

BE 18869

The Testing of Minority Language Students

Annette Archer Conroy

Background

During the early 1900s concerned citizens began to feel the need to assess the new waves of immigrants who threatened to contaminate American soil with new ideas and ways of life. Foremost among these citizens was H. H. Goddard who translated the recently developed Binet test into English and administered it to newly arrived immigrants on Ellis Island (Hakuta, 1986). In testing 30 adult Jews, he discovered that 25 of them were evidently feeble-minded. The test demonstrated to him that America was now attracting the "poorest of each race" and resulted in Goddard's (1917) recommendation that "if the American public

Annette Archer Conroy is an ESL specialist at the elementary level in the Highline School District in Seattle, Washington. She is on a one-year leave to work in a Bilingual Studies Doctoral Fellowship Program at the University of Washington.

wishes feeble-minded aliens excluded, it must demand that Congress provide the necessary facilities at the ports of entry." He boasted that "the number of aliens deported because of feeble-mindedness...increased 350 percent in 1913 and 570 percent in 1914 This was due to the untiring efforts of the physicians who were inspired by the belief that mental tests could be used for the detection of feeble-minded aliens" (p. 271).

During World War I, the draft enabled Goddard and others to conduct mass intelligence testing of two million recruits, many of foreign nationalities, which allowed for the comparisons of intelligence according to race (Hakuta, 1986). Carl C. Brigham (1923), a primary participant, found special significance in the differences in intelligence between foreign born draftees who had lived in the U.S. for over 20 years and those of shorter residence. Those who had lived in the United States the longest had significantly superior intelligence according to the tests. For Brigham, the data clearly pointed to the intellectual superiority of earlier immigrations:

Migrations of the Alpine and Mediterranean races have increased to such an extent in the last thirty or forty years that this blood now constitutes 70 percent or 75 percent of the total immigration. The representatives of the Alpine and Mediterranean races in our immigration are intellectually inferior to the representatives of the Nordic race which made up 50 percent of our immigration. (p. 197)

Brigham's data also indicated that Nordic immigrants from English-speaking countries were more intelligent than those from non-English-speaking countries (Hakuta, 1986).

To modern educators, the conclusions drawn by Goddard, Brigham, and others seem patently absurd. The classification of new immigrants as feeble-minded on the basis of intelligence tests given in English would receive no support, while it would be immediately obvious that a longer residency in the U. S. would result in a greater knowledge of the language and culture and, consequently, a higher score on an intelligence test. Brigham's dismissal of the idea of a language handicap in testing in spite of the higher scores of English-speaking new immigrants (Hakuta, 1986) would be scorned as unscientific ignorance. In fact, these conclusions are so evidently nonsensical to the modern educator that it seems pointless to bring them up. They should be relegated to the past along with outdated views of insanity or disease as interesting and amusing bits of history. But, tragically, they have not slipped into the past. These identical conclusions in a different guise are being drawn today in our modern schools by the same enlightened educators who would laugh at Goddard and Brigham. In spite of research, in spite of protests and law suits, minority

The greater the difference between the language of the student's community and that of the tester, the more likely it is that the student will be ignorant of the vocabulary or differ in his responses from the norm established by the tester.

The same difficulties apply to other commonly used assessments such as short term memory of sentences, analogies, and verbal reasoning which on the Stanford-Binet and other IQ tests include comprehension, recognition of absurdities, similarities, and verbal relations (Conoley & Kramer, 1989). As anyone can attest who has tried to repeat sentences in a foreign language, the more distinct word patterns are from the student's home language, the more difficult it is to repeat them. The recognition of analogies depends heavily on the language and experiences of a student's community. Sigel (1965) used the example from the Wechsler Intelligence Scale for Children (WISC) which asks "How are a scissors and a copper pan alike?" A score of one is given to "Both are metal." Since both responses are class concepts, he asks, "How does one decide which system is preferable?" Another test asked the respondent to pick the word which did not belong among clam, tree, oven, and rose. While oven was judged to be the correct response because the other items are living organisms, Sigel points out that clam, tree and oven could be classified as related to food preparation with rose excluded (p. 288).

