DOCUMENT RESUME

ED 300 440 TM 012 484

AUTHOR Rock, Donald; Chan, Kaling
TITLE Differential Item Functioning Analysis of Math Performance of Hispanic, Asian, and White NAEP Respondents.
INSTITUTION National Assessment of Educational Progress, Princeton, NJ.
PUB DATE Jul 88
NOTE 14p.; Appendix A is not available because it consists of information not made public; it is not included in the pagination.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Asian Americans; Cubans; Difficulty Level; Elementary Secondary Education; Ethnic Groups; Grade 3; Grade 7; Grade 11; *Hispanic Americans; *Item Analysis; *Mathematics Achievement; Mathematics Tests; Mexican Americans; Puerto Ricans; Test Bias; Test Items; *White Students
IDENTIFIERS *Differential Item Performance; Mantel Haenszel Procedure; *National Assessment of Educational Progress

ABSTRACT

The differential item functioning of mathematics performance of Hispanic, Asian, and White students on the National Assessment of Educational Progress (NAEP) was studied, using a modification of the Mantel-Haenszel Procedure. The sample for grades 3, 7, and 11, respectively, included: (1) 1,367, 1,570, and 1,580 Whites; (2) 265, 613, and 760 Asians; (3) 1,238, 1,602, and 1,022 Mexican Americans; (4) 566, 624, and 458 Puerto Ricans; and (5) 292, 347, and 566 Cubans. The data are from the 1985-86 NAEP regular assessment and the NAEP special supplemental study of language minority students. With the exception of third graders, no simple explanation could be offered as to why some items were more difficult or easier for ethnic group members. For the third grade, items with little or no text, involving simple arithmetical operations, were differentially easier for many ethnic group members. Overall, as many items favored the language groups as favored the majority comparison group at each grade level. Four tables summarize the study data. (SLD)

TM

DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF MATH PERFORMANCE OF HISPANIC, ASIAN AND WHITE NAEP RESPONDENTS

Donald Rock and Kaling Chan

National Assessment of Educational Progress
Educational Testing Service
Princeton, New Jersey

July 1988

A Differential Item Functioning Analysis of Math Performance of Hispanic, Asian and White NAEP Respondents

Introduction

Investigations of differential item performance-the problem of identifying items on which the performance of subpopulations is in some way not consistent with their performance on other items- has had a long history. Berk (1982), Shepard, Camilli and Averill (1981), Lord (1980) and Scheuneman (1979) discuss various approaches to identifying items that seem to "work" differently for various subgroups. The methods that have been proposed have often been referred to as "item bias" procedures. The more psychometrically acceptable procedures incorporate the notion of comparing the item performance of members of different subgroups who are in some sense comparable.

When comparing two populations on any criterion, it is important that only comparable members of the two groups be compared. For example, comparing Blacks and Whites' item performance when the groups' members may have come from different parent distributions with respect to grade in school, would lead to a meaningless contrast. A better approach would be to insure that only Black third graders were being compared to White third graders etc. What are the most important control variables depends to a certain extent on the purpose of the group contrasts.

The typical "item bias" approach to insuring comparability of group members is to use the total test score as a control variable. The question becomes: Do individuals from differing groups who have the same total test score perform the same way on a given item? If there is differential item performance under this control condition, then that difference in performance is interpreted as a difference attributable to characteristics of the particular item and not due to differences in the characteristics of the individuals as measured by the total score. If we do not control for the total score, then the finding of differential item difficulty confounds examinee characteristics with item characteristics and we are simply measuring impact and not differential item functioning (DIF). It should be kept in mind that this definition of Differential Item Functioning assumes that the mathematics test is unidimensional. Factor analysis of the NAEP mathematics items suggest that this is the case.

Marascuillo and Slaughter (1981) suggested that a good way of looking at differential item functioning would be to use multi-way contingency tables where at least one of the classification variables would be the control variable. Holland (1985) in the same spirit suggested the use of the Mantel and Haenszel's (1959) contingency table analysis procedure for identifying items with differential item functioning (DIF) characteristics. It was a modification of the Mantel-Haenszel procedure as suggested by Holland that was used here.

