Can the Search for "Fairness" Be Taken Too Far?

Elvis Wagner
Teachers College, Columbia University

The many works devoted to the issue of fairness in language testing (e.g., Kunnan, 1999, 2000; Shohamy, 2001; Spolsky, 1981) testify to the field's recognition of the importance of this issue. Brown (1996) defines fairness as "the degree to which a test treats every student the same or the degree to which it is impartial" (p. 31). The goal of language tests is to impartially measure individual test-takers' language ability. If, however, performance on those tests is influenced by factors other than language ability, then *bias* is introduced into the measurement.

Obviously, for ethical reasons, it is important to create fair and unbiased tests. But from a language testing standpoint, fairness and bias are also validity issues. Bachman and Palmer (1996) define construct validity as "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (p. 21). Messick (1989, 1996) describes how the introduction of construct irrelevant variance in a test threatens the construct validity of a test. Language tests are designed to measure an individual's language ability, but if the test-taker's employment background (for example) influences his or her performance on the test, then it is generally considered that construct irrelevant variance has been introduced into the measurement of the individual's language ability. For example, a reading test is given to a group of test-takers from diverse backgrounds. The text describes how to pilot a helicopter. It would seem that test-takers who happen to be helicopter pilots would have an unfair advantage on this part of the test, and the other test-takers would be unfairly disadvantaged. The inferences made about a person's reading ability based on the results of the test are of questionable validity. The bias in the test threatens the validity of the inferences made based on that test. A biased test cannot be a valid test. Subsequently, a test that is not valid can never be a fair test.

Bias, viewed as measurement error, introduces construct irrelevant variance. Test developers seek to eliminate this construct irrelevant variance by minimizing test bias. Bachman (1990) lists some of the test-taker characteristics of an individual that might lead to biased assessments, including cultural background, background knowledge, cognitive characteristics, native language, ethnicity, sex, and age. Bachman and Palmer (1996), in describing language test development procedures, created a framework allowing test developers to systematically review their tests to investigate (and minimize) sources of bias in their tests. Large scale test developers have bias committees that examine test items and materials for potential sources of bias.

However, in their zeal to eliminate bias from language tests, test developers might be going too far. As mentioned earlier, Messick (1989, 1996) describes how the introduction of construct irrelevant variance in a test (in this case, bias) can threaten the construct validity of that test. But Messick also lists another threat to the construct validity of a test: construct underrepresentation. In order to make valid inferences about the test-taker's ability based on the test results, the items on a test must be adequately representative of the ability and content domain that is being assessed. If the test is too restrictive, and is not representative of the ability and content domain the test purports to assess, the validity of that measure is suspect. For example, if a test developer created a test that purported to assess a student's overall proficiency in American history, but the test only had items that assessed American history since 1970, the inferences made from the results of that test could hardly be considered valid, because of content and construct underrepresentation. Or, looking at the issue from the opposite perspective, if a test developer created a test to measure helicopter pilots' ability to use English for aviation purposes, it would certainly be appropriate (and probably necessary) to include texts that involved content related to piloting a helicopter. If a test-taker did not have knowledge about piloting a helicopter, and this affected his or her performance on the test, this would still be considered construct-relevant variance, because this knowledge was part of the defined content domain for this test.

Some real-life examples might serve to make this issue clearer. I have worked for a large testing organization writing test items for state standardized K-6 ESL exams. One of the first rules given to me was never to write texts or items involving the topic of birthdays. The reason for this was because these items never made it past the bias committee. That is, because particular religious groups do not celebrate birthdays, the bias committee felt that items that included this idea would be unfair to members of those religious groups. There are two problems with the approach taken by the bias committee. First, considering how incredibly diverse and multicultural the U.S. school population is, it is virtually impossible to create texts or items with content that someone will not have objections to. Second, it would seem that birthdays are part of the content domain that is purportedly being assessed by these tests. The topic of birthdays is something that is prevalent in the academic and real-life language domains for this population of learners. By arbitrarily dictating that birthdays cannot be included in the content domain for these exams, threats to the validity of these exams due to content and construct underrepresentation are introduced. In striving to reduce threats to validity due to bias (construct-irrelevant variance), the test developers introduce a different threat to validity (content and construct underrepresentation).