Just as language reflects the home community, so does the knowledge embodied by that language. The knowledge sections of intelligence tests are based on the assumption that the intelligent child absorbs common knowledge by interacting with the environment. Without arguing the validity of the assumption, the tests themselves must be called into question on their attempt to define a standard body of knowledge. A body of knowledge based on the experiences of a white, middle class community cannot help but create a disadvantage for members of communities which differ from the mainstream. J.L. Ribeiro (1980), himself a Portuguese immigrant, determined in testing over 350 low-income rural Portuguese children with the WISC-Revised, that this section was the most biased. For example, the children had difficulty with the question "from what animal do we get bacon?" in spite of a diet which included pork as a staple. Bacon was not a familiar term in their community. But Ribeiro suggested that they would have had no difficulty if asked concerning the animal from which sausages come, since the latter is a part of their diet. The children found the item "Why does oil float on water?" to be easy compared to prior "easier" questions because as church-going Catholics, they were familiar with the lamp which traditionally burns on the alter with oil floating on water. It is difficult to see how either of these questions can be said to measure intelligence rather than experience.

Although achievement tests are supposed to measure learning, not innate ability, the issues are similar, if only because current scores are assumed to predict future performance. As with IQ tests, vocabulary and subject matter are drawn from the experiences of the "mainstream" community. But because the tests are designed to evaluate specific competencies, the dichotomy between what is purported to be tested and what is actually assessed is even more pronounced. On the MAT test (Prescott, Balow, Hogan, & Farr, 1984), for example, the Vocabulary section becomes a test of syntax as well as word meaning when students are required to choose correct words to complete sentences. Both the Vocabulary and the Reading Comprehension sections draw from material which would be most familiar to white, middle class students. Any reading teacher will acknowledge that students find reading more comprehensible when the subject matter and language are familiar. The language section is clearly intended to evaluate knowledge of standard English as is demonstrated by Sample D on the Primary 1 test which asks "Where ___ my new shoes" (p. 45)? The possible answers include both "are" and "be." Although speakers of certain Black dialects might use "be," only "are" is acceptable. Virtually every question in this section includes alternatives which are common in nonstandard English dialects, all of which are wrong answers. For the first graders to whom the Primary test is administered, the more closely their spoken language matches that of the written test, the greater their advantage in responding as required is.

Math Concepts is another section in which language interferes with evaluation of actual ability. For example, a child with a good understanding of ordinal numbers might still have difficulty with this item from the Primary 1 test often given to first graders in the fall:

Under the row of animals lined up behind the girl, you will see some animals with the pig at the top and a dog at the bottom. Mark next to the picture of the animal which is the sixth one away from the girl. (p. 29)

Any child might find this confusing, but especially one for whom standard English is not the home language. Rather than assessing knowledge of ordinals, it evaluates children on their ability to wend their way through the intricate language. Along with the assessment of basic academic skills, then, goes the assessment of a child's knowledge of the language of mainstream America.

The Students

Having established that much of what is assessed in standardized testing is based on a student's knowledge of standard American English, it follows

logically that students who are linguistically at a disadvantage would score lower as a group than mainstream, standard English proficient students. And, in fact, the two main groups of minority language students in our schools, those who speak English as their second language and those whose first language is a minority dialect of English, consistently score low in comparison. But when the difference in scores is displayed by the media, that difference is attributed to the schools for their failure in teaching these students, to the minority families for not taking a sufficiently strong stand for education, to the students themselves for their poor work habits. Testing is rarely questioned. The focus is invariably on raising test scores.

Duval County (Holmes, 1986), justifiably proud of its efforts in improving the achievement of minority students, nonetheless began with the assumption that improvement was to be achieved in focusing on test results and working with teachers, students, and parents to "close the gap" (p. 38) between the scores of minority students and those of mainstream students. Never was it considered that even if all else were equal, the tests themselves might be responsible for a gap.