The Mantel-Haenszel Procedure

The Mantel-Haenszel (MH) procedure first divides the reference group, say native English speakers, and the focal group, say Spanish speaking Mexican-Americans into subsets that are matched on the total test score before their performance on the items is compared. For any given item a matched subset can be formed, say individuals from either group that scored ten correct on the total test score, and then a 2 x 2 table can be formed where one dimension of the table is the two groups being compared and the remaining dimension is whether the individuals got the item right or wrong. The cell entries in the 2 x 2 table are the frequencies of rights and wrongs for the focal and reference group. For example the following table might reflect the frequencies for the two groups on the first item for those individuals who had a total scores of 10.

	Rights	Wrongs
Reference (White)	a	b
Focal (Ethnic)	c	d

The odds that a reference group member gets the item correct is a/b while the corresponding odds for the focal group member is c/d. The MH procedure measures the advantage (or disadvantage) that the reference group members have to their matched counterparts in the focal group by the ratio of their respective odds. This gives us the odds ratio estimate

$$\alpha_i = (a/b) \,/\, (c/d) \qquad (1)$$

where α_i estimates the population odds ratio for the ith matched group on this particular item. If the odds ratio in (1) is much greater than 1.0, then we infer that the item favors the reference group. Conversely if it is significantly less than 1.0 then we infer that it favors the focal group. For a given item the Mantel-Haenszel estimate α_{MH} is a weighted average of the odds ratios taken across all k matched groups where the weights depend on a proportionality function at each matched total score level.

$$\text{That is } \alpha_{MH} = \frac{\sum_{i=1}^{k} \frac{b_i c_i}{T_i} \alpha_i}{\sum_{i=1}^{k} \frac{b_i c_i}{T_i}} \qquad (2)$$

Where T_i = total numbers of individuals in the ith matched total score level.

Associated with the estimate α_{MH} is a chi-square test with one degree of freedom. The hypothesis being tested here is that all the odds ratios in the k matched subsamples for a given item are unity.

Sample

The data for this analysis come from the 1985-86 NAEP regular assessment and from the NAEP special supplemental study of language minority students.[1] All White students in NAEP who received math block 9M4 in third grade, block 13M7 in seventh grade, and block 17M8 in eleventh grade are contrasted with Asian, Mexican American, Puerto Rican and Cuban students in the supplemental study who received those same blocks.[2] Table 1 presents the sample for this study.

Table 1

Sample for this Study

	Grade 3	Grade 7	Grade 11
White	1367	1570	1580
Asian	265	613	760
Mexican American	1238	1602	1022
Puerto Rican	566	624	458
Cuban	292	347	566

[1]For a detailed discussion of the sampling procedures in NAEP and the special supplemental study see: Johnson, E., Kline, D., Norris, N., & Rogers, A. (1987). National assessment of educational progress 1985-1986 public use tapes: Version I users' guide. Princeton, NJ: Educational Testing Service, and National Assessment of Educational Progress (1988). Findings from the NAEP 1985-86 special study: the educational progress of language minority children. Princeton, NJ: Educational Testing Service.

[2]Although the supplemental study included a sample of Native Americans, their numbers were too small in each grade to generate reliable estimates in a DIF analysis. For this reason they are not included in this study. Appendix A presents the items used in this analysis.

Results

Table 2 presents the third grade mathematics results when the Asians, Mexican-Americans, Puerto Ricans, and Cubans are each compared with the White reference group who took the same block of mathematics items.

