The Development of Culturally Appropriate Tests

Taiko Tsuchihira

1. Purpose of the study

There has been a lot of controversy over how appropriate the tests are. Usually, the wider the area of the test is adopted, or the more people that take the test, the bigger the controversy is. There has been a lot of research done on the appropriacy or the fairness of certain tests, such as language tests. Some say that there is an unequal phenomenon, and others say there is not. Each conclusion sounds very simple, but it is very complicated, partly because each researcher employs different methods and also because each study has different views on this problem from the beginning. It is necessary for us to realize what has been done so far before considering the new testing methods. The purpose of this study, therefore, is to 1) provide an overview of our recognition about culture in testing by eras and issues and 2) to make some possible remarks on culturally appropriate testing in the future.

2. Culturally-sensitive tests

There has been much concern for culture-fair tests. Early work in language testing showed a concern for culture-fair tests when test were developed for monolingual/cultural groups. One of those studies is Briere (1968), and Briere and Brown (1971). Both studies are focused on

children of American Indians, and they both are a part of the project called the English Language Testing Project. This project aimed to provide certain measures for teachers to determine how much the American Indian students understand the material. At this stage, most of the studies in the United States had done on children of American Indians.

More studies have been done on how tests affect the performances of students from different cultures. For example, Farhady (1979), who argues the similarities between discrete-point tests and integrative tests, presents the potential variables influencing test results and creating test biases. Citing his precious work, he insists that there should be a significant difference in the performance of foreign students (e.g., Israeli and Taiwanese in his study), depending on their language and educational backgrounds, on how they score on discrete-point and integrative tests. On choosing the tests, as a conclusion, he suggests including different types of subtest to obtain an accurate picture of examinees' language ability.

3. Differential item functioning (DIF)

In the 1990s, there were abundant studies on test bias or differential item functioning focusing on different races and cultural groups (e.g., Berk 1982; Holland and Wainer 1993 etc.). According to Thissen et al. (1991), differential item functioning (DIF) is defined as a phenomena that "...some test items simply function differently for examinees drawn from one group or another or they measure different things for members of one group as opposed to members of another...". (Thissen et al. 1991) Here, DIF has been used almost in the same way

as item bias though the latter has a connotation for partiality while DIF sounds more neutral.

There are some ways to detect DIF. Some studies simply use means and standard deviations (Dorans and Livingston 1987). However, mainstreams are said to be 1) statistical standardizing methods (Dorans and Kulick 1986), 2) IRT (Item Response Theory) -based methods, and 3) methods using chi-square indexes including the Mantel-Haenznel procedure (Holland and Thayer 1988).

Among the abundant studies on DIF, there seems to be three main interest factors. First of all, many researchers attempted to detect DIF between male and female. Some succeeded in detecting DIF, and some did not. For example, Dorans and Livingston (1987) tried to detect DIF between two genders using the Scholastic Aptitude Test (SAT) though the result was not very clear. Doolittle and Cleary (1989) found DIF in math tests by genders, and McLarty, Candance, and Huntly (1989) detected the effect of item wordings on sex, and found unsystematic difficulty differences. In their extensive study, Ryan and Bachman (1992) tried to find DIF in the Test of English as a Foreign Language (TOEFL) between genders, though they could not detect any DIF.

The second interest is the influence caused by racial differences. Schmitt (1988) examined DIF between U.S. whites and Hispanics in SAT. Houston and Novick (1987) used data from the Armed Services Vocational Aptitude Battery (ASVAB) and tried to detect DIF, and found that the test was disadvantageous for Blacks. Schumitt and Dorans (1990) tried to detect DIF across forms of the questions and ethnic groups in SAT, and found some DIF. Freedle and Kostin (1990) examined the results of GRE and SAT across black and white students

according to their four vertical forms and successfully detected DIF. Moreover, Scheneman and Gerritz (1990) succeeded in detecting DIF both in different sexes and races, using SAT and the Graduate Record Examination (GRE). They discovered that certain types of items affect the results of the tests.