Another example might be useful in demonstrating this phenomenon. I was once discussing a test that a colleague was developing. She described how she was trying to eliminate bias from her exam by eliminating vocabulary words in the texts and test items that were cognates in some of the native languages (e.g., Spanish) of the test-takers, but that were not cognates in other of the test-takers' native languages. For example, she had decided not to use a reading text in which the topic was inflation, because *inflation* is a cognate in Spanish (*inflación*), and thus she felt that the use of this text was unfairly advantaging native Spanish speakers, while disadvantaging test-takers in whose native language *inflation* was not a cognate. But again, the result of her attempts to reduce bias ultimately introduced other threats to validity. If the concept of inflation was part of the content language domain that the test was purporting to assess, then to exclude it from the test resulted in content and construct underrepresentation.

A final example to illustrate this point is related to the use of videotexts in the testing of second language listening ability. Studies (e.g., Burgoon, 1994; Wagner, 2006) have presented evidence that individual listeners vary in their ability to utilize the nonverbal information transmitted by speakers in order to comprehend spoken texts. Some researchers (e.g., Buck, 2001) have interpreted these findings as indicating that videotexts should not be used in testing L2 listening ability. This interpretation seems to be based on the notion that by including the visual channel, one would introduce bias into the assessment by unfairly advantaging those test-takers that were able to understand and utilize this visual information, and disadvantaging those test-takers who were not able to understand and utilize this visual information. According to this interpretation, then, the inclusion of the visual channel would introduce construct irrelevant variance. However, an alternative interpretation is that the inclusion of the visual channel leads to construct relevant variance. Due to the fact that in most language use settings the listener is able to see the speaker (and utilize the nonverbal information that the speaker conveys), the test-takers' differing levels of ability to utilize the visual information is construct relevant variance. The ability to utilize the nonverbal information conveyed by the speaker to understand the spoken text appears to be part of the construct definition of listening ability (which seems to be true in settings other than listening to the radio or talking on the telephone). In this case, excluding it on a test (by using audio-only texts) results in construct underrepresentation, therefore making the validity of the inferences based on the results of that test suspect.

In their efforts to eliminate bias from language tests (to avoid threats to validity due to the introduction of construct irrelevant variance), test developers might be guilty of inappropriately narrowing the domain that is being assessed, and thus introducing threats to validity due to content and construct underrepresentation. Tests, by their very nature, are designed to differentiate the test-takers. Eliminating bias from tests in order to assure that a particular group of test-takers is not unfairly disadvantaged (because of their cultural or linguistic background, age, gender, etc.) is a worthwhile goal, but this must be undertaken in a logical and systematic manner. As Bachman (1990) argues, "differences in group performance in themselves do not necessarily indicate the presence of bias, since differences may reflect genuine differences between the groups on the ability in question" (p. 271). How the content domain and construct are defined and operationalized is of paramount importance. If the characteristic that influences test performance is part of the construct definition of the ability being assessed, including this characteristic in the assessment will lead to construct relevant variance. It is when these differences in performance on the test are associated with characteristics that are *not* inherent in the ability that is being assessed that bias occurs.

REFERENCES

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

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Buck, G. (2001). *Assessing listening*. Cambridge, UK: Cambridge University Press.

Burgoon, J. (1994). Non-verbal signals. In M. Knapp & G. Miller (Eds.), *Handbook of interpersonal communication* (pp. 344-393). London: Routledge.

Kunnan, A. J. (1999). Recent development in language testing. *Annual Review of Applied Linguistics, 19*, 235-253.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge, UK: Cambridge University Press.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 242-256.

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

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5-21). Frankfurt, Germany: Peter D. Lang.

Wagner, E. (2006). *Utilizing the visual channel: An investigation of the use of videotexts on tests of second language listening ability*. Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Dr. Elvis Wagner is a Lecturer in the TESOL and Applied Linguistics programs at Teachers College, Columbia University. His teaching and research interests include language testing, and foreign and second language teaching methodology.