Journal of Educational Measurement
Summer 1989, Vol. 26, No. 2, pp. 147–160

A Method for Studying Differential Distractor Functioning

Bert F. Green
Carolyn R. Crone
Johns Hopkins University
and
Valerie Greaud Folk
Syracuse University

A method of analyzing test item responses is advocated to examine differential item functioning through distractor choices of those who answer an item incorrectly. The analysis, called Differential Distractor Functioning, uses log-linear models of a three-way contingency table to examine whether there is an interaction of population subgroup and option choice when ability is held constant. The analysis is explained and is exemplified in an analysis of the Verbal portion of a recent Scholastic Aptitude Test.

Differences in test performance among ethnic groups have been a topic of extensive study and discussion in the testing profession in the past decade (Cole, 1981; Cronbach, 1976; Green, 1981; Jensen, 1980.) Ethnic groups often show large differences in average test scores, and some tests show gender differences. Critics have claimed that the differences in mean test scores signify that the tests are biased. But persons of different backgrounds and with different opportunities will have different educational and life experiences, and because tests assess present performance, the test scores may reflect real differences in knowledge, skills, and developed abilities. It is necessary somehow to separate the true differences from the artificial differences due only to the testing process.

Some interesting clues about group differences in the testing process are afforded by the particular incorrect responses that people make to test items. If different groups prefer different incorrect responses to an item, often called foils or distractors, then the item probably means something different to the different groups. Items that have different meanings to different groups would seem to be biased in a very fundamental sense. This paper discusses a way to analyze group differences in option choices.

Most studies of group differences in test responses have focused on correct answers. Two distinct lines of research have emerged. Studies of test bias (Berk, 1982; Bond, 1981; Cleary, 1986; Cole, 1973; Linn & Werts, 1971) address the question of whether test scores have the same meaning for different groups of examinees by determining if the relationship of the test scores to later performance in school or on the job is the same for each subgroup. By contrast, studies of differential item functioning (DIF), often called *item bias* in the earlier

This work was supported in part by the Educational Testing Service, Princeton, NJ.

literature, examine whether some of the items in the test may be favoring some subgroups in the sense that a larger proportion of subgroup members answered the item correctly than would have been expected from their average test performance (Angoff, 1982; Dorans & Kulick, 1986; Lord, 1980; Rudner, Getson, & Knight, 1980; Shepard, 1981, 1982; Tittle, 1982).

Our proposed method for studying DIF depends on the oft-stated suggestion that incorrect answer options might be differentially attractive to persons of different ethnic backgrounds (Frary & Giles, 1980; Veale & Forman, 1983). Although group differences in distractor choice have no effect on test scores, because all distractors are wrong, group differences might indicate that the item was functioning differently for the different subgroups. Extending the terminology of DIF, this analysis concerns differential distractor functioning (DDF).

Because distractor choices have no effect on test scores, it might seem that a DDF analysis would not need to control for total test scores. In fact, Osterlind (1983) described a distractor analysis that does not take ability into account. It involves a two-way contingency analysis of subgroup by option choice for each item. However, item analyses often find that different distractors are chosen by persons of different ability levels. Some distractors appeal to persons who know nothing whatever about the item's content; others provide reasonable but wrong alternatives that might be chosen by the partially informed. In the face of such an interaction of ability and option choice, ability must somehow be held constant when analyzing group differences in option choice, lest a detected group difference really turn out to be an ability difference in disguise.

Our method for analyzing incorrect item responses by subgroup is based on a three-way contingency table of item choice by subgroup by ability level. The method is explained in detail in the next section; then its use is illustrated by an analysis of the verbal section of a recent form of the Scholastic Aptitude Test (SAT).

Log-Linear Analysis of DDF

The method proposed here for studying DDF is much like the standardization method for stuyding DIF (Dorans & Kulick, 1986). Examinees are separated into several ability levels, and the frequency of each option choice is tabulated separately for each subgroup at each ability level, yielding a three-way contingency table. The ability levels can be defined by an independent measure of ability, if one is available, but usually the total test score is used. If the score is based on a large number of items, the minor dependence incurred from the self-correlation in item-ability relationships can safely be ignored. An advantage of using total test score is that the ability measured by the test is that tapped by each item, whereas an independent measure of ability might not be quite the same as that measured by the items.

