DOCUMENT RESUME

ED 389 721 TM 024 221

AUTHOR Dorans, Neil J.; And Others
TITLE The Standardization Approach to Assessing Differential Speededness.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-88-31
PUB DATE May 88
NOTE 24p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Asian Americans; Black Students; *Ethnic Groups; *Evaluation Methods; Hispanic Americans; *Item Bias; Item Response Theory; *Racial Differences; *Test Wiseness
IDENTIFIERS Omitted Responses; Scholastic Aptitude Test; *Speededness (Tests); *Standardization

ABSTRACT

An adaptation of the standardization approach to assessing differential item functioning that applies to all item responses, including omits and not reached, is described. Applications of this method to evaluate differential speededness show that there is evidence of differential speededness for Blacks and Hispanics, but not for Asian Americans. Data from a study by A. P. Schmitt and C. A. Bleistein (1987) and from a recent form of the Scholastic Aptitude Test were used. There may be a dependency between differential speededness and test section location. Differential speededness may be more noticeable when the test section is located at the beginning of a test. Implications of these findings for evaluations of content-related differential item functioning and on differential test-taking strategies are described. (Contains five figures and eight references.) (Author/SLD)

RESEARCH REPORT

THE STANDARDIZATION APPROACH TO ASSESSING DIFFERENTIAL SPEEDEDNESS

Neil J. Dorans
Alicia P. Schmitt
Carole A. Bleistein

Educational Testing Service
Princeton, New Jersey
May 1988

THE STANDARDIZATION APPROACH TO ASSESSING DIFFERENTIAL SPEEDEDNESS

Neil J. Dorans
Alicia P. Schmitt
Carole A. Bleistein

College Board Statistical Analysis
Educational Testing Service

May 1988

Abstract

An adaptation of the standardization approach to assessing differential item functioning that applies to all item responses, including omits and not reached, is described. Applications of this method to evaluate differential speededness show that there is evidence of differential speededness for Blacks and Hispanics, but not for Asian-Americans. There may be a dependency between differential speededness and test section location: Differential speededness may be more noticeable when the test section is located at the beginning of the test. Implications of these findings for evaluations of content-related differential item functioning and on differential test-taking strategies are described.

THE STANDARDIZATION APPROACH TO ASSESSING DIFFERENTIAL SPEEDEDNESS

The standardization approach to assessing differential item functioning (DIF), which is described in detail in Dorans (1987) and Dorans and Kulick (1983, 1986) for the analysis of the correct answer or keyed response, is readily adapted to all responses, including omits and not reached[1]. Schmitt and Bleistein (1987), in their analysis of the performance of Blacks on Scholastic Aptitude Test (SAT) analogy items, used the standardization method to examine DIF on distractors. In the process, they uncovered the phenomenon of differential speededness, i.e., differential response rates between focal group members and matched base group members to items appearing at the end of a section of a test. In the present paper, a description of the standardization approach to DIF as it generalizes to apply to all item options, including non-response, is presented. Then, data from the Schmitt and Bleistein (1987) study and a recent form of the SAT are used to illustrate how standardization uncovers phenomena such as differential speededness.

[1] When a candidate does not respond to an item, but responds to subsequent items, the non response to that item is referred to as an omit. If the candidate does not respond to any of the subsequent items, then the first non response and the subsequent non responses are characterized as "not reached".

Standardization and the Keyed Response

In the traditional standardization analysis, an item is said to exhibit differential item functioning when the probability of correctly answering the item is lower or higher for examinees from one group than for equally able examinees from another group. The focus of DIF analyses is on differences in performance between groups that are matched with respect to the ability, knowledge or skill of interest. The basic elements of a standardization analysis of the keyed response are proportions correct at each level of a matching variable, such as total score, in a base or reference group and a focal or study group. Plots of these conditional proportions correct against score level in the focal and base groups provide a visual indication of the extent of DIF that an item exhibits. A plot of differences in conditional proportions correct between the focal and base group portrays the degree of DIF more directly. In addition to these plots, standardization provides numerical indices for quantifying DIF.

The prime numerical DIF index that standardization computes is the standardized p-difference, which is defined as

$$(1) \qquad DSTD = \Sigma\{W_s[P_{fs} - P_{bs}]\} / \Sigma\{W_s\},$$

where $[W_s / \Sigma\{W_s\}]$ is the weighting factor at score level s used to weight differences in the proportions correct between the focal group (P_{fs}) and the base group (P_{bs}), and Σ is the summation operator

which sums these weighted differences across scores levels to arrive at $DSTD$, an index that can range from -1 to +1. Positive values of $DSTD$ indicate that the item favors the focal group, while negative $DSTD$ values indicate that the item disadvantages the focal group. $DSTD$ values between -.05 and +.05 are considered negligible. $DSTD$ values between -.10 and -.05 and between .05 and .10 are inspected to insure that no possible effect is overlooked. Items with $DSTD$ values outside the {-.10, +.10} range are more unusual and are examined very carefully.