Other studies examined DIF caused by various other factors. Johnson and Wallace (1989) examined how coaching test-taking strategies affects test results using math tests. Tatsuoka, Linn and Yamamoto (1988) studied DIF between the groups of different strategies, and proved that different strategies cause DIF.

Thus, although there are so many studies on DIF in this era, only a few studies are relevant to language learning. Ryan and Bachman (1992) is one of a few studies on DIF focusing on language tests. They tried to detect DIF in language tests between the learners of different native languages. They divided the learners according to their native language: Indoeuropean or non-Indoeuropearn, and examined DIF between the groups. As a result, they detected several items showing DIF. Kunnan (1990) examines DIF caused by different native languages and genders. They use UCLA's ESL placement examination (ESLPE), and detected some items functioning differently to different groups. Chen and Henning (1985) examined the Winter 1985 version of ESLPE to determine the nature, direction and extent of bias present for members of the Spanish and the Chinese native language groups, and found some test items exhibiting bias in the receptive skill domains. They found DIF, especially of the test items, consisted of English words for which close cognate forms existed in the Spanish language which were not similarly available in Chinese. Zeinder (1986) examined the test bias

between Arab students and Jewish students by using the English Aptitude Test which is often used for student selection and placement in Israel. However, the data provides only marginal evidences for DIF, and rather, negated the cultural bias hypothesis.

Thus, as we have seen, the group differences were extensively studied in this era. The term DIF was invented and more advanced statistical techniques were employed. Table 1 summarizes the studies of DIF according to the factor they researched and the results.

Table 1.

Studies	Tests Used	Factor	+/ −
Dorans & Livingston (1987)	SAT	Sex	+−
McCornack & McLeod (1988)	GPA		−
Ryan & Backman (1992)	TOEFL		−
Dolittle & Cleary (1989)	ACTM		+
Scheuneman & Gerritz (1990)	GRE & SAT		+
McLarty, Noble & Hunley (1990)	math test		+−
Houston & Novick (1987)	ASVAB	Race	+
Ryan & Bachman (1992)	TOEFL		−
Schmitt (1988)	SAT		+−
Schumitt & Dorans (1990)	SAT		+−
Freedle & Kostin (1990)	GRE & SAT		+
Scheuneman & Gerritz (1990)	GRE & SAT		+
Kunnan (1990)	ESLPE	NL	+
Chen & Henning (1985)	ESLPE		+
Zeinder (1986)	ELAT		+

As we can see in the table, towards the late 1980' s, many researchers came to seek the reason for the performance difference into cultural differences (e.g., Kunnan (1990); Chen & Henning (1985); Zeinder (1986, 1987) though the DIF was studied mainly on genetic

differences at the beginning. With that trend, it is noticeable that group differences are found in more studies.

4. Background and culture knowledge in testing

There is another group of studies which accessed the cultural bias on language testing from different approaches. They used quasi-experimental designs and attempted to see the differences caused by test-takers' background and culture knowledge. For example, Gatbonton and Tucker (1971) compared the Filipino students and the American students using experimental groups and control groups. They found out that cultural orientation affects literature appreciation. Chihara, Sakurai and Oller (1989) made an experimental comparison of the two original passages which includes minor changes such as names of persons and places, and examined if these minor changes in textual elements would result in a better understanding of the text. As a result, they found a significant difference between the unchanged and modified texts, and these changes led to a far better performance of the Japanese students. Kitao (1981) is another similar example of studies about cultural knowledge. He presents a multiple choice test of American culture to measure which items of American culture Japanese students know and examined to what degrees these items are understood.

These results are not very surprising, however, if we consider the fact that background knowledge is relevant to comprehension. It is very natural that the amount of knowledge about the context, in other words, cultural knowledge, naturally matters to comprehension.

