirical investigation of test performance. There are two aspects of test performance that we need to investigate in our evaluation of lest usefulness: the processes or strategies test takers use in responding to specific test tasks and die product of those processes or strategies - individuals' responses to the test tasks and the scores that they obtain. In order to evaluate the usefulness of a given test, we need to investigate both aspects. The investigation of die processes and strategies test takers employ provides important information for the evaluation of test usefulness.

It is the responsibility of *test developers* to go beyond mere assertions of reliability and construct validity, and to provide evidence to test users that *demonstrates* that their tests have the qualities the developers claim it has. Test developers must provide evidence that supports the claims they make about how test scores are lo be interpreted and used. Similarly, it is the responsibility of *test users* to require test developers to provide such evidence, and to use this evidence appropriately and ethically in their own selection and use of language tests.

Test developers and test users can employ many different procedures and activities to collect the evidence for assessing the usefulness of tests for the particular purposes, test takers, and situations for which they are intended. This evidence will ideally include both quantitative data, such as test scores, scores for items or tasks, or responses to questionnaires and self-ratings, and qualitative data, such as observations, verbal self- reports by test takers, or samples of language produced during the assessment, that provides information about the usefulness of a given test. This book will focus on the kinds of quantitative data that can be collected, and some of the statistical analyses that can be used to help us evaluate the usefulness of the tests we develop and use. The statistical procedures described in this book can be used with any quantitative data, and they are relevant to the investigation of the qualities of usefulness.

Types of decisions

Since the basic purpose of tests in educational programs is to provide information for making decisions, the various specific uses of tests can be best understood by considering the types of decisions to be made. These decisions are of two types: decisions about individuals, which we might call *micro-evaluation*, and decisions about the program, which we might call *macro-evaluation*.

Decisions about students

There are four types of tests based on which decisions about students are made:

1. *Selection (entrance, readiness)*

The first decision that may be made about students is whether or not they should enter the program. In many programs, such as in primary schools, entrance is nearly automatic with age, while other programs require a *selection*, or *entrance* test. If the purpose of this test is to determine whether or not students are ready for instruction, it may be referred to as a *readiness test*. One of the most common uses of language tests for making decisions regarding selection is in conjunction with measures of other abilities, such as grades from previous instruction, academic aptitude, and achievement.

2. *Placement*

In many language programs students are grouped homogeneously according to factors such as level of language ability, language aptitude, language use needs, and professional or academic specialization. In such programs, therefore, decisions regarding the placement of students into appropriate groups must be made. Probably the most common criterion for grouping in such programs is level of language ability, so that placement tests are frequently designed to measure students' language abilities.

3. *Diagnosis*

Information from language tests can be used for diagnosing students' areas of strength and weakness in order to determine appropriate types and levels of teaching and learning activities. Thus, virtually any language test has some

chapter 7

Theory and Methods of DIF

Introduction

A consideration in the use of any psychological or educational test for selection is that the test is fair to all applicants, and is not biased against a segment of the applicant population. This issue, *known as test bias*, has been the subject of a great deal of recent research, and a technique called *Differential Item Functioning* (DIF) analysis has become the new standard in test bias analysis.

As active users of psychological tests, it is necessary for researchers, policy makers, and personnel selection officers involved in the evaluation of tests to be conversant with the current thinking in test bias analysis.

The two most commonly used *scoring formats* for tests and measures are *binary* and *ordinal*. Binary scores are also referred to as dichotomous item responses and ordinal item responses are also referred to as graded response, Likert, Likert-type, or polytomous. The ordinal formats are commonly found in personality, social, or attitudinal measures.

It is important to note that it is not the question format that is important here but the scoring format. Items that are scored in a binary format are either: (a) items (e.g., multiple choice) that are scored correct/incorrect in aptitude or achievement tests, or (b) items (e.g., true/false) that are dichotomously scored according to a scoring key in a personality scale. Items that are scored according to an ordinal scale may include Likert type scales such as a 5-point strongly agree to strongly disagree scale on a personality or attitude measure.