The weights, $[W_s / \Sigma\{W_s\}]$, are the essence of the standardization approach. First, note that a common weight is applied to both P_{fs} and P_{bs}. This contrasts with what occurs in the computation of impact,

$$(2) \qquad IMPACT = P_f - P_b = \Sigma\{N_{fs}P_{fs}\}/\Sigma\{N_{fs}\} - \Sigma\{N_{bs}P_{bs}\}/\Sigma\{N_{bs}\},$$

where N_{fs} and N_{bs} are the frequencies of score level s in the focal and base groups. In addition, the particular set of weights employed for standardization depends upon the purposes of the investigation. Some plausible options are the following:

- $W_s = N_{ts}$, the number of examinees at s in the total group;
- $W_s = N_{bs}$, the number of examinees at s in the base group;
- $W_s = N_{fs}$, the number of examinees at s in the focal group;

or ➡ W_s = the relative frequency at s in some reference group.

In practice, $W_s = N_{fs}$ has been used because it gives the greatest weight to differences in P_{fs} and P_{bs} at those score levels most frequently attained by the focal group under study. Use of N_{fs} means that $DSTD$ equals the difference between P_f, the observed performance of the focal group on the item, and P_f, the imputed performance of selected base group members who are matched in ability to the focal group members.

Standardization and All Response Options

The generalization of the standardization methodology to all response options including omission and not reached is straightforward. It is as simple as replacing the keyed response with the option of interest in all calculations. For example, a standardized response rate analysis on option A would entail computing the proportions choosing A (as opposed to the proportions correct) in both the focal and base groups,

$$(3) \qquad P_{fs}(A) = A_{fs}/N_{fs};\ P_{bs}(A) = A_{bs}/N_{bs},$$

where A_{fs} and A_{bs} are the number of people in the focal and base groups, respectively, at score level s who choose option A. The next step is to compute differences between these proportions,

$$(4) \qquad D_s(A) = P_{fs}(A) - P_{bs}(A).$$

Then these individual score level differences are summarized across score levels by applying some standardized weighting function to these differences to obtain $DSTD(A)$,

$$(5) \qquad DSTD(A) = \Sigma\{W_s[P_{fs}(A) - P_{bs}(A)]\} / \Sigma\{W_s\},$$

the standardized difference in response rates to option A. In a similar fashion one can compute standardized differences in response rates for options B, C, D, and E, and for non-responses as well.

Differential Speededness

Application of the standardization methodology to counts of examinees at each score who did not reach the item culminates in a standardized not-reached difference,

$$(6) \qquad DSTD(NR) = \Sigma\{W_s[P_{fs}(NR) - P_{bs}(NR)]\} / \Sigma\{W_s\}.$$

For items at the end of a separately-timed section of a test, these standardized differences provide measurement of the differential speededness of a test. Differential speededness refers to the existence of differential response rates between focal group members and matched base group members to items appearing at the end of a section. Schmitt and Bleistein (1987) found evidence of this phenomenon for Blacks, as compared to a matched group of Whites, on

analogy items. Schmitt and Dorans (1987) reported that this effect was also found for Hispanics. In the balance of this paper, differential speededness results for Black, Hispanic and Asian-American focal groups, compared to a White base or reference group, are presented and their implications are discussed.

Figure 1 depicts standardized differential not-reached rates on the last ten items of the two verbal sections of the November 1983 form of the SAT that were observed for Blacks, Mexican-Americans, Puerto Ricans and Asian-Americans. For the purposes of this paper cross-group comparisons were made cautiously because different standardization weights were used for each ethnic group. It is evident from Figure 1a that Blacks reach the last ten items of the 45-item Verbal 1 section at a lower rate than a matched group of White examinees. The standardized differential not reached rate, $DSTD(\mathrm{NR})$, for Blacks hovers around .05 for all ten items. In contrast, the $DSTD(\mathrm{NR})$ values for Asian-Americans on these same ten items are close to zero, indicating the absence of differential speededness. The $DSTD(\mathrm{NR})$ rates for the two Hispanic groups are closer to the Black rates than the Asian-American rates, indicating that differential speededness exists for Hispanics as well as Blacks, but not for Asian-Americans.

Insert Figure 1 about here

Figure 1b depicts the standardized differential not reached rates for the last ten items of the 40-item Verbal 2 section. In contrast to Figure 1a, the differential speededness phenomenon builds up from no effect for any group at item 31 to a clear separation of the groups by item 40. Once again, the effect is most pronounced for the Blacks and non-existent for the Asian-Americans. The effect is minimal for Mexican-Americans, and it approaches the .05 level for Puerto Ricans on the last few items.

Figure 2 depicts standardized differential not reached rates on the last ten items of the two verbal sections of the November 1984 form of the SAT that were observed for Blacks, Mexican-Americans, and Puerto Ricans. Figure 2a depicts the rates for the last ten items on the 45-item Verbal 1 section, while Figure 2b displays the rates for the last ten items on the 40-item Verbal 2 section. All three ethnic groups have standardized differential not reached rates near or above the .05 level on the Verbal 1 section with the Black group having the higher rates and the Puerto Rican group having the lower rates. This pattern is similar to that seen for the November 1983 form with the exception that the two Hispanic groups have exchanged locations on the plot. On the Verbal 2 section in Figure 2b, differential speededness is noticeable only for the Black group and, as in Figure 1b, builds up from near zero on item 31 to over .05 on items 38 to 40.

Insert Figure 2 about here

Figures 3, 4 and 5 depict standardized differential not reached rates on the last ten items of the two verbal sections of the November 1986 form of the SAT that were observed for Asian-Americans, Blacks, and Hispanics[2]. Figure 3a portrays the $DSTD$(NR) rates for the last ten Verbal 1 items when the Verbal 1 section was the first section in the test, while Figure 3b displays the $DSTD$(NR) rates for the same ten items when the Verbal 1 section appeared as the third section of the test. As was the case for the November 1983 form, differential speededness is non-existent for Asian-Americans. For Blacks and Hispanics, the size of the differential speededness effect depends on the location of the section in the test: Differential speededness is more pronounced where Verbal 1 was the first section in the test, as seen in Figure 3a.

Insert Figure 3 about here

Figure 4a contains the rates for the last ten Verbal 2 items when the Verbal 2 section was the fourth section in the test, while Figure 4b contains the rates for the exact same items when the Verbal 2 section appeared as the first section of the test. Once again, no evidence of differential speededness for Asian-Americans exists. As was the case on the November 1983 and November 1984 forms, differential speededness on the Verbal 2 section for Blacks grows from near zero at item 31 to near the .05 level by item 38. This buildup is most evident

[2] Analyses for the November 1986 form combined all Hispanic examinees into one category.

on the second order where the Verbal 2 section appeared as the first section on the test. The Hispanic rates for the two different orders also demonstrate the importance of section location. In Figure 4a, differential speededness is virtually nonexistent for Hispanics; in Figure 4b , evidence of differential speededness exists for Hispanics. Differential speededness is more pronounced where Verbal 2 was the first section in the test. In contrast to what is observed for Blacks and Hispanics, the absence of differential speededness for Asian-Americans seems to generalize across section location.

Insert Figure 4 about here

Figure 5 depicts average differential speededness rates for the Verbal 1 (Figure 5a) and Verbal 2 (Figure 5b) sections across the two orders of test booklets. In these average plots, differential speededness on the Verbal 1 is less pronounced for Blacks than it was in the two earlier November forms, which reflects the lower levels of differential speededness observed when the items do not appear in the first section of the test. (In both November 1983 and November 1984, Verbal 1 was the first section.) One sees in Figure 5 that, once again, differential speededness is non-existent for Asian-Americans.

Insert Figure 5 about here

Implications

There are both methodological and substantive implications of the differential speededness phenomenon.

Methodological

If undetected, the differential speededness phenomenon may produce evidence of differential item functioning that might be misconstrued to be content-related when it is actually a function of item position. In other words, differential speededness may induce differential item functioning on items at the end of a test section simply because those items are at the end of the section. Differential speededness can confound the assessment of content-related differential item functioning, as Schmitt and Bleistein (1987) discovered in their analysis of differential item functioning for Blacks on SAT analogy items. In an attempt to adjust for differential speededness effects, a slightly altered version of the standardized p-difference is computed in which examinees at each score level who do not reach the item are excluded from the analysis. Schmitt and Bleistein (1987) employed this correction and found that it reduced much of the differential item functioning that had been evident on the sets of ten analogy items that appear at the end of the 45-item Verbal 1 section. This finding led the authors to conclude that differential speededness was a major contributing factor to the appearance of differential item functioning for Blacks on SAT analogy items, especially those which appeared towards the end of a section. It is an empirical question

whether the statistical adjustment for differential speededness yields a DIF index that accurately reflects what would be seen under unspeeded conditions. Recent research (Dorans, Schmitt & Curley, 1988) suggests that the correction used by Schmitt and Bleistein mitigates the effect of speed but does not eradicate it.

The existence of differential speededness also has implications for the quality of matching that can be attained. Matching on a total score that is contaminated by differential speededness is likely to influence *DSTD* values in a small but systematic way. Bleistein and Schmitt (1987) found that as the unidimensionality of the matching variable increases, fewer item are flagged for DIF. In order to avoid the artifactual detection of DIF, it may be necessary to devise ways of removing the speed component from the matching score.

Substantive

The existence of differential speededness also has important implications for the advice given to test-takers and for test specifications. It appears that the speededness of the SAT-Verbal sections differentially affects matched groups of Whites and Blacks, and Whites and Hispanics. Differential speededness may be a consequence of differential test-taking strategies employed by different groups. For example, Whites may be more likely to skip over difficult items when confronted with them than would a matched group of Blacks. As a consequence, the matched Whites may be more likely to reach items at

the end of the test than are the Blacks. Indirect evidence for this differential strategy hypothesis can be garnered from examination of standardized differential omit rates, as was done by Rivera and Schmitt (1986) who found that Hispanics tended to omit less than matched Whites.