(Table 2 about here)

The three columns under each ethnic group show the effect size, the Mantel-Haenszel chi-square, and the odds ratio respectively. As indicated in the table when the odds ratio is less than 1. and the effect size is positive, then there is some suggestion that the item is favoring the particular focal group, i.e., the ethnic group is doing better on average on this item than expected from their total test score. Conversely, if the odds ratio is greater than 1. and the effect size is negative, than there is some suggestion that the item is favoring the reference group (Whites). Items that are marked with an asterisk show statistically significant differential item functioning. However, in order to protect against possible over-interpretation in the presence of repeated statistical tests, we will only attempt to interpret those differences that are statistically significant and have an absolute effect size of 1.5 or greater. These effect sizes are on a scale of difference of item difficulty as measured by the ETS delta scale (Holland & Thayer, 1985). One can loosely interpret a differential item function difference (effect size) of 1.5 as one and half standard deviations difference on the delta scale. It has been ETS' experience that differences in differential performance of this magnitude often can be explained by the content or cognitive demand required to solve the item.

Inspection of Table 2, the third grade results, indicates that items 1 and 2 are both differentially easy for the ethnic groups (with the possible exception of the Mexican-American group on item 1), while conversely item 12 is differentially hard for the ethnic groups when compared to the White reference group. Inspection of items 1 and 2 indicate that they are simple single operation arithmetic (subtraction of whole numbers) problems with no text involved. Conversely item 12 is a more difficult item and has considerable text in the stem. In addition item 12 requires a basic notion of the concept of probability. One cannot separate how much of this differential difficulty is due to the vocabulary in the text and how much may be due to lack of exposure to the concept of probability.

Table 3 presents the reference-focal group contrasts for the seventh grade mathematics items.

(Table 3 about here)

Inspection of Table 3 indicates that only two items meet both criteria-statistical significance and effect sizes whose absolute values are equal to or greater than 1.5. Given these criteria, Item 8 would appear to be

differentially hard for the Asians when compared to the White reference group. Conversely item 1 appears to be differentially easy for the Puerto Ricans. Inspection of item 8 indicates a somewhat heavier load of textual material than many of the other items. While not significant other text based items such as items 2, 3, and 7 also have negative effect sizes suggesting that the text may indeed make the item differentially difficult for Asians.

Item 1 appears to favor the Puerto Ricans, and it indeed is a simple arithmetic operation with a small amount of text involved. But, the text is quite easy in terms of familiar vocabulary and the item solution only requires simple arithmetic. Overall there does not appear to be any systematic bias against the ethnic groups that can be explained by a language problem. If one simply counts the number of statistically significant comparisons, there are 20 favoring members of ethnic groups and 14 favoring the reference group.

Table 4 presents the reference-focal group comparisons on the eleventh grade mathematics test.

(Table 4 about here)

Inspection of the Table 4 results suggest that items 28 and 25 seem to be differentially difficult for most if not all the ethnic groups. Inspection of the table suggests that many of the negative effects sizes occur among the more difficult items. Many of these items would be difficult if one were less likely to be exposed to the concepts involved. For example, item 25 requires an understanding of the concept of probability, while item 28 requires a basic knowledge of geometry. It may well be that at least some members of the above ethnic groups are less likely to be exposed to curriculum that covers probability concepts than are their White counterparts. Item 28 requires knowledge of both a basic definition in geometry as well as an inference about what the item writer means by the term "combined" in the particular context of the problem. It would seem that item 28 is differentially hard for all ethnic groups due to this inference from the text.

It is interesting to note that one item (item 4) that was differentially easy for Mexican-Americans and Puerto Ricans consisted of reading and making a selection among graphs. It is not clear why this is the case except there is very little text involved in the pictorial presentation.

Conclusions

With the exception of the third graders, no simple explanation can be put forth with any confidence with respect to why some items were differentially difficult or differentially easy for the ethnic group members. However, in the case of the third graders, items which had little or no text and involved simple arithmetical operations were differentially easier for many of the ethnic group members. In general, however, there was no consistent evidence for differential difficulty at any grade level. That is, there tended to be as many items favoring the language groups as there were items favoring the majority group at each grade level.

References

Berk, R. (Ed.) (1982) Handbook of Methods for Detecting Test Bias.Baltimore and London: John Hopkins University Press.