Current conceptions of validity theory with an eye to item bias

Technological and theoretical changes over the past few decades have altered the way we think about test validity. This section will briefly address the major issues and changes in test validity with an eye toward bias analysis.

Evaluating the measures: Validity and Scale Development

The concept of method and process of validation are central to evaluating measures, for without validation, any inferences made from a measure are meaningless. Throughout this presentation, the terms measure, observation, score, test, index, indicator, and scale will be used interchangeably and in their broadest senses to mean any coding or summarization of observed phenomenon.

Two central features in contemporary thinking in validation are:

- First, it is not a measure that is being validated but rather the *inferences* one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from a scale may appear at first blush subtle but in fact it has significant implications for the field of assessment.
- The second central feature in the above paragraph is the clear statement that all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an objective human resource/health/social indicator such as number of sick days, or a more psychological measure such as a depression scale or a measure of life satisfaction and well-being, one must be concerned with the validity of the inferences.

In recent years, there has been a resurgence of thinking about validity in the field of testing and assessment. This resurgence has been partly motivated by the desire to expand the traditional views of validity to incorporate developments in qualitative methodology and in concerns about the consequences of decisions made as a result of the assessment process.

The Contrasts between the traditional and more current views of validity are presented below:

The traditional view of validity focuses on:

- whether a scale is measuring what we think it is,
- reliability as a necessary but not sufficient condition for validity
- validity as a property of the measurement tool,
- validity as defined by a set of statistical methodologies, such as correlation with a gold-standard,
- a measure is either valid or invalid, and
- various types of validity -- usually four – in practice the test user or researcher assumes that they only need to demonstrate one of the four types to have demonstrated validity. Table 7.1 describes the traditional view of validity.

Table 7.1. The traditional categories of validity

Type of Validity	What do we do to show this type of validity?
Content	Ask experts if the items (or behaviors) tap the construct of interest.
Criterion-related:	
A. Concurrent	Select a criterion and correlate the measure of interest with the criterion obtained in the present
B. Predictive	Select a criterion and correlate the measure of interest with the criterion obtained in the future
Construct	Can be done several different ways. Some common ones are (a) correlate to a "gold standard", (b) a statistical technique called factor analysis, (c) convergent and discriminant validity

The process of validation then simply becomes picking the most suitable strategy from Table 7.1 and conducting the statistical analyses. For example, if we were conducting a human resource study on work environment and had developed a new measure of job satisfaction a common strategy would be to conduct a study to see how well the new measure correlates with some gold standard (e.g., the quality of work life scale) and if the correlation is sufficiently large then we would be satisfied that the measure is valid. This correlation with the gold standard is commonly referred to as a validity coefficient. Ironically, of course, the gold standard may have been developed and validated with some other gold standard.

It is important to note that there is, of course, nothing inherently wrong with the traditional view of validity. The purpose of *the more current view of validity* is to expand the conceptual framework of the traditional view of validity.

To help us get a better working knowledge of the more current conceptions of validity let us restate the traditional features listed above in the following way:

- construct validity is the central most important feature of validity and one must show construct validity;
- there is debate as to whether reliability is a necessary but not sufficient condition for validity; my view is that this issue is better cast as one of

measurement precision so that one strives to have as little measurement error as possible in ones inferences;

- validity is no longer a property of the measurement tool but rather of the inferences made from that tool;
- the validity conclusion is on a continuum and not simply declared as valid or invalid;
- validity is no longer defined by a set of statistical methodologies, such as correlation with a gold-standard but rather by an elaborated theory and supporting methods;
- consequences of *test decisions* and *use* are an essential part of validation; and
- there are no longer various types of validity so that it is no longer acceptable in common practice that the test user or researcher assumes that he/she only needs to demonstrate one of the four types to have validity.

As an example to help motivate our understanding of the more current thinking in validity let us consider the example of validating a job satisfaction measure. First, one is interested in gathering evidence supporting the trustworthiness that the scale actually measures job satisfaction and not some other related construct (such as coping with emotional stress, mental distress, general life dissatisfaction). To do that, one might consider:

- correlations with other theoretically related (i.e., convergent) and unrelated (i.e., discriminant) constructs,
- factor analysis of the scale,
- focus groups of target samples/groups of men and women, and expected age, or other differences to explore the consequential basis of validity.