The degree of differential speededness observed may be partially dependent upon section location. More differential speededness was evident on both verbal sections in the November 1986 data when these sections were the first in the test booklet. One plausible explanation for this location effect may be that the Black and Hispanic examinees as a group have less experience taking tests than the matched group of Whites and consequently may be less adept at pacing themselves early in the test. As a reaction to running out of time on the first section, they may quicken their pace through the later sections of the test and consequently dampening the differential speededness effect.

Differential speededness is not a desirable test property. Its impact on test scores needs to be investigated. Differential speededness is bound to affect test scores when easy items are involved, as is the case for the first items among the the last ten analogy items on the 45-item Verbal 1 section of the SAT, because these easy items are likely to be answered correctly if they are reached. Test specifications need to be reexamined to ascertain what changes can be made to mitigate the impact of differential speededness.

References

Bleistein, C. A., & Schmitt, A. P. (1987, April). *The effect of criterion selection on the evaluation of differential item functioning on SAT analogy items between Black and White examinees.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Dorans, N. J. (1987). Two new approaches to assessing differential item performance: Standardization and the Mantel-Haenszel method. In A. P. Schmitt & N. J. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test.* (RM-87-1). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach.* (RR-83-9). Princeton, NJ: Educational Testing Service.

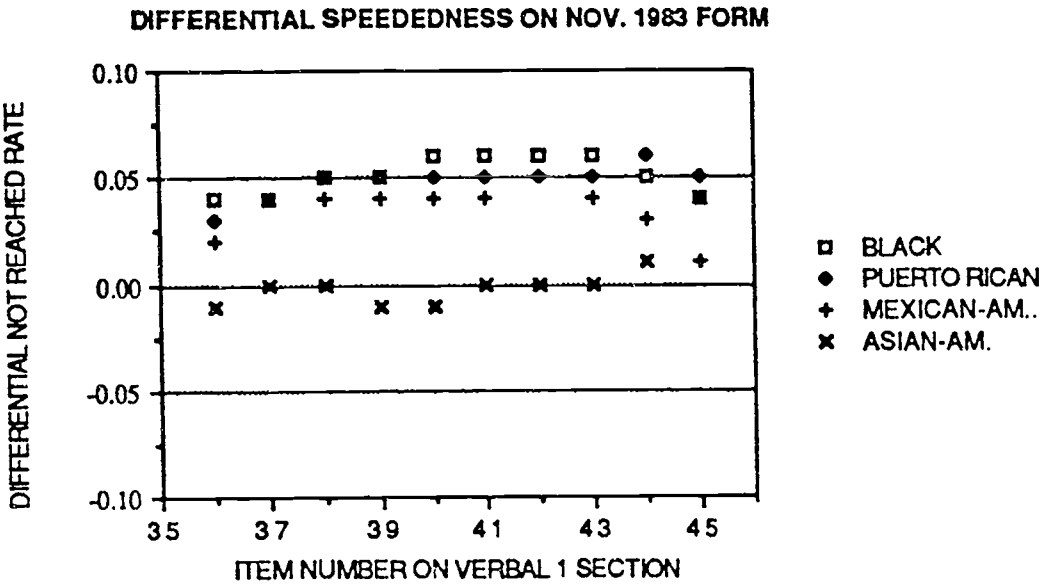
Dorans, N. J., Schmitt, A. P., & Curley, W. E. (1988, April). *Differential speededness: Some items have DIF because of where they are not what they are.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Rivera, C., & Schmitt, A. P. (1986, April). *A comparison of Hispanic and White students' omit patterns on the Scholastic Aptitude Test.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

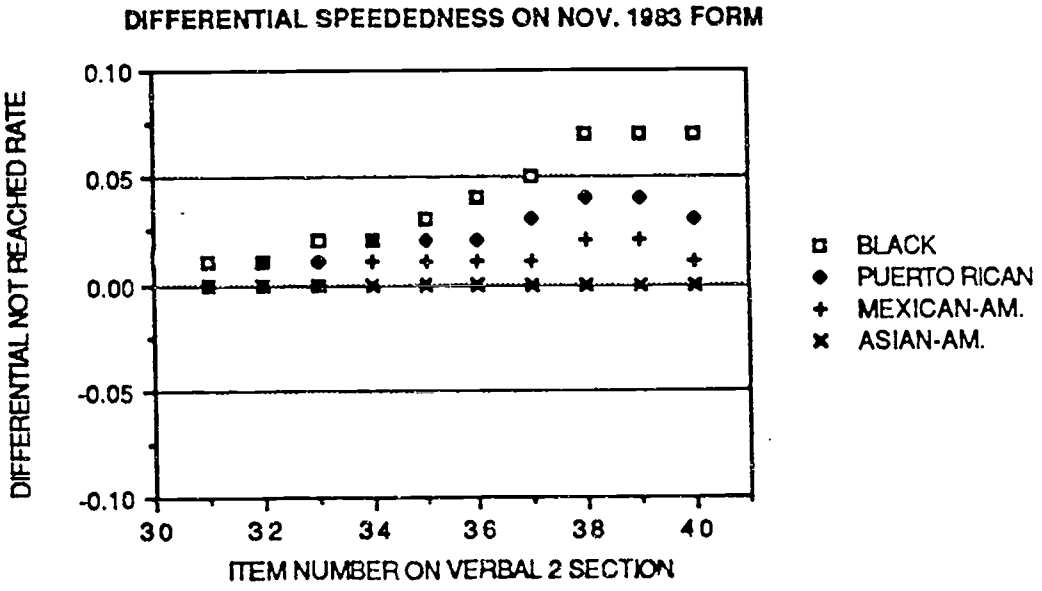
Schmitt, A. P., & Bleistein, C. A. (1987). *Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items.* (RR-87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A. P., & Dorans, N. J. (1987, August). *Differential item functioning for minority examinees on the Scholastic Aptitude Test.* Paper presented at the annual meeting of the American Psychological Association, New York.

FIGURE 1

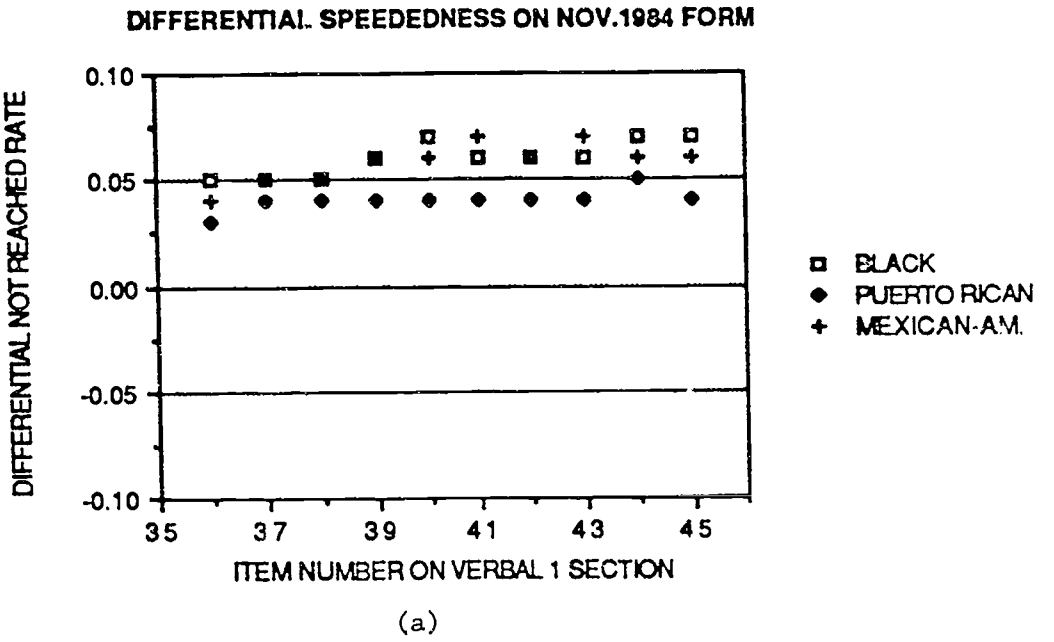

DIFFERENTIAL SPEEDEDNESS ON NOV. 1983 FORM
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
35
37
39
41
43
45
ITEM NUMBER ON VERBAL 1 SECTION
BLACK
PUERTO RICAN
MEXICAN-AM..
ASIAN-AM.

(a)

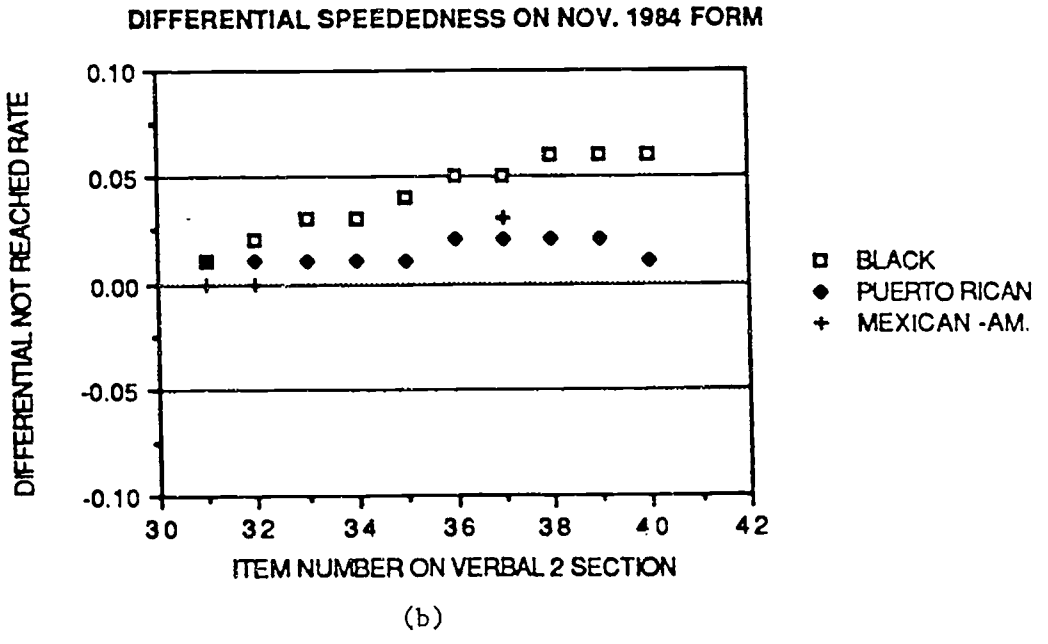

DIFFERENTIAL SPEEDEDNESS ON NOV. 1983 FORM
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
30
32
34
36
38
40
ITEM NUMBER ON VERBAL 2 SECTION
BLACK
PUERTO RICAN
MEXICAN-AM.
ASIAN-AM.

(b)

FIGURE 2

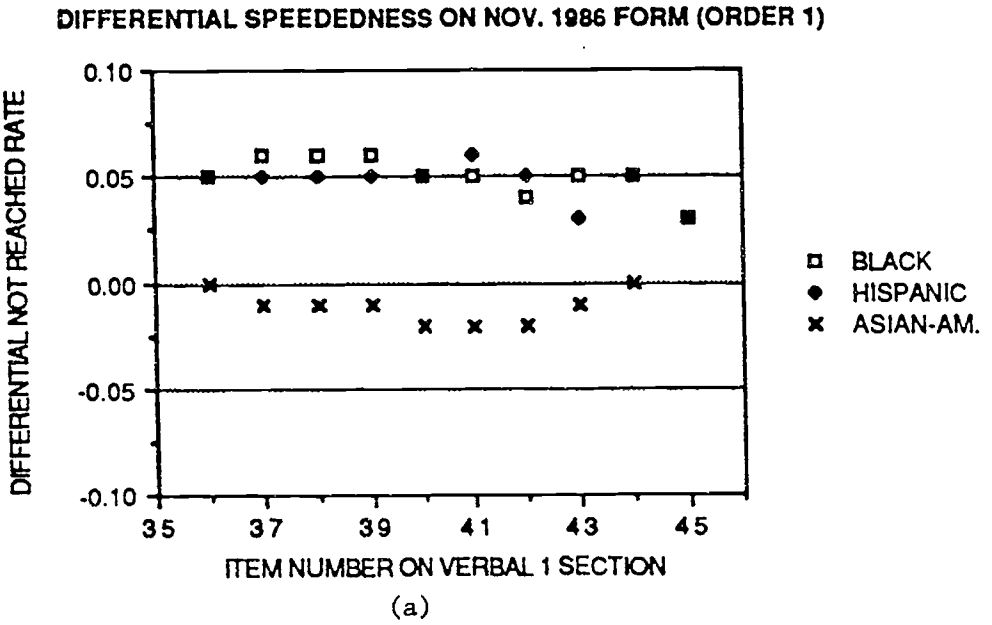

DIFFERENTIAL SPEEDEDNESS ON NOV.1984 FORM
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
35
37
39
41
43
45
ITEM NUMBER ON VERBAL 1 SECTION
BLACK
PUERTO RICAN
MEXICAN-AM.

(a)

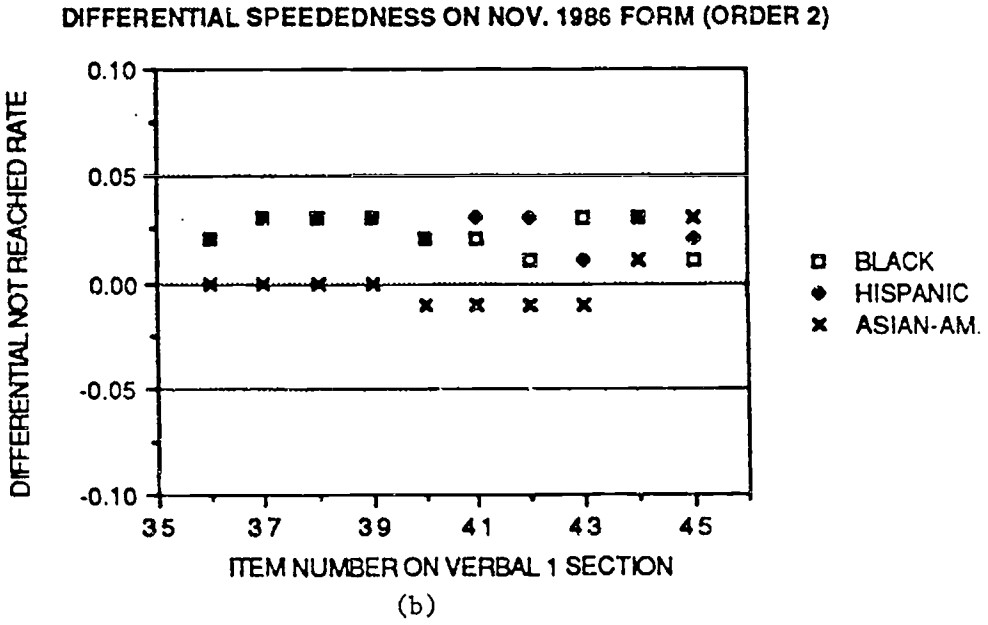

DIFFERENTIAL SPEEDEDNESS ON NOV. 1984 FORM
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
30
32
34
36
38
40
42
ITEM NUMBER ON VERBAL 2 SECTION
BLACK
PUERTO RICAN
MEXICAN -AM.