Subgroup comparisons are addressed through an application of the log-linear analysis of contingency tables (Bishop, Fienberg, & Holland, 1975; Kok, Mellenbergh, & van der Flier, 1985; Marascuilo & Slaughter, 1981; Mellenbergh, 1982; SPSS-X, 1983). This analysis is a logical extension of chi-square analyses for two-way contingency tables to higher order tables. The log-linear

analysis provides for the separation of various main effects and interactions of the main variables being categorized, in the style of the analysis of variance. The observed frequency in each tabular cell is modeled as a product of not two but several factors. In addition to factors for the three one-way marginal distributions, the analysis also includes factors to explain the various two-way interactive effects and the three-way combined effect.

For example, the particular cell that indicates the frequency of white, low-ability examinees choosing Option 2 would have a certain expectation because of the overall number of whites, as well as an expectation based on the low-ability level and on the overall number of people choosing Option 2. Each of these expectations is based on a main effect. Two-way interactions can be used as additional adjustments, such as the number of whites who are at the low-ability level, the tendency of low-ability students to choose Option 2, and the specific tendency of whites to choose Option 2, over and above the other tendencies. The last-named effect, the interaction of subgroup and option choice, is the critical issue. The central question is whether the data require a factor to account for the differential tendency of different subgroups to choose particular options.

In this analysis, each effect represents differences in proportions. The main effect of subgroup merely accounts for any difference in the number of examinees in each subgroup in our study. The main effect of ability likewise accounts for any difference in the sizes of the different ability groupings. Both of these effects are arbitrary, depending on the selection of groups and the definition of ability levels. The interaction of these two effects, ability level and subgroup, may be of interest in other contexts, but it should be set aside from an analysis of DDF.

Our interest focuses on the option choices and the interactions of option choice with the other two factors. The main effect of option choice accounts for the overall differences in the popularity of the different options. The interaction of option choice with ability level is mainly interesting to test authors. It indicates the extent to which examinees of different abilities are drawn to different wrong answer options. Some test authors try to write items that do just that. Normally, we would expect many items to show an interaction of option preference by ability level. But the major question is whether, in addition, examinees of different subgroups choose the various options more or less frequently than would be expected by their ability level. This is indicated by the interaction of subgroup and option choice.

Thus our analysis will test the statistical hypothesis that the interaction between subgroup and option choice is not needed to explain the observed item responses. That is, the analysis will assess the adequacy of a model that includes main effects of subgroup, ability level, and option choice, as well as the interaction of subgroup by ability and the interaction of option choice by ability, but *not* the interaction of option choice by subgroup.

This model will be compared with a model that includes the pivotal interaction, subgroup by answer option. We will infer DDF only if the data *are not* explained without the pivotal interaction, and *are* explained when that interaction is included. If the more complete model still does not account adequately for

the data, then there is some idiosyncratic deviation of some option, in some group, at some ability level. Such an irregular occurrence could not be treated as systematic, and does not indicate DDF.

Analysis Specifications

The analysis as so far described is generic. Several specific choices and decisions must be made to apply it in any instance. First, the analysis may be done for all answer options, or for the incorrect options only. In the examples to be reported, we have chosen to analyze only the incorrect choices. Those persons answering the item correctly are not included in the analysis of that item. We have found that inclusion of the correct choices sometimes tends to overwhelm the smaller, subtle differences among the distractors. For symmetry, a separate analysis can be done using correct-incorrect in place of incorrect answer option to determine DIF in the same context. Such an analysis is virtually the same as the standardization approach of Dorans (1986).

The number of ability groups is also somewhat arbitrary. The group can be separated into as many groups as there are raw scores on the test, but this seems likely to pick up unimportant differences among examinees who differ by as little as one score point. In our work we have chosen to use only five grouped categories of test score.

Finally, any number of subgroups may be included, and the samples of each group may be completely random, or may be matched on ability. We have chosen to match on ability, fearing that otherwise the main effect of group might overwhelm more subtle differences.