Holland, P.W. and Thayer, D.T. (1985) Alternative definition of the ETS delta scale of item difficulty. Princeton, NJ, Educational Testing Service, Research Report RR-85-43.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysisof ^R ata from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Marascuillo, L. and Slaughter R. (1981) Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

Scheuneman, J. (1979) A new method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L., Camilli, G., and Averill, M. (1981) Comparisons of six procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

TABLES

Table 2

WHITE ITEM PERFORMANCE REFERENCE GROUP CONTRASTED WITH THAT OF EACH OF FOUR LANGUAGE (FOCAL) GROUPS AT GRADE 3

	Asians N=265			Mexican Americans N=1238			Puerto Ricans N=566			Cuban Americans N=292		
Items	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio
1	2.26	12.56*	.38	1.31	21.56*	.57	2.84	48.10*	.30	2.51	24.14*	0.34
2	1.92	9.24*	.44	1.67	32.37*	.49	1.77	21.91*	.47	1.78	14.49*	0.47
3	1.94	18.53*	.44	0.56	5.94*	.79	0.30	0.96	.88	0.92	5.95*	0.68
4	-0.10	0.01	1.04	0.46	3.17	.82	0.59	3.35	.78	0.28	0.44	0.89
5	2.26	15.20*	.38	0.91	12.51*	.68	0.74	4.90*	.73	0.63	2.17	0.76
6	-0.25	0.20	1.11	0.43	2.94	.83	0.45	1.88	.83	0.32	0.53	0.87
7	0.07	0.01	0.97	-0.04	0.02	1.02	-0.91	10.39*	1.47	-0.32	0.63	1.15
8	0.04	0.00	0.98	0.43	3.57	0.83	0.75	5.54*	.73	0.24	0.31	0.90
9	-0.95	4.42*	1.50	-1.11	20.66*	1.60	-2.01	38.79*	2.35	-1.19	7.87*	1.66
10	-0.70	2.65	1.34	-1.10	21.06*	1.59	-1.29	14.64*	1.73	-1.57	14.23*	1.94
11	-1.15	6.18*	1.63	-0.57	5.25*	1.27	-0.80	5.69*	1.41	-0.65	2.24	1.32
12	-2.34	31.96*	2.71	-1.58	37.48*	1.95	-1.54	18.03*	1.92	-2.01	19.37*	2.35
13	0.34	0.65	0.86	0.26	1.13	0.89	1.14	11.63*	.62	-0.31	0.43	1.14
14	0.38	0.74	0.85	0.61	4.85*	0.77	1.34	14.41*	.57	1.03	5.94*	0.64
15	-1.02	3.89*	1.55	0.03	0.00	0.99	0.07	0.01	.97	0.06	0.00	0.97
16	0.40	0.51	0.85	-0.59	2.23	1.29	-1.30	4.29*	1.74	-0.77	1.44	1.38
17	0.82	0.44	0.71	0.59	0.60	0.78	0.33	0.03	.87	0.40	0.03	0.84
18	-0.49	1.35	1.23	-1.10	25.92*	1.59	-1.43	26.38*	1.84	-0.54	2.17	1.26

[1] If the effect size is ≥ 1.5, it is interpreted as a practical difference in favor of the focal minority group. The relationship is symmetric and an effect size ≤ -1.5 would indicate favoring the White reference group.

[2] Odds ratios >1.0 suggest the item is favoring the reference group. Odds ratios <1.0 suggest the opposite.

*Significant M-H x^2 at the alpha .05 level or better.

