ED 314 461 TM 014 330

AUTHOR Pike, Gary

TITLE The Performance of Black and White Students on the ACT-COMP Exam: An Analysis of Differential Item Functioning Using Samejima's Graded Model. Research Report 89-11.

INSTITUTION Tennessee Univ., Knoxville. Center for Assessment Research and Development.

PUB DATE 89

NOTE 31p.

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Black Students; College Entrance Examinations; *College Students; Comparative Testing; Difficulty Level; Higher Education; *Item Response Theory; Multiple Choice Tests; Objective Tests; *Racial Differences; Standardized Tests; Test Bias; *White Students

IDENTIFIERS *American College Testing Program; Binary Data Analysis; *Differential Item Performance; Graded Response Model

ABSTRACT

Responses to American College Test College Outcome Measures Program (ACT-COMP) items by 481 black and 9,237 white students at the University of Tennessee (Knoxville) were analyzed using F. Samejima's graded model to determine the level of differential item functioning (DIF). Students had been tested using Form 8 of the ACT-COMP objective test either as freshmen or as seniors. The test contains 60 multiple-choice items, each of which has two correct answers. The model developed by Samejima (1969) for graded responses, which uses a series of binary models to describe polychotomous data, was used to assess the data. Student response patterns were fitted to the graded model and five items that did not fit the model were dropped. The remaining items were analyzed using threshold parameters and their standard errors to calculate difficulty-shift coefficients. Results indicate that: (1) for 32 of the 55 remaining items, significant instances of DIF are present; (2) instances of DIF are not evenly distributed among the six subscales of the ACT-COMP test; (3) questions designed to assess explanation skills produce higher rates of DIF than do questions designed to assess skills related to identification and description; and (4) activities that rely on blueprints, require interpretation of satire, or use a radio news format to produce high levels of DIF. Four data tables and nine graphs are provided. (TJH)

CENTER FOR ASSESSMENT RESEARCH AND DEVELOPMENT

-- RESEARCH REPORT --

RR 89 - 11

The Performance of Black and White Students on the ACT-COMP Exam: An Analysis of Differential Item Functioning Using Samejima's Graded Model

by

Gary Pike

Center for Assessment Research and Development
The University of Tennessee, Knoxville
1819 Andy Holt Avenue
Knoxville, Tennessee 37996-4350
(615) 974-2350

The Performance of Black and White Students on the ACT-COMP Exam: An Analysis of Differential Item Functioning Using Samejima's Graded Model

Gary R. Pike
Associate Director
Center for Assessment Research and Development
University of Tennessee, Knoxville

Surveys of current assessment practice consistently find that colleges and universities ma.<e extensive use of student achievement data to evaluate the quality and effectiveness of their education programs (Boyer, Ewell, Finney, & Mingle, 1987; El-Khawas, 1988; National Governors' Association, 1988). These achievement data almost always are used to examine differences between institutional means and nacional norms or differences between programs at the same institution.

Differential item functioning (Dif) refers to a situation in which an identifiable subgroup performs better (or worse) on a set of test questions than do other subgroups. Such a situation represents a serious threat to the validity of the comparisons made in assessment research because differences in the performance of subgroups may produce variance in achievement scores that is not related to program quality (Thissen, Steinberg, & Wainer, 1988). Consequently, programs may be incorrectly judged to be effective or ineffective depending on whether certain subgroups are overrepresented in the programs.

The performance funding guidelines adopted by ti Tennessee Higher Education Commission (THEC) in 1983, and revised in 1986, provide an example of how differential item functioning can adversely effect assessment efforts. These guidelines currently provide a financial supplement of up to 5% of an institution's budget for instruction, and the standard on learning in general education determines one-fifth of this total, or approximately $1 million (Pike & Banta, 1987). Awards in general education are based, in part, on institutional means (national percentile ranks) on the College Outcome Measures Program (COMP) examination (Banta, 1988). In addition, the performance funding standard on corrective measures requires that institutions use subscores on the COMP exam to implement program changes that will improve total scores on the exam.