To continue with the example, next, one would want to amass information about the value implications of the construct label itself, the broader theory of job satisfaction in women, and even broader ideologies about work life. For example, what is the implication of labeling a series of behaviors or responses to items as "high job satisfaction" and what does this mean for human resource policy, "do all people have to be happy and satisfied with their job, or is it sufficient to simply be able at their job tasks?."

And, if we want to use the measure in decision-making (or, in fact, simply use it in research) we need to conduct research to make sure that we do not

have bias in our measures. Where our value statements come in here is that we need to have organizationally and socially relevant comparison groups (e.g., gender or minority status).

As you can see from the simple example of job satisfaction, the traditional view of validity (as it is usually used in research; that is, either a correlation or just a factor analysis) is quite meager compared to the more comprehensive current approach.

Furthermore, the traditional view of validity is not tossed out but rather built on and expanded. Basically, the current view of validity makes validation a core issue that is not resolved by simply computing a correlation with another measure (or even a factor analysis of the items).

Another implication of the current view of validity is that we now need explicit statistical studies examining bias and concept-focused and policy studies of value implications for research use and decision making. Much of this has gone on in either informal or unsynthesized ways but now we need to bring this all together to address the issue of the inferences we make from measures.

Messick clarifies the issue of consequences. He states that it is not the obvious misuses of measures that is the issue but rather that we need to think about the unanticipated (negative and positive) *consequences* of the legitimate use and/or *interpretation* for decision making from measures that can be traced back to test invalidity- such as *construct under-representation* and/or *construct irrelevant variance.* Item bias studies are examples of the sort of questions that need to be asked. That is, the validation process begins at the *construct definition stage* before items are written or a measure is selected, continues through *item analysis* (even if one is adopting a known measure), and needs to continue when the measure is in use.

Bias and differential item functioning: Definition and history

The use of the term "bias" in assessment research fundamentally follows its popular usage, which conveys a skewed and unfair inclination toward one side (group, population) to the detriment of another. The notion of bias is directly tied to fairness, in popular usage as well as assessment: A biased judgment unduly takes into account factors other than those that should be informing it. For assessment, bias can be seen in traditional validity terms as *construct-irrelevant variance* that distorts the test results and therefore makes

conclusions based on scores less valid. Specifically, a test or an item is biased if test takers of equal ability but from different groups score differently on the item depending on their *group membership*. In this case, group membership introduces systematic construct-irrelevant variance, which has a consistent effect on scores. Another way to look at this is to *consider bias a factor that makes a unidimensional test multidimensional*: The test measures something in addition to what it is intended to measure, and the result is a confound of two measurements.

Whereas any construct-irrelevant variance is harmful to valid interpretations, bias systematically harms one group by inflating one group's scores and depressing the other group's scores. In a broader sense, biased tests harm all stakeholders because students might get exempted from language programs although they would benefit from them, others do not get admitted to a program in which they would excel, universities or employers reject perfectly qualified applicants and accept less qualified ones, and society is deprived of potentially excellent doctors, lawyers, language teachers, or electricians and must make do with mediocre ones.

A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society. The first two characterizations relate fairness to *absence of bias* and to *equitable treatment of all examinees* in the testing process. There is broad consensus that tests should be free from bias and that all examinees should be treated fairly in the testing process itself (e.g., afforded the same or comparable procedures interesting, test scoring, and use of scores). The third characterization of test fairness addresses the *equality of testing outcomes* for examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature. A more widely accepted view would hold that examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership. The fourth definition of fairness relates to *equity in opportunity to learn* the material covered in an achievement test. There would be general agreement that adequate opportunity to learn is clearly relevant to some uses and interpretations of achievement tests and dearly irrelevant to others, although three interrelated aspects contribute to absence of bias quality: