PERFORMANCE PATTERNS OF BILINGUAL CHILDREN TESTED IN BOTH LANGUAGES

David L. McArthur

CSE Report No. 164
1981

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ABSTRACT

The testing of bilingual students poses particular problems for analyses of performance, item bias, and test adequacy. When children are selected for their facility in two languages, and the same test is administered in both languages, a special arena is provided for the study of these problems. A widely-used test, the Comprehensive Tests of Basic Skills, is available in both English and Spanish. The vocabulary subtest was administered to 1162 second-graders in bilingual education programs throughout the Southwest, as part of a larger study; 58 of those students received both versions of the test because they were deemed equally proficient in both languages. Results show that patterns of performance for these students differ markedly between the two versions, and suggest that the test differs in important dimensions even though the Spanish version is a rather faithful translation of the English original.

INTRODUCTION

Severe problems confront the evaluation of bilingual program students from the standpoint of both individual performance measurement and the potential for bias in testing. Assessing the student in the majority language runs one set of risks; assessing in the native tongue runs another. The number of studies which have successfully assessed a single skill in two languages for the same individuals is exceedingly small (Duran, 1980). Resolution of these problems is not aided by the current controversy surrounding both the definition and measurement of bilingualism itself (De Avila, 1978.) Moreover, thoroughly contradictory findings emerge from studies of the acquisition of French by native English-speaking children in Canada (Lambert & Tucker, 1972), of Swedish by native Finnish-speaking children in Scandinavia (Skutnabb-Kangas & Toukomaa, 1976), and of English by native Spanish-speaking children in the U.S. (Fischer & Cabello, 1978). The integration of such differences may rest in part on linguistic, developmental, and/or sociocultural interpretations (Troike, 1978); a practical level of shared bilingualism or dominance of one language over the other in the community may also play a strong role (Laosa, 1975). Finnish-speaking children from the populous southern districts find, and potentially model, both Finnish and Swedish in almost every shop window, while the politics of separatism are explicit in Quebec and de facto in many areas of the American Southwest, so children from these regions may encounter the second language with mixed emotions. As-

sessing even a relatively simple arena like vocabulary skills becomes multiply compounded when dealing with students who must cope with two languages.

Measuring the skills of bilingual program students also means assessing whether tests developed for the monolingual-English student are appropriate for making decisions about bilingual or limited-English proficient students characteristically found in such programs, and of minority groups who tend to be overrepresented there. Some educators believe that many tests are intrinsically unfair to minorities because the values they reflect are those of the majority only (Cervantes, 1975). Others, however, hold that tests of culturally defined content and vocabulary are not biased because achievement itself is language and culture specific (Ebel, 1975). But the impetus for testing continues:

> The problem now becomes not whether to test bilingual students, but rather how to do it in a manner that accurately assesses their specific abilities and in a manner that does not create a bias either against them or in favor of them (Cooper, 1978, p. 2, italics original).

We turn attention specifically to assessment in Spanish-English bilingual programs at the primary level, and encounter two factors which strongly mitigate against simple solutions to the problems noted above. The first is that exceedingly few instruments are available at present which are both culturally appropriate and technically sound for this purpose. "The problems are particularly acute with respect to English language measure, but are often equally pervasive in instruments

that are simply translations from English language versions" (Burry, 1979, p.8). The second is that English-language instruction in reading, listening comprehension, and vocabulary may be intrinsically more difficult for Spanish-speaking children than for their native English-speaking counterparts because of the increased rhythmic and phonological complexity of English. Fundamental linguistic skills for understanding Spanish are frequently inadequate for comprehending English. Even a relatively simple phrase like "I c'n take it home fer ya," (/âykn̥tèykťthówmfťryť↓/ for the English listener) is likely to be heard by the native Spanish-speaking child as /'aintekrómfia↓/, resulting in the obliteration of six out of seven words in the sentence (Matluck & Mace, 1972). The quantity of purely linguistic differences between Spanish and English suggest that the Spanish-speaking child is at no small disadvantage; especially in the primary grades, appropriate language skills testing must not ignore such difficulties.

The Comprehensive Tests of Basic Skills/Spanish (CTBS, 1974/1978), is in large measure a direct translation of its English counterpart, which has been widely used as a primary skills evaluation tool. The CTBS/S has been presented as a major attempt to meet the needs of native Spanish-speaking children (Finch, 1979). With such a test, the teacher can select the language appropriate for a child with some assurance that the instrument is valid, reliable, and unbiased (Hoepfner & Christen, 1979). Thus, the CTBS and CTBS/S

should provide a good vehicle to examine individual performance patterns in either language for students in bilingual programs. However, recent evidence based on the performance of English- and Spanish-speaking pupils suggests that the tests contain multiple sources of bias (McArthur, 1981), so a particularly interesting situation for research obtains when both versions of the CTBS are administered to the same children. That is, if a group of children who possess similar levels of knowledge in both English and Spanish are tested on both instruments, will individual performances be the same across the two? Will the results of such dual language testing reflect patterns which can be interpreted as the direct result of item bias? Will direct translation hold up as a viable strategy for fair testing of primary pupils in Spanish as well as English?

METHODS

Subjects

As part of a larger study (CSE, 1979), almost 1200 children in bilingual education programs in 26 school districts spread over five southwestern states were administered a series of educational achievement tests by their teachers. Programs were designed to provide instruction in reading and mathematics at the upper primary level. Teacher reports from these programs indicate that the time spent using Spanish as the language of instruction was approximately equal to the time spent using English. Ninety-three percent of

the program teachers had earned at least a BA or BS; 94% were full-time employees of the school district, and 88% had prior experience in bilingual education. Assignment of students to these special programs relied primarily on teacher evaluations and language dominance tests. Achievement tests were infrequently used to determine remediation placement, and intelligence test scores were generally excluded altogether from placement considerations. Thus the programs represented a major effort, competently staffed, to provide special attention in a bilingual setting to student educational needs. Most of the students were rated by their teachers as having some skills in both English and Spanish. Overall only one child in ten from these classes was considered monolingual Spanish while only one in nine was rated as monolingual English.

Instruments

While a large number of instruments were used in the investigation of programs, only the CTBS is of concern in the present study. It was selected because test content between the two language versions is virtually identical. The CTBS-Spanish was the first test by a major publisher to be subjected to a four-step editorial procedure designed to reduce bias; included were studies of content validity, application of editorial guidelines in item construction, reviews for bias, and separate ethnic group pilot studies. The developers of the Spanish-language version tried to keep the test content and measurement features intact, thus building a test which

was similar in rationale, administration, and interpretation to its parent version in English. What differences exist are the result primarily of problems of literal translation.

The children in the study were given a large number of standardized tests of achievement during the course of the regular school year by their teachers. With regard to the CTBS, the important instruction made to teachers was that they decide in advance on an individual basis whether each child would receive the English-language or Spanish-language version of the test. This decision was left totally to the discretion and best judgment of the classroom teachers. A total of 1162 completed test forms were returned, 814 in English and 348 in Spanish. Fifty-eight students in the sample were found to have been tested in both languages; that is, one student in every nineteen was given both forms of the test because the teachers felt unable to distinguish in advance which language these students should be tested in. No evidence is available to suggest that any selection bias or other external circumstance might have contributed to obtaining this sample. Order of administration was apparently random. For purposes of this report, only the Vocabulary subscale of test level C, consisting of 33 items selected in response to the teacher's verbal directions, is considered.

Methods of analysis

Two techniques for analysis of response patterns were utilized in this study. The first relies on the work of Sato (1980) and colleagues in Japan; they have generated a systematic method

of appraisal of test performance based on the S-P (Student-Problem) Chart, a matrix of right and wrong answers, coded 1 or 0, for each respondent for each item. The N x n matrix has the additional characteristics that students have been sorted by descending total score and items have been sorted by increasing difficulty. Thus the top row of the S-P Chart is a representation of the pattern of correct and incorrect responses to this sample of items by the most capable student in the group, the bottow row by the least capable. The left-hand column shows the pattern of responses to the easiest item in the set of items, and right-hand column shows the most difficult. From this matrix are generated two statistics, one related to the group pattern for the group as a whole, the other related to individual performance vis-á-vis both the group and the configurationof items, for each individual. The first is an "index of discrepancy," D*, which ranges from 0.00 for a matrix of perfect symmetry between student capabilities and item difficulties, to 1.00 for a matrix representing exclusively random responding.[1] The second is a "caution index," c_i, which ranges from 0.00 for an individual whose response pattern is perfectly fitted to that reflected in the order of item

---

1. $D^* = \frac{A\ (N,n,\bar{p})}{A_B\ (N,n,\bar{p})}$

where the numerator is a discrepancy between cumulative probability ogives obtained from the S-P chart, and the denominator is an analogous discrepancy as modeled by cumulative binomial distributions, both with the same number of cases, number of items, and average passing rate. (Sato, 1980).

difficulties as determined by the group, to 1.00 for an individual whose pattern of responses is totally antithetical to the order of item difficulties, and thus is quite unlike the representative average respondent in the group.[2]

The second analytic tool used in this study is a statistic from Goodman and Kruskal called lambda, which has been applied elsewhere to the detection of differences in response patterns in testing (Veale & Forman, 1976). Here the focus is on differences between groups in the attractiveness of incorrect responses within the multiple-choice format of one correct and three incorrect responses per item. Lambda is an index of the pattern of choice for the incorrect responses. If the value of lambda is 0.00, the two groups use about the same pattern of selection of the incorrect responses. As the value increases, one group is using a different strategy for selection of incorrect responses than the other. The computation of lambda is independent of the actual proportions within each group who select the correct response to the item. In this paper, values of lambda above .10 are considered noteworthy.[3]

---

2. $$c_i = 1 - \frac{\mathrm{cov}(x_{ij}, Y_j)}{\mathrm{cov}(u_{ij}, Y_j)}$$

where the numerator is the covariance over problems of the i-th student's score on the j-th problem with the number of students who correctly answer that j-th problem, and the denominator is the covariance over problems of the i-th ideal student's score on the j-th problem with the number of students who correctly answer that j-th problem (Sato, 1980).

3. $$\lambda = \frac{\Sigma \mathrm{max.f}_{jk} - \mathrm{max.f}_{.k}}{N - \mathrm{max.f}_{.k}}$$

where $\mathrm{max.f}_{jk}$ is the larger frequency of the two groups for any single wrong choice, $\mathrm{max.f}_{.k}$ is the larger marginal frequency of the two groups across all wrong choices, and N is the total number of observations.

Details of the computation and use of these approaches in the context of testing and item bias detection research have been set out elsewhere (McArthur, 1981). The usual test-retest and reliability statistics are not appropriate here, because of the attention to deciphering specific performance patterns rather than whole-group performance.

Hypotheses

Because of process of respondent selection, specific hypotheses about their performance on the English-language and Spanish-language versions of the Vocabulary subtest were, first, that the achieved scores between tests would be perfectly correlated. Additionally, the S-P charts for the two versions would be similar, as shown by equal indices of discrepancy, D*. At the level of the individual respondent, it was hypothesized that the achieved total score in English would equal the achieved total in Spanish, and that the caution index generated for each individual in the English-language S-P chart would be equal to the caution index obtained by the same individual from the Spanish-language S-P chart.

RESULTS

Total scores on the English-language Vocabulary subtest averaged 75.34% correct with a range of 6 - 33. On the Spanish-language version, the average was 37.56% correct with a range of 4 - 25. The total scores are significantly ($p<.05$) correlated, $r = .48$. Median improvement from Spanish to English is

13 answers correct. Only three of the 58 participants did not show improvement in their total scores from Spanish to English.

Two of the 33 items yielded higher percentages of correct responses in the Spanish-language version than in the English. For the remainder of the items, students were able to select the correct response less frequently in the Spanish-language version, often by substantial margins. The ratio of Spanish correct to English correct for each item is shown in the first column of Table 1. The consistency with which students picked the correct

---

Insert Table 1 about here

---

answer in both languages ranged from moderately high (65% of the respondents chose the correct answer to item 8 in both languages) to very low (only 7% chose the correct response to item 31 in both languages). The consistency of selection of incorrect responses was generally extremely low, reaching 14% for items 24 and 31. The proportions of joint correct and joint incorrect proportions are shown in columns 2 and 3 of Table 1.

Those incorrect answers to items which garnered at least 10% more responses than the next most frequently chosen incorrect response were termed "popular distractors". Three popular distractor items were found in the English-language version, while twelve were found in the Spanish. The average percentage of respondents who chose the correct answer to an item in English but were swayed to choose the popular distractor (incorrect) response to that same item in Spanish was 35%. The reverse, choosing a popular distractor response in English although selecting the correct response to that same item in Spanish was 30%. Whether a specific item contained a popular distractor, and if

so the percentage of respondents correct on the same item in the other language but who chose that popular distractor, is indicated in the next four columns of Table 1.

The data to this point quite clearly indicate that the Spanish-language version of the CTBS presented a far more difficult task for these respondents than did the English-language version. Only infrequently did any vocabulary item from one version have both an equal percentage of incorrect selections. Examination of the S-P charts is necessary to show whether the difference in performance patterns is systematic.

The Spanish-language version generated a D* of .53, a relatively high level of randomness of responses, while the English-language version yielded a D* of .24, reflecting a much more orderly fit of subject capabilities to item difficulties. No exact test of significance exists for the size of, or differences between, D* values, but in this instance they represent configurations of the S-P charts which are distinctly different visually. The difference is supported by reference to the caution indices which for individual respondents to the English-language version averaged .17, but to the Spanish-language version .25. That is, on average the respondents were more consistent in selecting correct answers to easy items and incorrect answers to difficult items in the English-language version. In fact, the number of respondents with caution indices of 0.00 is much higher in English. Of particular interest is that the correlation between the two indices computed across the 58 participants is nonsignificant. Changes in caution indices from

one language version to the other are uncorrelated.

The computation of lambda, which details differences in selection patterns for wrong answers, showed that twelve out of 33 items had large discrepancies in the obtained configuration. That is, for a large number of items, the respondents shifted their choice from one incorrect answer to another across language versions, rather than picking the same incorrect responses on both occasions. The last column of Table 1 indicates those items with such shifts in incorrect answers.

DISCUSSION

The findings of this study in general comport with earlier research on the CTBS in English and Spanish using independent groups of bilingual program respondents (McArthur, 1981). The distributions of total subscale scores, the higher D* indices for the Spanish-language version, and the number of popular distractors and of lambda values exceeding .10 are all similar. That the two versions of the test do not produce equal outcomes even when the actual respondents are identical seems clear from the present data. If there was to have been equivalence of total subscale scores, of group or individual patterns of correct scores, or of selection of wrong answers between the English- and Spanish-language versions, the number of discrepancies emerging from the statistical computations would have been far smaller. In its present configuration, these data suggest that children do not show the same performance patterns in response to the two versions of the test. Review of data contained in Table 1 suggests that many of the items may be suspected of somehow biasing the

choice of correct response, and that such potentially biasing items are more prevalent in the Spanish-language version.

The relatively small number of individuals represented in this study makes these results necessarily tentative: they are presented neither as a representation of majority vs. minority responses to a specific test, nor as an indication in any way of a measure of true ability among bilingual program students. Rather, the unusual trial of a purportedly decent test in two languages, a purportedly equal-ability student sample, and a classroom experience for that sample equally divided into the use of English and Spanish, demands thoughtful attention to the appraisal of testing. In the present investigation, one weakness is the absence of an independent and unambiguous assessment of bilingual capability, and the ensuing reliance on the accuracy of teacher selection of students equally competent in two languages. DeAvila and Duncan (1978) have pointed out numerous shortcomings in teacher ratings of language competence. However, for this study, students were not drawn for their equally high abilities or for the purposes of assembling a homogeneous sample, but only for their language abilities to be equally high or low in both languages. Nothing is known about the relative levels of exposure to English or Spanish outside the school, nor about the relative strengths and weaknesses of the texts in both languages used in the program. However, the teachers' close personal supervision of students and the even division between English and Spanish as the language of instruction in these programs suggest that the childrens' levels of

readiness for vocabulary would be roughly similar. Another weakness is the relatively small number of items included in this investigation. However, the CTBS appears to represent the state of the art in English/Spanish testing of vocabulary skills at this level, and no other instrument is known to be a closer approximation to neutrality. The present results support the contention that the method of direct translation from English to Spanish for bilingual vocabulary testing may not be fully adequate for the needs of the bilingual program student.

Table 1

Summary of Findings for the CTBS and CTBS/S

Item number	Ratio of Spanish correct to English correct	Percent joint correct	Percent joint wrong	Popular distractor		Percent who move from correct in one language to popular distractor in the other		Lambda greater than .10
				English	Spanish	S to E	E to S	
1	.45	42	0	--	--	--	--	yes
2	.32	32	4	--	yes	--	32	--
3	.64	47	0	--	--	--	--	--
4	.63	60	2	--	--	--	--	--
5	.60	42	2	--	--	--	--	yes
6	.74	56	4	--	--	--	--	yes
7	.62	47	9	--	yes	--	17	--
8	.73	65	0	--	yes	--	12	--
9	.64	46	0	--	--	--	--	yes
10	.37	33	2	--	--	--	--	--
11	.29	25	0	--	--	--	--	--
12	.37	30	0	--	--	--	--	--
13	.25	19	4	--	--	--	--	--
14	.53	35	4	--	yes	--	31	yes
15	.46	30	9	--	--	--	--	--
16	.11	9	7	--	yes	--	41	yes
17	.13	9	4	--	yes	--	54	--
18	.67	54	4	--	--	--	--	--
19	.23	18	9	--	yes	--	40	--
20	.63	49	2	--	--	--	--	--
21	.22	16	5	--	--	--	--	--
22	1.06	37	9	--	--	--	--	yes
23	.77	44	5	--	--	--	--	--
24	.55	19	14	--	--	--	--	--
25	.53	23	5	--	--	--	--	--
26	.23	12	9	--	yes	--	43	yes
27	.51	33	4	--	--	--	--	yes
28	.37	12	4	--	--	--	--	--
29	.36	5	9	yes	yes	56	60	yes
30	.51	30	7	yes	yes	5	28	--
31	.40	7	14	yes	yes	30	48	yes
32	.70	42	9	--	yes	--	19	yes
33	1.41	12	11	--	--	--	--	--

REFERENCES

Burry, J. Evaluation in bilingual education. Evaluation Comment, 1979, 6, 1-14.

Cervantes, R.A. Self-concept, locus of control, and achievement in Mexican-American pupils. Unpublished doctoral dissertation, Union Graduate School-West, San Francisco, 1975.

Cooper, E. Test selection in bilingual education evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Comprehensive Tests of Basic Skills (CTBS) Examiner's Manual and Espanol Examiner's Manual. Monterey, CTB/McGraw-Hill, 1974/1978.

Center for the Study of Evaluation: Final Report: Basic Skills Learning Centers evaluation. Los Angeles, UCLA, 1979.

DeAvila, E., & Duncan, S.E. Definition and measurement, the east and west of bilingualism. Larkspur, California, DeAvila, Duncan & Associates, mimeo, 1978.

Duran, R.P. Bilinguals' skill in solving logical reasoning problems in two languages. Princeton: Educational Testing Service, 1980. (ERIC Document Reproduction Service No. ED 198 724).

Ebel, R.L. Constructing unbiased achievement tests. Paper presented at the National Institute of Education Conference on test bias, Baltimore, 1975.

Finch, F.L. At last: a Spanish version of CTBS. Paper presented to the California Association of Bilingual Educators, Fresno, 1979.

Fischer, K.B., & Cabello, B. Predicting student success following transition from bilingual programs. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Hoepfner, R., & Christen, F. Measures of academic growth. Santa Monica: System Development Corporation, mimeo, 1979.

Lambert, W.W., & Tucker, G.R. Bilingual education of children: The St. Lambert experience. Rowley, Massachusetts: Newbury House, 1972.

Laosa, L.M. Bilingualism in three United States Hispanic groups: Contextual use of language by children and adults in their families. Journal of Educational Psychology, 1975, 67, 617-627.

Matluck, J.H., & Mace, B.J. Language characteristics of Mexican-American children: Implications for assessment. Journal of School Psychology, 1973, 11, 365-386.

McArthur, D.L. Detection of item bias using analyses of response patterns. CSE Report No.163 . Center for the Study of Evaluation, University of California, Los Angeles, 1981.

Sato, T. The S-P chart and the caution index. NEC (Nippon Electric Company, Japan), Educational Informatics Bulletin, 1980.

Skutnabb-Kangas, T., & Toukomaa, P. Teaching migrant children's mother tongues and learning the language of the host country in the context of the socio-cultural situation of the migrant family. Helsinki, Finnish National Commission for UNESCO, mimeo, 1976.

Troike, R.C. Research evidence for the effectiveness of bilingual education. Washington, D.C., National Clearinghouse for Bilingual Education, mimeo, 1978.

Veale, J.R., & Forman, D.I. Cultural variation in criterion referenced tests: A global item analyses. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.