Because public institutions in Tennessee vary greatly in terms of the characteristics of their student populations, differences in the performance of subgroups on the COMP exam may significantly influence program improvement efforts and the money received through the performance funding guidelines. For example, if black students perform differently than whites on the COMP exam, judgments about program effectiveness and allocations of money will be influenced by the proportion of black students an institution tests during a given year.

Phillippi (1989) reports that performance on the COMP exam is significantly different for black and white students at the University of Tennessee, Knoxville (UTK). In separate analyses for freshmen and seniors, he finds that the mean total score on the COMP exam is 10 points lo\ r for blacks than whites, even after controlling for the effects of entering achievement levels

(ACT Assessment scores) and age. Phillippi also notes that there are significant differences in the means for black and white students on the subscales of the COMP exam. While these results strongly suggest that items on the COMP exam function differently for blacks and whites, they do not indicate which items are involved nor do they provide information about the magnitude of the differences.

Although the analysis of covariance techniques employed by Phillippi can be used to identify instances of differential item functioning, several authors suggest that the techniques of item response theory (IRT) are superior to analyses based on general linear models (Burrill, 1982; Camili & Shepard, 1987). Accordingly, the present research uses techniques from item response theory to evaluate differential item functioning for blacks and whites on the COMP exam. In the context of item response theory, differential item functioning is defined as statistically significant differences in the item characteristic curves (ICCs) for black and white subgroups (Thissen, Steinberg, & Wainer, 1988).

Methods

The Students

Analyses of the questions on the COMP exam are based on the responses of 481 black and 9237 white students at UTK who have been tested using Form 8 of the COMP Objective Test either as freshmen or as seniors. Approximately 52% (5040) of the total sample is comprised of freshmen, with 304 (6%) of the freshmen being black and 4736 (94%) of the freshmen being white. Of the 4678 seniors tested, 177 (4%) are black and 4501 (96%) are white.

The COMP Exam

In 1976, the American College Testing Program (ACT) organized the College Outcome Measures Program (COMP) to develop a measure of "knowledge and skills relevant to successful functioning in adult society" (Forrest, 1982, p. 11). Since its development, the COMP exam has been administered at least once on more than 500 college campuses, and it is used annually by approximately 100 four-year institutions in the evaluation of their general education programs (American College Testing Program, 1987).

The COMP exam is available in two forms: the Objective Test (consisting of multiple-choice questions) and the Composite Examination (consisting of multiple-choice items and exercises requiring students to write essays and record speeches). ACT staff report that the correlation between the two forms of the exam is .80, allowing the Objective Test to serve as a proxy for the Composite Examination (Forrest & Steele, 1982). Most institutions use the Objective Test because it is easier to administer and score (Banta, Lambert, Pike, Schmidhammer, & Schneider, 1987).

The Objective test contains 60 multiple-choice questions, each with two correct answers. The questions are divided among 15 separately timed activities drawing on material (stimuli) from television programs, radio broadcasts, and print media. Students taking the COMP exam are instructed that there is a penalty for guessing (i.e., incorrect answers will be subtracted from their scores), but that leaving a question blank will not be counted against them.

The combination of two correct answers for each item, the guessing penalty, and no penalty for not answering a question means that the score range for each of the 60 items is from -2 to 2 points. A score of -2 represents two incorrect answers, while a score of -1 represents one incorrect answer and one answer left blank. A score of 0 can represent either both answers left blank or one correct and one incorrect answer. A score of 1 represents one correct answers and a blank, and a score of 2 represents two correct answers. For convenience, scores for each item are recoded to produce a range from 0 to 4 points, making the maximum possible score on the Objective Test 240 points and a chance score 120 points.

In addition to a total score, the COMP exam provides three content subscores (Functioning within Social Institutions, Using Science and Technology, and Using the Arts) and three process subscores (Communicating, Solving Problems, and Clarifying Values). Content subscores may be further subdivided based on the 15 stimulus activities (five activities for each content area). For each content subscore, two of the activities require identification or description, and three of the activities require explanation (Forrest & Steele, 1982).