A similar emphasis on the value of test scores can be seen whenever educational reform and accountability are discussed by the mass media. Except for the most limited English proficient (LEP) students, the language factor in test results seems to be dismissed. Standardized test scores are still the primary indicators of the success of schools and students, and this trend shows no immediate indication of major change. Fred M. Hechinger, encouraged by the work of many advocates of equitable testing, prophesied in a 1988 column for the *New York Times* that "an era is ending. Dissatisfaction with the old order of testing has prepared the ground for change. The banner of the new order is in place. Attempts at standardized packaging of everything ... seem to have led to a counter-revolution against standardized testing" (p. 1). But, less than three years later, President Bush unveiled his blueprint for educational reform with plans for standardized testing on a national level at the fourth, eighth, and twelfth grades. If his plan is implemented, the results of standardized tests will have greater impact than ever before. According to *U.S. News and World Report* (May 6, 1991), the Bush administration calls for scores to be used not only as the basis for admission to institutions of higher education and qualification for federal scholarships but even for hiring by businesses. In an attempt to ensure accountability for learning, the administration is calling for an assessment system which does not accommodate language differences, whose tests cannot accurately measure the achievement of minority language students, and which could have a devastating effect on their futures since low scores are interpreted

to mean lack of ability.

Cummins (1984), during the course of his extensive research of minority language students, found that the majority of teachers, school psychologists, and policy makers considered two years to be sufficient for LEP students to acquire English proficiency. Since students were usually able to converse fluently face to face in that amount of time, it was assumed that subsequent poor academic performance must be due to cognitive or cultural deficiencies. However, his examination of data forced him to draw a strong contrast between what he called basic interpersonal communicative skills (BICS) and cognitive/academic language proficiency (CALP), two different levels of language proficiency. While the first referred to the proficiency required in everyday communication, the second referred to the "manipulation of language in decontextualized academic situations" (p. 137), precisely the kind of language proficiency measured on standardized tests. He learned, in fact, in analyzing data from the Toronto Board of Education on over 1200 immigrant students, that children who arrived in Canada at ages 6-7 took between five and seven years to reach grade level norms in English verbal-academic skills which led to the following condition:

The failure of school personnel to distinguish between the development of conversational or surface fluency and cognitive/academic aspects of English proficiency can result in low academic performance being attributed to deficient cognitive or personality traits within minority language students. (p. 136)

Thus, when LEP students who have achieved basic communications skills in SAE score poorly on tests, they are assumed to have acquired full command of the language, and the low scores are attributed to student deficiencies.

For minority dialect students, the language factor can create even greater difficulties than it does for LEP children. Because these students are born in the United States and speak English as their first language, it is assumed that language should not be a problem, and their language differences are seen, not as variant dialects, but as incorrect usage of standard English. Lindfors (1987) points out that "these less prestigious dialects are often regarded not as intact linguistic systems (which they are), but as error-ridden, garbled, inadequate attempts of a group of people to use regional standard dialect (which they are not)" (p. 397) with the result that not only is the students' language regarded as deficient but also their cultural background and cognitive development. Among these students, Blacks have been predominantly studied, but the same issues pertain to Hispanics, members of rural communities, and any other group using a dialect of English clearly different from SAE.

From a linguistic point of view, it would be as appropriate to say that mainstream American children speak an incorrect version of Black English (BE) as to say that black students speak improper English. Gleason (1989) reminds the educator that "the form of a language that becomes the standard typically is not selected by virtue of any inherent linguistic factors, but because it is used by people with higher social status" (p. 334-335). She adds that, to linguists, all forms of a language are equally valid, and that a child's competence cannot be rated on the basis of dialect. Fromkin and Rodman (1988), in examining the group of dialects spoken by American blacks and Hispanics, discovered that speakers followed consistent grammatical patterns systematically different from SAE. They also noted that most speakers of Chicano English or BE were bidialectal, changing dialects according to the social occasion. Thus, they pointed out, "the use of these dialects...is clearly not evidence of language deviance but of language expertise" (p. 271).

Perhaps the most famous linguist to deal with the issue of BE and other dialects is William Labov. Labov (1972) extensively studied urban Black children and found that, contrary to popular views, Black children received considerable verbal stimulation, lived in a highly verbal culture, had the same capacity for conceptual learning, and used the same logic as speakers of standard English. He thoroughly debunked the myth of verbal deprivation which has so often been used to account for low achievement rates among minority students. In its place, he called for the teaching of SAE as a second dialect to minority students, much as English is taught to LEP students (Samuda, 1973). His work, along with that of other dedicated researchers, indicates clearly that for minority dialect students, as for LEP students, language must be considered in evaluating test results. As Samuda writes,

If Labov and the Baratzes are right, then it follows that standardized IQ tests do not, and cannot as they presently exist, measure the true potential of black children whose language and life styles are largely determined by the conditions of the ghetto. For such tests depend heavily on vocabulary and language usage which place the minority child at a distinct disadvantage.
(p. 9)