Table 3

WHITE ITEM PERFORMANCE REFERENCE GROUP CONTRASTED WITH THAT OF EACH OF FOUR LANGUAGE (FOCAL) GROUPS AT GRADE 7

	Asians N=613			Mexican American N=1602			Puerto Ricans N=624			Cuban Americans N=347		
Items	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio
1	0.54	1.13	0.80	1.25	18.42*	0.59	1.62	17.84*	0.50	0.79	2.66	0.71
2	-0.61	3.10	1.30	-0.39	3.27	1.18	-1.21	19.20*	1.67	0.19	0.23	0.92
3	-0.15	0.33	1.07	0.25	1.83	0.90	0.13	0.24	0.95	0.07	0.03	0.97
4	0.42	1.59	0.84	0.56	7.06*	0.79	0.58	4.46*	0.78	0.84	5.60*	0.70
5	0.00	0.00	1.00	-1.09	26.82*	1.59	-1.41	22.66*	1.82	-0.53	2.26	1.25
6	0.00	0.00	1.00	0.11	0.27	0.96	0.12	0.18	0.95	0.30	0.71	0.86
7	-0.72	4.96*	1.36	-1.23	33.56*	1.68	-0.38	1.76	1.18	-0.20	0.25	1.09
8	-1.51	32.08*	1.90	-0.74	13.42*	1.37	-1.10	15.86*	1.60	-0.95	7.49*	1.50
9	-0.09	0.06	1.04	-0.60	10.01*	1.29	-0.46	3.33	1.22	-0.64	4.25*	1.31
10	0.14	0.18	0.94	0.12	0.33	0.95	0.07	0.05	0.97	-0.18	0.24	1.08
11	0.12	1.07	0.95	-0.28	1.54	1.13	0.59	3.82	0.78	0.63	2.29	0.75
12	-0.21	0.52	1.09	-0.07	0.10	1.03	-0.26	0.96	1.12	-0.37	1.33	1.17
13	0.70	5.50*	0.74	0.60	8.70*	0.77	1.08	14.79*	0.63	0.62	2.66	0.77
14	0.91	8.07*	0.68	0.84	13.67*	0.70	1.08	11.90*	0.63	0.42	1.06	0.84
15	1.03	10.65*	0.65	0.81	13.05*	0.71	0.99	10.02*	0.66	0.73	3.25	0.73
16	0.98	8.36*	0.66	0.99	19.60*	0.66	1.05	11.54*	0.64	-0.05	0.00	1.01
17	-0.57	3.74	1.28	-0.46	4.37*	1.22	-0.14	0.15	1.06	-0.56	2.00	1.27
18	-0.05	0.01	1.02	-0.37	2.55	1.17	-0.12	0.12	1.05	-0.10	0.04	1.04
19	0.64	5.06*	0.76	0.46	3.89*	0.82	0.91	8.64*	0.68	0.47	1.44	0.82
20	0.10	0.10	0.96	-0.14	0.39	1.06	-1.14	13.32*	1.62	-0.10	0.05	1.04
21	-0.47	2.21	1.22	0.42	3.57	0.83	-0.07	0.03	1.03	-0.03	0.01	1.01
22	-0.35	1.36	1.16	-0.18	0.56	1.08	-0.95	8.72*	1.50	-0.53	1.79	1.25

[1] If the effect size is ≥ 1.5, it is interpreted as a practical difference in favor of the focal minority group. The relationship is symmetric and an effect size ≤ -1.5 would indicate favoring the White reference group.

[2] Odds ratios >1.0 suggest the item is favoring the reference group. Odds ratios <1.0 suggest the opposite.

* Significant M-H x^2 at the alpha .05 level or better.