To illustrate the nature of the analysis, a complete log-linear analysis is provided in Table 1 for one item (Item 13, Section 1, of SAT Form 6K). The three-way frequency table, shown in Table 2, has five ability levels (A), three

Table 1

Complete Log-linear Analysis for Item 13

Line	Effects	Chi square	Degrees of Freedom	Prob.
1	A	568.18	59-4 = 55	.000
2	B	638.78	59-2 = 57	.000
3	C	296.84	59-3 = 56	.000
4	A,B,C	215.62	59-4-2-3 = 50	.000
5	A,B,C,AB	210.86	50-8 = 42	.000
6	A,B,C,AC	58.99	50-12 = 38	.016
7	A,B,C,BC	192.12	50-6 = 44	.000
8	A,B,C,AB,AC	54.24	50-8-12 = 30	.004
9	A,B,C,AB,BC	187.36	50-8-6 = 36	.000
10	A,B,C,AC,BC	35.49	50-12-6 = 32	.307
11	A,B,C,AB,AC,BC	30.83	50-12-8-6 = 24	.159

Table 2

Frequency Distribution of Incorrect Responses for Item 13

Ability Group	Ethnic Group	Incorrect Response 1	2	3	4	Total
1	Black	37	59	87	57	240
	Hispanic	62	37	82	53	234
	White	31	61	96	56	244
2	Black	30	54	96	50	230
	Hispanic	52	43	82	39	216
	White	24	67	100	58	249
3	Black	25	54	80	43	202
	Hispanic	15	39	74	37	165
	White	19	48	76	32	175
4	Black	23	61	69	25	178
	Hispanic	31	55	57	22	165
	White	19	61	66	36	182
5	Black	28	84	34	15	161
	Hispanic	20	73	36	16	145
	White	30	81	47	21	179
Total	Black	143	312	366	190	1011
	Hispanic	180	247	331	167	925
	White	123	318	385	203	1029

ethnic backgrounds (B), and four incorrect option choices (C), or a total of $5 \times 3 \times 4 = 60$ cells. Because the overall number of incorrect choices of each item is considered fixed, there are 59 degrees of freedom. Like the analysis of variance, the log-linear model includes factors for the main effects A, B, and C, and for each of the two-way interactions, A × B, A × C, and B × C. However, the analysis does not partition a sum of squares, as in the analysis of variance. Rather, each statistical test is a test of the fit of some particular model to the data table. Each line of Table 1 lists the effects included in the model examined in that line, the actual chi-square value obtained, its degrees of freedom (number of independent data proportions less the number of independent model parameters that are estimated), and the probability that a chi square as large or larger than this would be obtained by chance.

For example, to test the main effect of ability, we fit a model that accounts for each cell frequency only in terms of the ability level of that cell. The model can be expressed as

$$\frac{f_{ijk}}{n} = \frac{f_{ioo}}{n},$$

where f_{ioo} indicates a sum over the second and third indices (for variables B and C) and where $f_{ooo} = n$, the total number of cases. Line 1 in Table 1 shows that the chi-square statistic for the fit of this model is 568.18, which indicates a very poor fit. We therefore conclude that factor A alone cannot account for the incorrect choices made to this item.

The first three lines of Table 1 show the result of separately testing each main effect for the corresponding item. In no case is a main effect alone a suitable model. Line 4 results from fitting a model with all three main effects but no interactions. Specifically, it models the data as

$$\frac{f_{ijk}}{n} = \frac{f_{ioo}}{n} * \frac{f_{ojo}}{n} * \frac{f_{ook}}{n}.$$

Note that the item is not fit well by this model.

Lines 5, 6, and 7 each add a different second-order interaction. For example, Line 5 adds the interaction of ability and background, A × B. The cell is modeled as

$$\frac{f_{ijk}}{n} = \frac{f_{ioo}}{n} * \frac{f_{ojo}}{n} * \frac{f_{ook}}{n} * \frac{n * f_{ijo}}{f_{ioo} * f_{ojo}}.$$

Recall that we matched the groups to avoid an interaction of ability and background. However, if the item should happen to show DIF, that is, have different numbers of correct responses by persons of equal ability in the different groups, then our DDF analysis, which includes only the incorrect responses, could show an interaction of ability by background.

The lines 8–10 consider pairs of interactions together, along with the main effects. Our major hypothesis is that the best model for the data is given in line 8, which includes the ability by choice and ability by background interactions, but not the interaction of background and choice. Item 13 (Table 1) is not well fit by this model. Line 10 indicates that the model with both the background by choice (BC) and the ability by choice (AC) interactions fits this data table.

Finally, Line 11 shows the result of fitting a model with all three two-way interactions. The chi square for this line necessarily must be lower than for any of the previous models. If the data can be accounted for without a B × C interaction, then they certainly can be accounted for if that interaction is included. Similarly, if the data *cannot* be accounted for by all three interactions, then it certainly cannot be accounted for without the crucial B × C interaction. Thus we conclude that the item distractors are functioning differentially only when a model without the B × C interaction *does not* fit, whereas a model that includes that interaction does fit.