Assessment

What is the place, then, of standardized tests in assessments of minority language students? The Encyclopedia of Educational Research (Mitzel, 1982) defends intelligence tests for minority language students while admitting that

test content is drawn from the experience and language background of white, middle class subjects:

To the degree that the experiences of individuals differ from this average expectable background, the tests will be biased as measures of native endowment. They may be quite sensitive measures, however, of relative acculturation. It is hard to see what genuine social ends are served by banning tests that identify children who need special help, just because the tests do not support egalitarian beliefs. (p. 940)

This view is a common one among educators. The companion view is that, while tests may not indicate capacity to learn, they do measure a student's current abilities.

The fallacy in these views is two-fold. First, a cultural and linguistic background which differs from the "average expectable" background, which is not sufficiently acculturated into the mainstream, is not an indicator that a child needs "special help," and it is outrageous to view it as a deficiency to be remediated. On the contrary, it should be viewed as a resource for enriching the school environment as a whole. Second, while test results may indeed indicate a child's ability to perform in the context-reduced setting of the test, they do not predict how a child will perform in a classroom which provides an environment rich in language and contextual learning. At most, they predict the environment in which the child is likely to fail.

The most damaging use of standardized tests, then, is the classification of students for remediation of skills. Mercer's California research found that the placement of minority language students in special classes resulted in "irreparable harm and injury" as expressed by the U. S. District Court (Samuda, 1973). Samuda protested against the tracking done in the name of special assistance, pointing out that "it is a well known fact that classes for low achievers offer a severely limited and low quality curriculum in which reading, spelling, and mathematics are reduced to a minimum and the stimulation of a challenging program and higher achieving peers is lacking entirely" (p. 5-8). Stigmatized as slow learners with low expectations from their teachers, the students usually fulfill expectations and fall further and further behind until they leave school with their role in life already defined as a failure.

A valid use of standardized tests, especially achievement tests, may be evaluation of programs, not in the sense that Duval County and the media has done, but to see if lack of success in testing situations is duplicated in the classroom. If a child who has scored poorly in the context-reduced situation of the test is successful in the classroom, that program must be meeting the needs of the child. That success must be duplicated whether or not it achieves high test

scores. Attempting to change the program to fit a test which has been shown to lack validity is absurd.

The issue of assessment, especially for minority language students, is not likely to be resolved. The school population is becoming increasingly diverse, and the difficulty of finding an accurate tool for evaluating minority language students has also increased. Every proposal, from testing in the student's home language to nonverbal tests, has been shown to be inadequate. Rather than despair or revert to familiar methods, it is time for the educational community to admit that no simple solution is to be found. For all children the time for depending on standardized tests for quick evaluations is past. If such testing is required, then scores must be used judiciously. A variety of factors must be called into play whenever a child is to be evaluated: classroom progress as well as classroom performance, cognitive abilities demonstrated in the home community, accomplishments in diverse tasks, and observations from teachers and others who interact with the student. An examination of the types of errors on a test may prove to be of more value in evaluation than any score. As Sigel (1965) noted in studying Piaget's error analysis, "we do not know what may account for the correct response, but analysis of errors could tell us what kinds of difficulties the child is having" (p. 296).

This type of assessment is complex and time consuming, but educators deal with complex individuals. If the system cannot accommodate the accurate assessment of these individuals, then it is the system which must be changed. Educators cannot continue to put children at risk through their own blind acceptance of test scores.