(b)

FIGURE 3

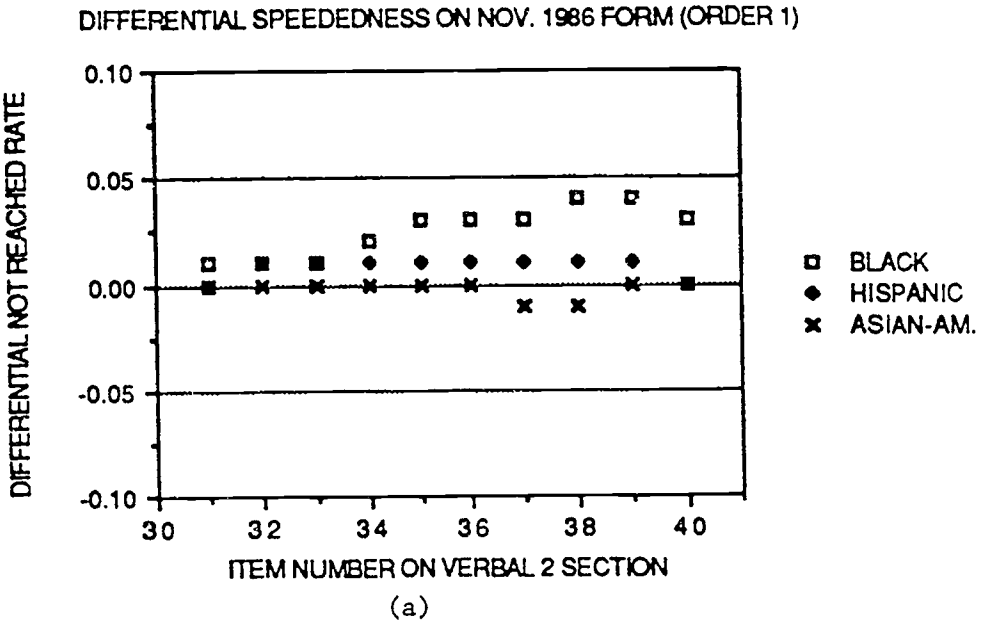

DIFFERENTIAL SPEEDEDNESS ON NOV. 1986 FORM (ORDER 1)
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
35
37
39
41
43
45
ITEM NUMBER ON VERBAL 1 SECTION
BLACK
HISPANIC
ASIAN-AM.

(a)

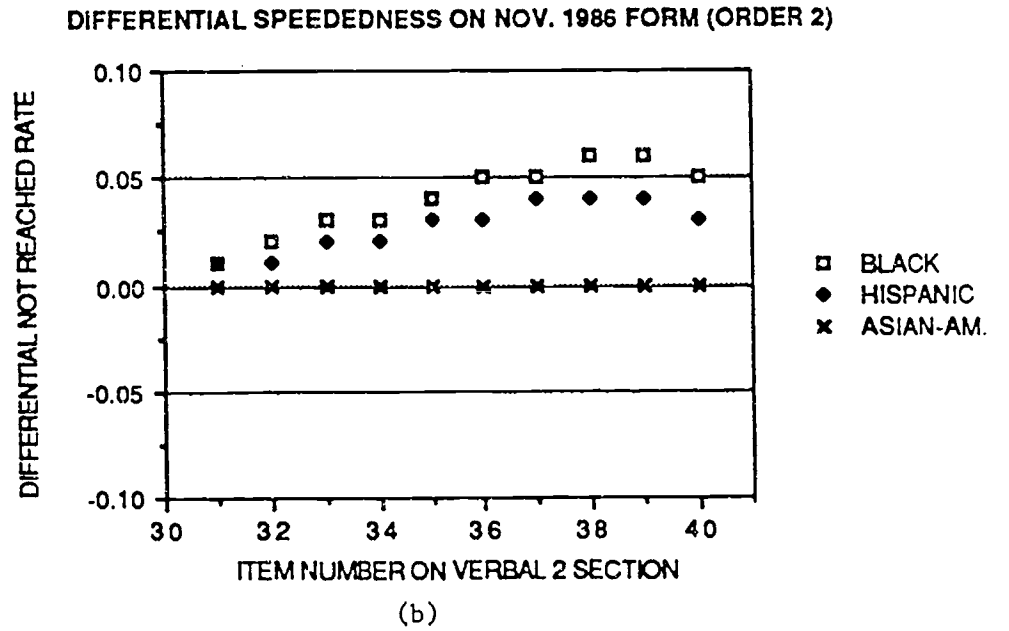

DIFFERENTIAL SPEEDEDNESS ON NOV. 1986 FORM (ORDER 2)
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
35
37
39
41
43
45
ITEM NUMBER ON VERBAL 1 SECTION
BLACK
HISPANIC
ASIAN-AM.