To give some sense of the size of the effects, Figure 1 (a–d) shows the proportions of choices of distractors for persons of each background at each ability level who choose a wrong response to the sample item. That is, the frequencies in Table 2 are normalized to sum to 1.0 across the four options within each ability level and background group. Apparently, Hispanics tend to choose Option 1 more than the other groups do, particularly in the lower two ability

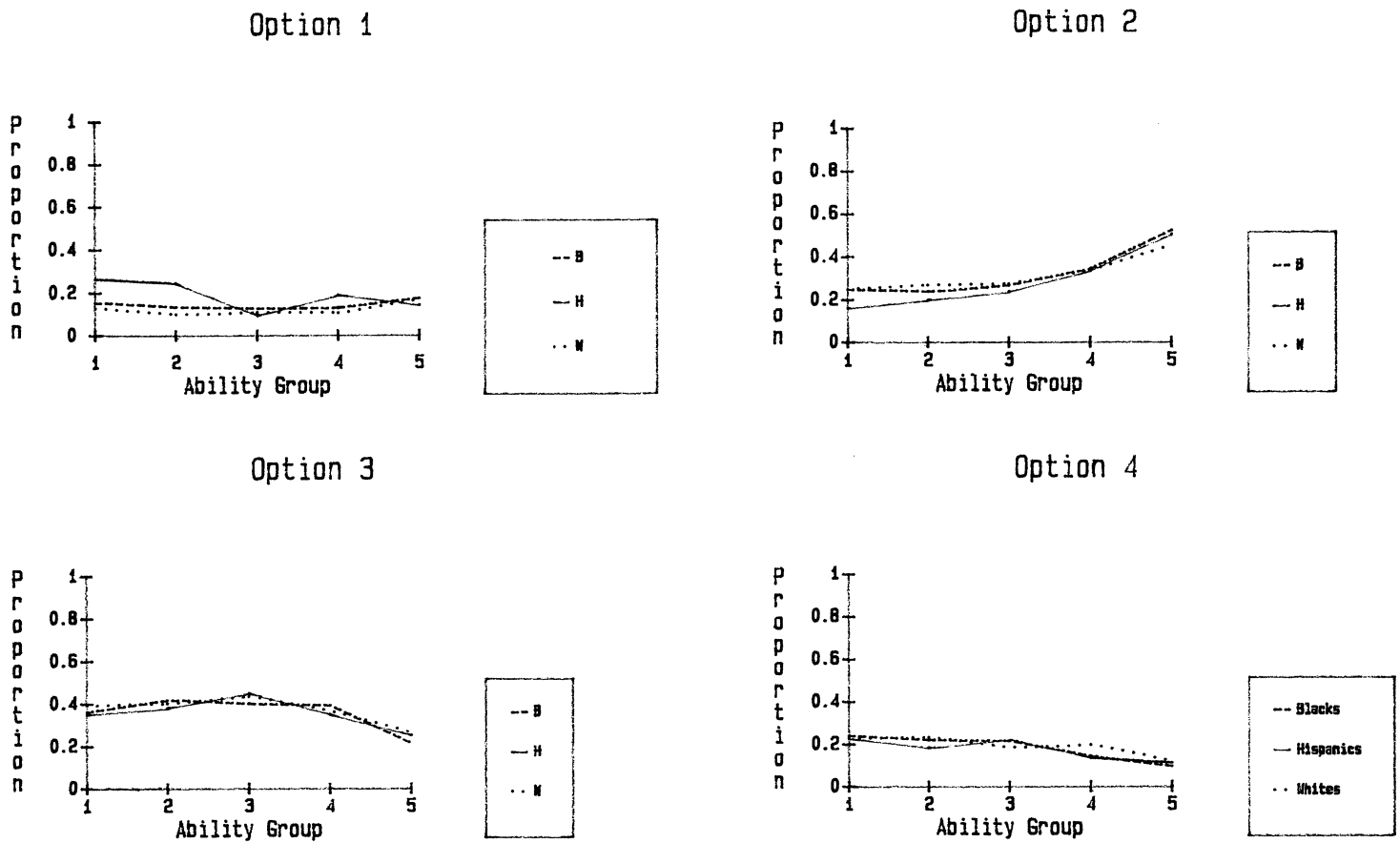

FIGURE 1. *Proportion of three matched groups, at five ability levels, choosing each incorrect option of Item 13. (1a) Option 1. (1b) Option 2. (1c) Option 3. (1d) Option 4.*

groups, whereas Hispanics differentially tend to avoid Option 2. The differences are remarkably small.

Example: The SAT-Verbal

The utility of the log-linear DDF method can be demonstrated by an analysis of the items in the verbal section of the College Entrance Examination Board's Scholastic Aptitude Test (SAT-V). The data were obtained from the public use tape for Form 6K of the SAT-V, obtained in the November 1986 national administration. This SAT form subsequently has been disclosed, so that the items are in the public domain (although still protected by copyright.) The tape includes background questionnaire data supplied voluntarily by the examinees, permitting a comparison of groups of self-identified Whites, Blacks, and Hispanics.

A total verbal score was obtained from the 85 items on the verbal section (Sections 1 and 4 on Form 6K); this raw score served as the measure of ability in the analysis.[1] Population score distributions were made for Whites, Blacks, and Hispanics. The score distributions of Blacks and Hispanics were averaged to

Table 3

DDF Analysis of the Scholastic Aptitude Test, Form 6K; November 1986 Administration.

Item		Models				
		A,B,C,AB,AC		A,B,C,AB,AC,BC		
#	Type	Chi Square	p	Chi Square	p	Difference
Section 1						
1	Op	70.402	.000	15.887	.892	54.515
6	Op	67.612	.000	32.047	.126	35.565
10	Op	61.250	.001	28.520	.239	32.730
13	Op	54.238	.004	30.835	.159	23.403
18	SC	59.204	.001	16.789	.858	42.415
20	SC	149.968	.000	29.250	.211	120.718
22	SC	65.845	.000	15.922	.891	49.923
24	SC	54.069	.005	24.650	.425	29.419
42	An	119.266	.000	38.928	.028	80.338
Section 4						
9	Op	62.776	.000	36.319	.051	26.457
12	SC	57.064	.002	29.257	.211	27.807
14	SC	73.419	.000	31.851	.131	41.568
23	An	51.111	.009	29.721	.194	21.390
24	An	85.929	.000	36.552	.048	49.377

Note Only Items showing significant DDF by our criteria are listed. The analysis includes incorrect options only; omitted items are excluded.
Type code: Op, Opposites; SC, Sentence Completion; An, Analogies; PC, Paragraph Comprehension.

serve as the target distribution. Random samples of 2,000 cases each of Whites, Blacks, and Hispanics were drawn with the restriction that they have the target distribution. This was done by calculating for each score the number needed at that score, as a proportion of the total 2,000, so that the target distribution would be matched. Then that number of cases was selected randomly from those with that score in the respective populations. Finally, the ability scores were divided into approximate quintiles, yielding five matched ability groups of Blacks, Hispanics, and Whites for the analysis.

The DDF analysis described previously then was applied to each of the 85 items on the verbal portion of the test (Sections 1 and 4). For each item, only the incorrect option choices were analyzed. Each item thus involved a table with five ability groups, four incorrect options, and three ethnic groups, for a total of 60 cells. Because we were analyzing so many items, some were likely to appear statistically significant by chance, so the stringent .01 significance level was used. That is, an item was not judged statistically significant unless the probability was less than 1 in 100 that the results could have occurred by chance.[2] Moreover, DDF is indicated only when the model without the background by option choice interaction has a significant chi square, indicating lack of fit, whereas a model that includes that interaction fits the data satisfactorily. The differences are also chi squares, generally on six degrees of freedom. Some of the differences are statistically significant, indicating that the model fit better with than without the critical interaction, but if a model without that interaction fits the data, the fact that the more inclusive model fits even better is not considered indicative of DDF.

Tables 3 and 4 show the results of the analyses. Table 3 shows chi-square values for the critical models for the items that meet our criterion for DDF: Items 1, 6, 10, 13, 18, 20, 22, 24, and 42 in Section 1; Items 9, 12, 14, 23, and 24 in

Table 4

Distributions of Chi Square Values for all 85 Items in the Verbal Sections of Form 6K of the SAT, and for Items Not Showing DDF.