References

- Brigham, C. C. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.
- Conoley, J. C., & Kramer J. J., (Eds.). (1989). *The tenth mental measurements yearbook*. (pp. 768-775). Lincoln, Nebraska: University of Nebraska Press.
- Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. San Diego, CA: College-Hill Press, Inc.
- Fromkin, V., & Rodman, R. (1988). *An introduction to language* (4th ed.) New York: Holt, Rinehart and Winston, Inc.
- Gleason, J. B. (1989). *The development of language* (2nd ed.). Columbus, Ohio: Merrill Publishing Company.
- Goddard, H. H. (1917). Mental tests and the immigrant. *Journal of Delinquency*, 243-277.
- Hakuta, K. (1986). *Mirror of language*. New York: Basic Books.
- Hartle-Schutte, D. (1991). Standardized tests as measures of success. In Kenneth S. Goodman, (Ed.), *The Whole Language Catalogue* (pp. 264). Santa Rosa, CA: American School Publishers.
- Hechinger, F. M. (1988, Dec. 7). About education. *The New York Times*. Reproduced in *FairTest*, December 1988.
- Hilliard III, A. G. (1979, Winter). Standardization and cultural bias as impediments to the scientific study and validation of "intelligence." *Journal of Research and Development in Education*, 12, 47-58.
- Holmes Jr., W. P. (1986), October). Here's how our school board is boosting black student achievement. *The American School Board Journal*, pp. 38-39.
- Labov, W. (1972). Academic ignorance and black intelligence. *Atlantic Monthly*,

pp. 59-67.

Lindfors, J. W. (1987). *Children's language and learning* (2nd ed.). Englewood Cliffs, N. J.: Prentice-Hall, Inc.

Mercer, J. (1971). Institutionalized Anglocentrism: Labeling mental retardates in the public schools. *Race, Change, and Urban Society*, 5, 311-338.

Mitzel, H. E. (Ed.). (1982). *Encyclopedia of educational research*. New York: Collier MacMillan Publishers, 939-941.

Prescott, G., Balow, I., Hogan, T., & Farr, R. (1984). *Metropolitan achievement tests: Directions for administering - primary 1, form L, elementary, form m*. San Diego, CA: Harcourt Brace Jovanovich, Inc.

Ribeiro, J. L. (1980). Testing Portuguese immigrant children: Cultural patterns and group differences in response to the WISC-R. In D. P. Macedo (Ed.), *Issues in Portuguese Bilingual Education*. Cambridge, Mass.: National Assessment and Dissemination Center for Bilingual Education.

Samuda, R. J. (1973, May). Racial discrimination through mental testing. *IRCD Bulletin*, Number 42, 1-12.

Schooling's big test. (1991, May 6). *U.S. News and World Report*, p. 63.

Sigel, I. E. (1965) How intelligence tests limit understanding of intelligence. In Ira J. Gordon (Ed.), *Human Development* (pp. 287-297). Glenview, IL.: Scott, Foresman, and Co.

Sternberg, R. J. (1984, June). Testing intelligence without I. Q. tests. *Phi Delta Kappan*, pp. 694-698.

language students continue to be classified as intellectually deficient on the basis of standardized testing.

Is it appropriate to compare modern assessment with that of Goddard and Brigham? Are minority language students indeed at a disadvantage when tested in standard American English (SAE)? Mercer (1971) determined that four and half times as many Mexican Americans and twice as many blacks were placed in EMR (educable mentally retarded) classrooms than would be expected from their proportional population, and that reevaluation of tests showed 75% of them to be mislabelled. But is language rather than cultural deficits or poor environments a major factor in poor test results? The answers lie in a look at minority language students and the tests they take.

The Tests

Two types of standardized tests are used to assess students: achievement tests such as the Metropolitan Achievement Test (MAT) and intelligence and aptitude tests such as the Stanford-Binet. While achievement tests are purported to measure what a student has learned, and intelligence tests are said to reflect innate abilities, Hartle-Schutte (1991) points out that "much of the current content of IQ tests is indistinguishable from achievement tests, aptitude tests and even developmental screening tests" (p. 264). The content of both types of tests is based heavily on the use and comprehension of language, specifically standard English.

The vocabulary section common to intelligence tests is undoubtedly the most language based. Sternberg (1984) found that vocabulary was the single best predictor of overall IQ scores. While indicating that "there is good reason to believe that vocabulary is such a good measure of intelligence because it measures, albeit indirectly, children's ability to acquire information in context" (p. 697), he also acknowledges that few other tests are more based on achievement than is the vocabulary section of the intelligence tests. If an intelligent child learns vocabulary in context, he will learn the vocabulary present in his own context. As Hilliard (1979) said,

Verbal meanings (definitions) are modified by and are embedded in particular cultural environments. Since words have no intrinsic meaning but derive their meaning from the habits of use by a particular cultural group, it is a gross error to presume that a particular word can have universal meaning to test takers. Therefore, it makes no scientific sense to have a standardized test of vocabulary when there is no such thing as a standard vocabulary. (p. 5)