(b)

FIGURE 4

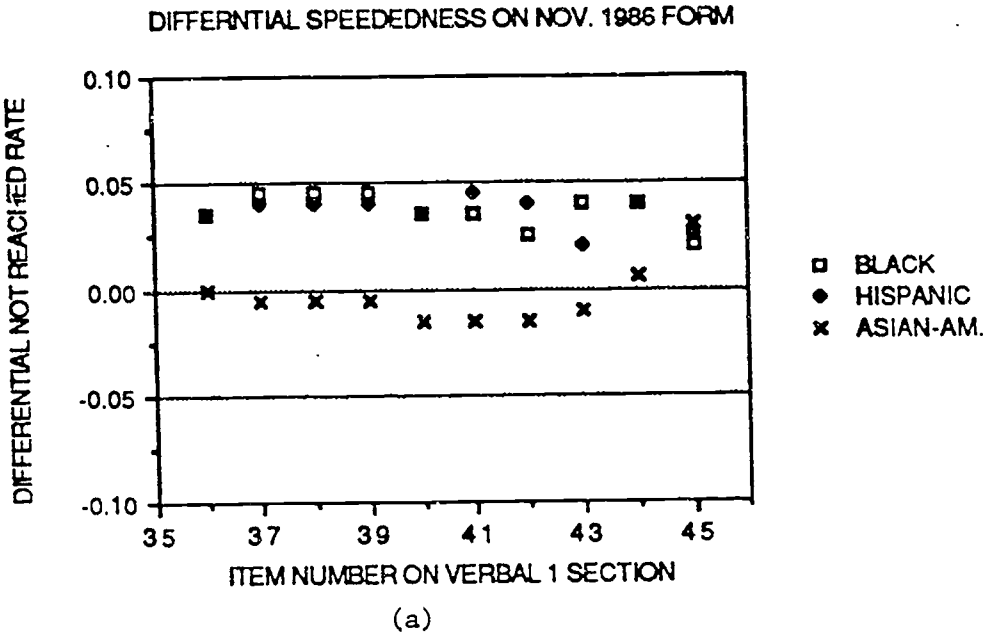

DIFFERENTIAL SPEEDEDNESS ON NOV. 1986 FORM (ORDER 1)
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
30
32
34
36
38
40
ITEM NUMBER ON VERBAL 2 SECTION
BLACK
HISPANIC
ASIAN-AM.

(a)

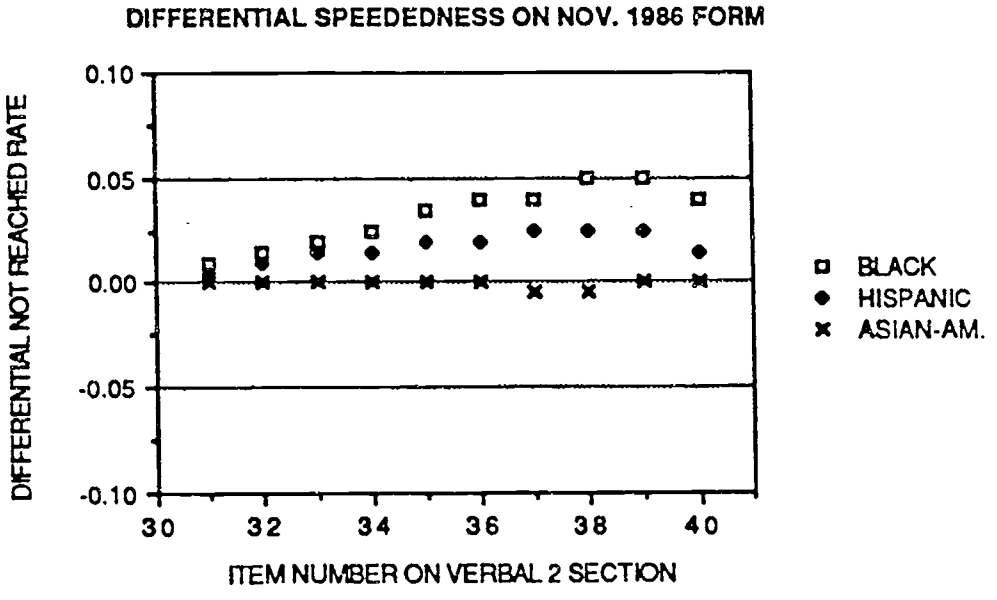

DIFFERENTIAL SPEEDEDNESS ON NOV. 1986 FORM (ORDER 2)
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
30
32
34
36
38
40
ITEM NUMBER ON VERBAL 2 SECTION
BLACK
HISPANIC
ASIAN-AM.

(b)

FIGURE 5


DIFFERENTIAL SPEEDEDNESS ON NOV. 1986 FORM
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
35
37
39
41
43
45
ITEM NUMBER ON VERBAL 1 SECTION
BLACK
HISPANIC
ASIAN-AM.

(a)


DIFFERENTIAL SPEEDEDNESS ON NOV. 1986 FORM
DIFFERENTIAL NOT REACHED RATE
0.10
0.05
0.00
-0.05
-0.10
30
32
34
36
38
40
ITEM NUMBER ON VERBAL 2 SECTION
BLACK
HISPANIC
ASIAN-AM.

(b)