Chi Square Probability	Models A,B,C,AB,AC All	A,B,C,AB,AC non-DDF	A,B,C,AB,AC,BC All	A,B,C,AB,AC,BC non-DDF	Difference All	Difference non-DDF
Interval	f	f	f	f	f	f
.00 - .009	15	1	1	1	30	16
.01 - .049	12	12	5	2	9	9
.05 - .249	14	14	16	9	14	14
.25 - .499	19	19	26	25	14	14
.50 - .749	13	13	21	21	9	9
.75 - .949	7	7	14	11	9	9
.95 +	5	5	2	2	0	0
Total	85	71	85	71	85	71

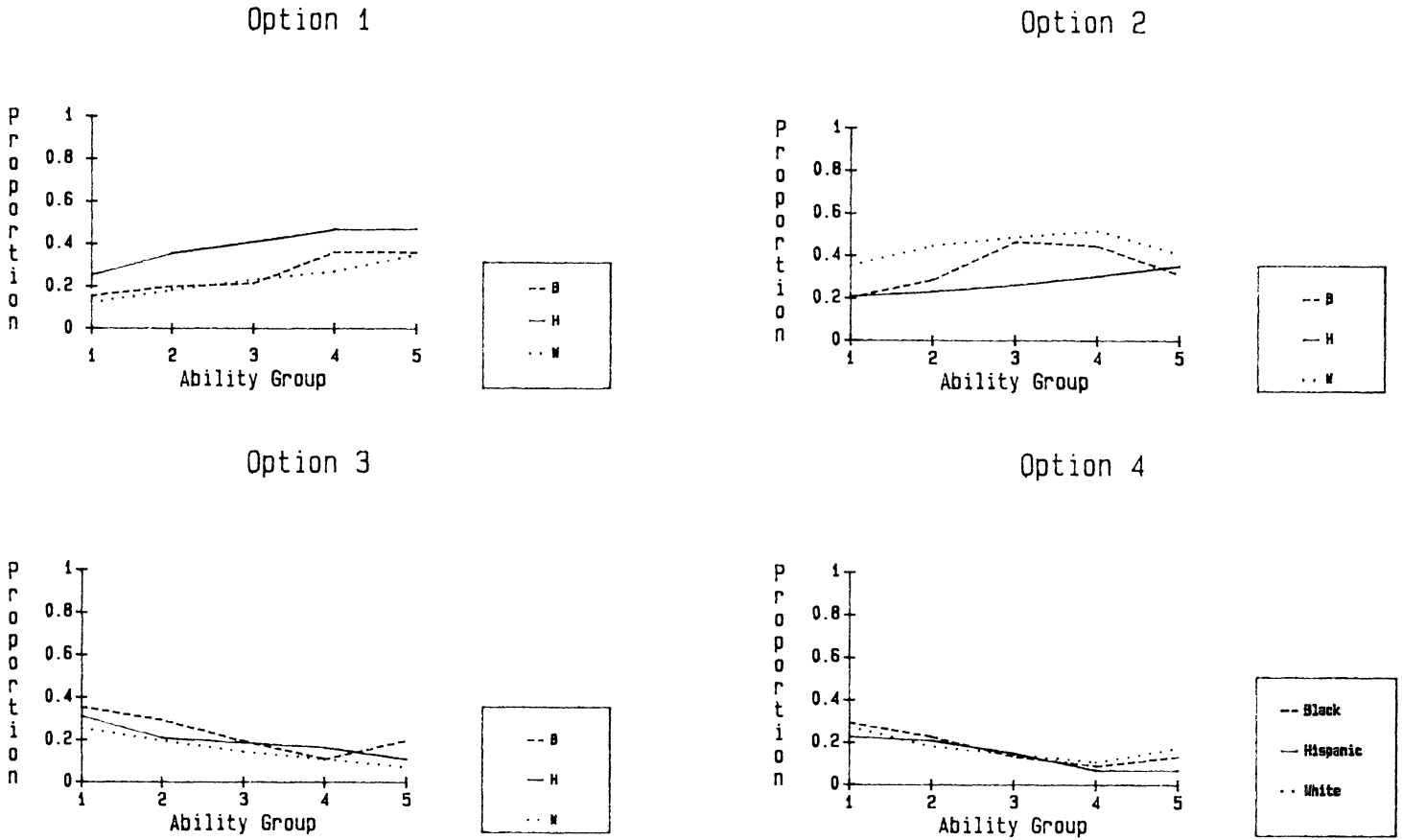

FIGURE 2. *Proportion of three matched groups, at five ability levels, choosing each incorrect option of Item 20. (2a) Option 1. (2b) Option 2. (2c) Option 3. (2d) Option 4.*

Section 4. The model did not fit Item 25 in Section 1, either without or with the critical interaction, so Item 25 was not considered to show DDF.