Interestingly, however, Angoff (1989) conducted a similar experiment between the examinees who had lived in the U.S. for more

than a year and those who had spent little time in the U.S., and he suggested completely opposite results in the Research Reports for the TOEFL. He employed five raters to rate whether the test items contains explicit reference to some aspect of Americana, and the "Americana" scores were compared with the Mantel-Haenznel indices which have often been used in the DIF studies. Then, as a result, he states that there is no relation found between them. Therefore, it was concluded that there is no support for the hypothesis that the test items that make reference to American people, places tend to advantage those who have lived in the U.S. for more than a year. This conclusion suggested by Angnoff (1989) is very surprising, and it is something which is contradictory to many intuitions. Although the study was extensive and thorough, more discussions and replicated studies might be necessary in order to accept the conclusion in a general sense.

5. Culturally appropriate test methods in the future

5. 1 What is a fair test?

So far, the present study shows that there has been much research and discussion on the appropriacy of some language tests. It has been found that there are certain validity problems in language tests and other tests in general, and it is essential that we consider appropriate test methods. Here, however, before considering culturally appropriate test methods, it is meaningful to discuss some arguments raised in past literature.

The first argument is how we should utilize the implications of those studies in making language tests. In many studies, the researchers state that we should remove or improve the items which caused the

group differences. However, there are some cases that we cannot remove those items, according to Chen and Henning (1985). As is stated before, they examined the ESLPE, and found that some test items are more advantageous for the members of the Spanish native language groups than others. However, they found DIF especially of the test items consisted of English words with close cognate forms existing in the Spanish language but not similarly available in Chinese, and they argue whether they should remove all those items.

What does this argument mean? This fact shows us a crucial problem in studying group differences. That is, if they include essential parts of English language, some items cannot be removed even though we found the group differences on those items. In the case of Chen and Henning (1985), the items are English vocabulary which include Spanish cognates. If we consider this, it is quite obvious that there are many English words which include Spanish cognates, and therefore look similar to Spanish words. It is impossible to remove all those words since they are too many. Furthermore, even though we remove those items it is very doubtful if we can call the rest of the items as the very English that native speakers use.

This thought brings us to the second argument. That is, no matter how often we repeat removals or improvements of items, there must be always crucial group differences or handicaps if the test takers are from the different groups. If the culture of the test-takers is similar or close to that of the test writers, then the problem is small, and if the culture of the test-takers is distant from that culture of the test writers, then the problems become apparent. As long as there is a distance between the culture of test-takers and that of the test writers', the appropriacy

problem unceasingly exists at a fundamental level no matter how we try to cope with biased items. Is it possible to make fair tests at all?

5. 2 Studies on testing minority students

Among the studies on testing minority students, there is a certain trend to criticize standardized testing. Bordeaux (1995) argued that standardized norm-referenced testing is no longer universally accepted as the best method. Moreover, Geisinger and Carlson (1992) claimed the dis-equalised situation for language minority students. They claim that schools often use inappropriate standardized instruments to determine the English language fluency of limited English proficient (LEP) and language minority children. They also state that school employees with little or no knowledge of the child's first language or culture often administer these instruments. They conclude that schools need to notice the cultural diversity and use an unbiased, fair method of measurement.

In addition to them, there are several studies on evaluating properly students from American Indian communities. In summarizing the literature on the cultures of American Indians and the testing problems, Neely and Shaughnessy (1984) listed six problems in using tests with minority students: 1) inappropriate content; 2) inappropriate standardization sample; 3) examiner and language bias, 4) inequitable social consequences; 5) measurement of different constructs; 6) differential predictive validity. Brescia and Fortune (1988) state that many American Indian students fail to exhibit successful test-taking behaviors such as reading directions correctly, test-taking skills, and cognitive structure to respond to certain items. They insist on the necessity of providing proper test-taking preparation with American

Indian students. They add that cultural beliefs in some Indian tribes may even bar competitive behaviors in an academic setting, and that some students might underestimate the seriousness of the tests. They also claim that an acculturation problem exist in testing American Indian students, and note that many American Indian students are experiencing poverty, low parental education, broken homes, and non-standard English backgrounds. They also suggest this acculturation problem relate to a motivation problem. In other words, it is not very possible to have motivation for higher marks in tests if the students are living in discouraging, deprived situations.