Table 4

WHITE ITEM PERFORMANCE REFERENCE GROUP CONTRASTED WITH THAT OF EACH OF FOUR LANGUAGE (FOCAL) GROUPS AT GRADE 11

	Asians N=760			Mexican Americans N=1022			Puerto Ricans N=458			Cuban Americans N=566		
Items	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratic	Effect[1] Size	M-H x^2	Odds[2] Ratio	Effect[1] Size	M-H x^2	Odds[2] Ratio
1	0.07	0.00	0.97	0.47	1.56	0.82	-0.26	0.25	1.11	0.41	0.50	0.84
2	0.11	0.07	0.96	-0.04	0.01	1.02	0.36	1.02	0.86	-0.29	0.67	1.13
3	1.31	16.14*	0.57	-0.55	4.84*	1.26	0.26	0.59	0.89	0.28	0.67	0.89
4	0.97	5.31*	0.66	2.16	31.49*	0.40	2.28	18.51*	0.38	0.88	3.76	0.69
5	0.16	0.34	0.93	0.13	0.26	0.95	0.40	1.64	0.84	-0.01	0.00	1.01
6	0.62	1.68	0.77	0.01	0.00	0.99	0.39	0.73	0.85	0.27	0.26	0.89
7	2.28	50.20*	0.38	0.70	8.02*	0.74	1.25	14.48*	0.59	0.33	1.10	0.87
8	0.53	3.22	0.80	0.45	3.40	0.82	0.62	3.75	0.77	0.92	7.57*	0.68
9	-0.49	3.15	1.23	0.02	0.00	0.99	0.17	0.25	0.93	0.12	0.14	0.95
10	-0.25	0.56	1.11	-0.61	6.01*	1.30	0.33	0.86	0.87	0.81	5.00*	0.71
11	0.14	0.22	0.94	0.31	1.74	0.88	0.39	1.68	0.85	-0.53	3.58	1.25
12	1.81	38.65*	0.46	0.50	4.43*	0.81	0.66	4.14*	0.76	0.19	0.35	0.92
13	0.81	10.48*	0.71	0.52	5.24*	0.80	0.53	2.83	0.80	1.13	18.82*	0.62
14	0.74	8.38*	0.73	0.41	3.22	0.84	0.74	5.84*	0.73	0.72	7.35*	0.74
15	-0.33	1.31	1.15	-0.30	1.57	1.14	-0.24	0.56*	1.11	-0.06	0.02	1.03
16	-1.57	18.37*	1.95	-0.80	6.77*	1.41	-0.53	1.50	1.25	-0.72	2.64	1.36
17	1.62	29.09*	0.50	0.06	0.04	0.98	0.18	0.28	0.93	0.41	1.86	0.84
18	-0.10	0.08	1.04	0.22	0.73	0.91	0.15	0.17	0.94	0.38	1.43	0.85
19	0.34	1.36	0.87	-0.53	5.31*	1.25	0.03	0.00	0.99	0.07	0.03	0.97
20	-0.41	2.53	1.19	-0.31	1.75	1.14	-0.37	1.23	1.17	-0.86	10.16*	1.44
21	-0.38	1.94	1.17	0.48	4.13*	0.81	0.45	1.79	0.83	0.54	3.50	0.80
22	-1.57	29.32*	1.95	-0.25	0.96	1.11	-1.14	11.67*	1.62	-0.57	3.26	1.28
23	0.35	1.93	0.86	0.43	3.12	0.83	0.35	1.01	0.86	0.13	0.17	0.95
24	-0.72	7.72*	1.36	-0.45	3.46	1.21	-0.43	1.62	1.20	-0.24	0.70	1.11
25	-2.62	84.58*	3.05	-0.67	7.30*	1.33	-2.37	46.36*	2.74	-1.54	28.69*	1.92
26	-0.36	1.42	1.16	-0.09	0.08	1.04	-0.62	2.08	1.30	-0.27	0.66	1.12
27	0.06	0.03	0.97	-0.31	1.36	1.14	-1.30	11.25*	1.74	0.17	0.28	0.93
28	-1.72	47.73*	2.08	-0.80	13.79*	1.41	-1.97	44.30*	2.31	-1.60	37.63*	1.98

[1] If the effect size is ≥ 1.5, it is interpreted as a practical difference in favor of the focal minority group. The relationship is symmetric and an effect size ≤ -1.5 would indicate favoring the White reference group.

[2] Odds ratios >1.0 suggest the item is favoring the reference group. Odds ratios <1.0 suggest the opposite.

*Significant M-H x^2 at the alpha .05 level or better.

APPENDIX A

Note Bene: These items are not to be circulated. They have not been released to the public.