Table 4 shows a summary of the chi-square values for all items and for the remaining 70 items for the critical models. Once the items meeting our criterion are removed, the remainder have chi-square values that are about what would be expected by chance. The difference chi square is difficult to interpret. It is a legitimate chi-square statistic, with six degrees of freedom, and the chi squares are somewhat high, with correspondingly low probabilities. But we probably should not be concerned if an item that is well fit by a model without the BC interaction is fit significantly better when that interaction is included.

To give an idea of the size of the differences, Figure 2 presents the worst case; Figure 1 (described earlier) presented as an example the item showing DDF with the *smallest* significant chi-square value. Figure 2, for Item 20, Section 1, shows differences of 10%, among those answering incorrectly, as to which error is preferred. Whites prefer Option 2; Hispanics and Blacks tend to prefer Option 1. All groups seem to react about the same to Options 3 and 4.

By our criterion, 15 of the 85 items on the verbal sections showed significant DDF. When the separate item types were examined, the types fared quite differently. Statistically significant DDF was found in only 3 of the 20 verbal analogies items (15%), 5 of the 25 opposites (20%), and none of the 25 paragraph comprehension items. By contrast, 6 of the 15 sentence completion items (40%) showed DDF.

Detailed examination of the data showed that in most cases, the observed differences were very small, only 2% or 3%. In a few cases, differences of 10% were noted (as in Item 20, Section 1), but these were exceptional. For the most part, the three ethnic groups responded remarkably similarly to all the items, even where their responses were different by our statistical criterion.

A closer look at the six sentence completion items exhibiting DDF shows that in general it is the Hispanic group that is different. On only one item is the Black group disparate, and on no item are the Black and Hispanic groups alike yet different from the White group. We suggest the obvious hypothesis that differences in Spanish sentence structure may be at the root of the differences.

Apart from the differences among item types, examination of the content of specific items revealed no clear reasons for the differences observed. A few hypotheses might be advanced, but they would not be very convincing. In this respect, our analysis is in roughly the same position as many others who have studied DIF with the SAT (Alderman & Holland, 1981; Angoff & Ford, 1973; Angoff & Sharon, 1974; Dorans & Kulick, 1986; Harnisch & Linn, 1980; Wild & McPeek, 1986).

The main impression one gets from examining the detailed data is that the differences between the ethnic groups are small. Although the differences are statistically reliable, they are small in absolute magnitude; the large sample of cases probably contributes to the reliability of such small statistical results. Sentence completion items seem to offer more opportunities for DDF, mainly with the Hispanics. DDF seems somewhat more likely at low ability levels.

Ancillary Analyses

Some additional analyses have been done. It could be argued that 5 ability levels are too few, resulting in too much ability variation within a level. Consequently, a set of analyses was done with 10, rather than 5, ability levels. There were essentially no changes in the items found to exhibit DDF.

Another set of analysis included "omits" as a fifth incorrect item option. Certainly omits are different from one of the actual alternative answers, but they represent another way of getting the item wrong. In fact, including the omits as a fifth option leads to a few more items showing DDF. Because we believe that respondents might have a quite different rationale for omitting an item than for choosing an incorrect answer, including the omits as another option might lead to an ambiguous interpretation of DDF. We believe that omitting is a different kind of behavior, so we have not pursued the issue.

A number of additional analyses are planned but have not been completed yet. An analysis is planned in which ability is based only on the scores for item types other than sentence completion. We also wish to examine the Mantel-Haenzel analysis of the same items and a log-linear analysis of the items scored right-wrong (1-0).

Discussion

The DDF analysis provides an index of the extent to which subgroups approach an item differently. In principle, this index is statistically independent of indices of DIF, which treat all wrong answers alike, and which pit the wrong answers against the correct answer. The DDF analysis intentionally ignores the correct answers, examining only the incorrect responses.