Furthermore, Blanchard and Reedy (1970) attempted to identify factors contributing to the poor achievement levels of American Indian students in standardizing tests. They administered Test of English as a Foreign Language (TOEFL), Iowa Test of Educational Development (ITED), Tennessee Self Concept Scale (TSCS), and the Southwestern Indian Adolescent Self-Concept Scale (SIASS) and analyzed the relationships of each other. They found interrelation factors between educational retardation, low self-concept, and skill in the English language, and this suggests that we should consider that language, culture and self-concept are inextricably interwoven.

Considering the examples above, it is possible to say that the phenomena of poor academic achievement by minority students has resulted from the overuse of the single measurement method such as standardized norm-referenced test. Culture, language, and students' environment are all intertwined, and those multiple factors are occurring simultaneously to the students. These studies are good examples of bringing a norm of other culture to one culture automatically causing a

distance which is fairly difficult to adjust completely, as is shown in 5.1. Therefore, allowing diversity and differences are necessary rather than devising a new adjustment to maintain a norm. In other words, a new way of testing should be devised in the future.

One might say, however, that minority students have to adapt to the society majority. We should consider two approaches depending on how we consider languages. If the goals of language learners are to be successful as a fair member of the society, then the culture of test maker might be still applicable. The minority student wants to accomplish the goal in the end. However, if we think of an identity of learners, motivation, and their life related to the learning situations, then evaluation should take the form of the culture of test-takers. With the descriptions of the education of Chicanos, Trevino (1973) strongly suggests to supporting bilingual-bicultural opportunities at schools.

In the discussion on language education in the U.S., Postman (1980) states that the improvement of reading scores is not a legitimate educational goal, and notes that reading abilities are inseparable from other modes of linguistic expression. In other words, language is embedded in our life itself, and hardly separable from it. Navarette et al. (1990) suggests holistic, informal on-going assessment of students' growth. She proposes a combination of unstructured and structured assessment. According to her, unstructured assessment is, for instance, writing samples, homework, journals, games, and debates, and structured assessment is based on checklists, tests, rating scales, questionnaires, structured interviews. Using student portfolios is also recommended, and she suggests some guidelines.

The method of evaluating language ability reflects how test-

constructors perceive languages just in the same way as a teaching methodreflects the teschrs'idea about languages.Languages are so embedded in our life, intertwined with culture and society that we should not rely too much on a simplistic method, which sorely distorts the reality of language itself. It is necessary to reconsider our policy and method of testing language abilities from this perspective.

REFERENCES

Angoff, W. H. (1989). Context Bias in the Tesa of English as a Foreign Language. *Research Report* 2, Educational Testing Service.

Berk, R. A. (ed.)(1982). *Handbook of methods for detecting test bias.* Baltimore: The John Hopkins University Press.

Blanchrd, J. d. and Reedy, R. (1970). The Relationship of a Test of English as a Second Language to Measures of Achievement and Self-Concept in a Sample of American Indian Students. *Research snd Evaluation Report Series,* 58. (ERIC Document Reproduction Service No. ED 147090).

Bordeaux, Roger. (1995). Assessment for American Indian and Alaska Native Learners. ERIC Digest. (ERIC Document Reproduction Service No. ED 385424).

Brescia, W. and Fortune, J. C. (1988). Standardized Testing of American Indian Students. ERIC Digest. (ERIC Document Reproduction Service No. ED 296813).

Brire, E. J. (1968). Testing ESL among Navajos children. In J. A. Upshur and J. Fata (eds.) *Problems in foreign language testing. Langusge Learning,* Special Issue 3:11-21.

Briere, E. J. and R. H. Brown. (1971). Norming tests of ESL among

American Indisn children. *TESOL Quarterly*, 5: 327-34.

Cargill-Power, C. (1980). Cultural Bias in Testing ESL. ERIC Digest. (ERIC Document Reproduction Service No. ED 192621).