A practical point of view about test construction holds that test scores depend on correct answers, and that *which* wrong answer someone chooses is irrelevant. Indeed, because DDF already acknowledges that people of different abilities naturally would be expected to pick different wrong answers, why not also acknowledge that people of different backgrounds might prefer different distractors. Why be concerned with which wrong answer someone chooses, since that has no effect on their test score?

We argue that a test is not ethnically blind if people with different ethnic backgrounds are attracted to different distractors. Likewise, if men prefer different distractors from women, the test is not gender-blind. When a test shows substantial DDF, that seems to us to be evidence that the test items mean something different for different groups, and consequently that the test scores cannot be interpreted in the same way for the different groups.

On the other hand, a few items with significant DDF would not be a matter for concern. Almost certainly, any given item could be made to exhibit DDF by a carefully devised method of grouping people. An item has to be about *something*. A strike-out has a clearer meaning to a baseball buff than to a cricketer. Fifty miles is a longer distance in the eastern U.S. than in the western states. Students of physical science are more likely to know the meaning of *cohere*. Only when there is a substantial pattern of DDF, as with the sentence completion items, is there cause for concern.

Notes

[1]The formula score was avoided because of the inconvenient fractional scores, which complicated the score distributions.

[2]A current practice among researchers in DIF is to establish a criterion for significant DIF by conducting a comparable analysis using equivalent samples all from the reference population. In our case, that would mean three samples of 2,000 Whites, constrained to the specified ability distribution. We have not done that analysis, and have no plans to do so. We would expect about 5% of the items to be significant at the 5% level of significance, and about 1% at the 1% level. Departures from this result would raise questions about the sampling and about the independence of item responses, which at present we have no reason to raise.

References

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the test of English as a foreign language* (Rep. No. ETS-RR-81-16). Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–106.

Berk, R. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Bond, L. (1981). Bias in mental testing. In B. F. Green, Jr. (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias.* Washington, DC: Jossey-Bass.

Cleary, T. A. (1986). Test bias: Prediction of grades of Negro and White students in integrated schools. *Journal of Educational Measurement, 5,* 115–124.

Cole, N. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

Cole, N. (1981). Bias in testing. *American Psychologist, 36*(10), 1067–1077.

Cronbach, L. J. (1976). Equity in selection—Where psychometrics and political philosophy meet. *Journal of Educational Measurement, 13,* 31–41.

Dorans, N. J. (1986). *Two new approaches to assessing unexpected differential item performance: Standardization and the Mantel-Haenszel method.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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*(4), 355–368.

Frary, R. E., & Giles, M. B. (1980). *Multiple-choice test bias due to answering strategy variation.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Green, B. F., Jr. (Ed.). (1981). *Issues in testing: Coaching, disclosure, and ethnic bias.* Washington, DC: Jossey-Bass.

Harnisch, D. L., & Linn, R. L. (1980). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133–146.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*(4), 295–303.

Linn, R. L., & Werts, E. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on X statistics. *Journal of Educational Measurement, 18*(4), 229–248.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105–118.

Osterlind, S. J. (1983). *Test item bias.* Beverly Hills, CA: Sage.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Measurement, 17*(1), 213–233.

Shepard L. A. (1981). Identifying bias in test items. In B. F. Green, Jr. (Ed.), *Issues in testing: Coaching, disclosure, and ethnic bias.* Washington, DC: Jossey-Bass.

Shepard, L. A. (1982). Definitions of bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

SPSS-X: User's Guide. (1983). New York: McGraw-Hill.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

Veale, J. R., & Foreman, D. I. (1983). Assessing cultural bias using foil response data: Cultural variation. *Journal of Educational Measurement, 20*(3), 249–258.

Wild, C. L., & McPeek, W. M. (1986). *Performance of the Mantel-Haenszel and standardization methods of detecting differential item performance.* Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Authors

BERT F. GREEN, Professor, Psychology Department, Johns Hopkins University, Baltimore, MD 21218. *Degrees:* AB, Yale University; MA, PhD, Princeton University. *Specialization:* psychometrics.

CAROLYN R. CRONE, Associate Measurement Statistician, Educational Testing Service (32E), Princeton, NJ 08541. *Degrees:* BA, Loyola College; MA, Johns Hopkins University. *Specialization:* quantitative psychology.

VALERIE GREAUD FOLK, Associate Measurement Statistician, Educational Testing Service, Princeton, NJ 08541. *Degrees:* BA, MA, PhD, Johns Hopkins University. *Specialization:* psychometrics.