Chen, Z. and G. Henning. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155-63.

Chihara, T. Sakurai, A.,and Oller, J. (1989). Background and Culture as Factors in EFL Reading Comprehension. *Language Testing*, 6(2), 143-151.

Doolittele, A. E. and Cleary, T. A.(1987). Gender-based differential item performance in mathematics achievement items, *Journal of Educational Measurement*, 24(2), 157-166.

Dorans, N. J. and Livingston, S. A. (1987). Male-female differences in SAT-verbal ability among students of high SAT-mathematical ability. *Journal of Educational Measurement*, 24(1), 65-71.

Farhady, H. (1979). The disjunctive fallacy between discrete-point and integrative test. *TESOL Quarterly*, 13, 347-58.

Freedle, R. and Kostin, I. (1990). Item difficulty of four verbal item types and sn index of differential item functioning for black and white examinees. *Journal of Educational Measurement*, 27(4), 329-343.

Gatbonton, E. C.,and Tucker, G. Richard. (1971). Cultural Orientation and the Study of Foreign Literature *TESOL Quarterly*, 5(2), 137-143.

Geisinger, K. F. and Carlson, J. F. (1992). Assessing Language-Minority Students. ERIC Digest. (ERIC Document Reproduction Service No. ED 356232).

Holland, P. W. and Wainer, H. (1993). *Differential Item Functioning.*

Hillsdale:Lawrence Erlbaum Associates.

Houston, W. M. and Novick, M. R. (1987). Race-based differential Prediction in air force technical training programs. *Journal of Educational Measurement,* 24(4), 309-320.

Kitao, K. (1981). The Test of American Culture. *NALLD Journal*;15(2), 25-44.

Kunnan, A. J. (1990). Differential item functioning and native language and gender groups:The case of an ESL Placement examination. *TESOL Quarterly,* 24, 741-746.

McCornack, R. L. and McLeod, M. M. (1988). Gender bias in the prediction of college course performance. Journal of Educational Measurement, 25(4), 321-331.

McLarty, R., Noble. C. A., and Huntly, R. M. (1989). Effects of item wording on sex bias. *Journal of Educational Measurement,* 26(3), 285-293.

Neely, R. and Shaughnessy, M. F. (1984). Assessment and the Native American. (ERIC Document Reproduction Service No. ED 273889).

Postman, N. (1980). Language education in a knowledge context. *ETC.: A Review of General Semantics,* 37(1), 25-37, Spring 1980.

Ryan, K. E. and Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing,* 6, 13-30.

Scheyneman, J. D. and Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement,* 27(2), 109-131.

Schmitt, A. P. (1988). Language and cultural characteristics that

explain differential item functioning for Hispanic explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement,* 25(1), 1-13.

Schmitt, A. P. and Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement,* 27(1), 67-81.

Smith, J. B. (1994). The Dis-Equalizing Impact of Standardized Testing on Language-Minority Children. (ERIC Document Reproduction Service Service No. ED 374664).

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item response theory in the study of groups differences in trace lines. In H. Wainer and H. I. Braun (eds.), Test Validity. Hillsdale, NJ:Erlbaum.

Trevino, Robert E. (1973). Is Bilingual Education Shortchanging the Chicano? (ERIC Document Reproduction Service No. ED077617).

Upshur, J. A., and Fata, J. (eds). (1968). Problems in Foreign Language Testing; Proceedings of a Conference Held at the University of Michigan, September 1967. *Language Learning,* Special Issue No. 3 August.

Welch, C., Doolittle, A. and McLarty, J. (1989). Differential performances on a direct measure of writing skills for black and white college freshman. *ACT Research Report Series,* 8. Iowa City: American College Testing Program.

Zeinder, M. (1986). Are English language sptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3: 80-98.

Zeinder, M. (1987). A comparison of ethic, sex, and age bias the

predictive validity of English language aptitude test: Some Israeli data. *Language Testing*, 4: 55-71.