Process subscores can be subdivided into 20 skills (six each for Communicating and Clarifying Values, and eight skills for Solving Problems). The six skill areas for the Communicating subscore evaluate the ability to receive and send information from oral presentations, written materials, and numerical/graphic representations. The skill areas for Solving Problems and Clarifying Values represent the skills of identification and analysis (Forrest & Steele, 1982). Because the 6 subscales of the COMP exam form a matrix using the same test questions, activities requiring identification and description correspond to the skills of identification, while activities requiring explanation correspond to the skills associated with analysis.

The Dif Test

While item response theory provides a superior method of detecting instances of differential item functioning than do traditional GLM procedures, the binary item response models typically used for this purpose are not appropriate for the COMP exam with its five possible response categories for each question. Although scores on each question could be recoded to conform to a binary model (e.g., only giving credit for two correct answers), recoding the questions would change the nature of the COMP exam and sidestep the issue of whether the COMP exam, as used in performance funding, evidences differential item functioning for black and white students.

The use of ordered scores (from 0 to 4) for each question on the COMP exam suggests that a polychotomous item response model would be more appropriate for analyzing this test. Samejima's (1969) model for graded responses uses a series of binary models to describe polychotomous data. The item response functions in the graded model represent the probability of a correct response in a given category (k) and all higher categories (k~). According to Thissen (1988), the probabilities associated with a particular response function ($P_{k\sim}$) can be represented mathematically as:

$$P_{k\sim} = 1 / \{1 + \exp[-a_{k\sim}(\theta \cdot b_{k\sim})]\}$$

where $a_{k\sim}$ is the slope of the function, $b_{k\sim}$ is the threshold of the function, and θ is the latent ability or achievement level of the respondent. Because the probability for the lowest response and all higher responses is unity, *n* response categories can be described by *n-1* functions.

Thissen and Steinberg (1986) describe Samejima's model for graded responses as a difference model because the probability of a given response (k) is the differences between the probability for the function k~ and the next highest function (m~):

$$P_k = P_{k\sim} - P_{m\sim}$$

Figure 1 presents graded model response functions for a hypothetical question on the COMP exam. These functions depict two important assumptions of graded response models. First these models assume that responses are ordered (i.e., that 2 is greater than 1). If this assumption is not met at all levels of the latent ability/achievement variable, the difference formula; a will yield negative probabilities. The second assumption of the graded model is that the slopes of the functions are all equal (Thissen, 1988). Unequal slopes produce functions that will cross at some point on the ability/achievement continuum, and the difference formula again will yield negative probabilities (Thissen & Steinberg, 1986).

Insert Figure 1 about here

The twin assumptions of ordered responses and unequal slopes parallel the assumptions of one-parameter binary item response models. In one parameter models, slopes for the items are assumed to be equal (usually 1.00) and only the thresholds (item difficulty levels) vary. Because of the wide variety of tests for differential item functioning that are available for one-parameter models, treating the response functions of the graded model as a series of one-parameter models is particularly helpful (Ironson, 1982; Thissen, Steinberg, & Wainer, 1988).

Among the most popular tests for differential item functioning is the difficulty-shift statistic (Lord, 1977). This test makes use of a *z* statistic (i.e., a value for the standard normal distribution) and calculates differences in difficulty values after equating parameters onto the same latent ability/achievement scale (Ironson, 1982). The difficulty-shift statistic is defined as:

$$z = (b_1 - b_2) / (SE_1{}^2 + SE_2{}^2)^{1/2}$$

where b_1 and b_2 are threshold (difficulty) parameters, and SE_1 and SE_2 are the standard errors for the difficulty parameters.

Application of the difficulty-shift statistic to graded models provides an empirical test of the assumption that thresholds (difficulty levels) are the same across subgroups. Nonsignificant results indicate that difficulty levels are similar across subgroups, while significant results indicate that the difficulty of achieving a given score is different for the subgroups. In terms of the present research, differential item functioning is operationally defined as statistically significant differences in the threshold (difficulty) parameters for the response functions of blacks and whites on the 60 items of the COMP exam.

The Data Analyses

Analyzing students' test responses was a two-step process. First, response patterns for each item were fitted to the graded model using the MULTILOG computer program (Thissen, 1988). The responses of blacks and whites were analyzed separately, and the slopes of the response functions were fixed at 1.00. Five questions did not fit the model and were dropped from further analyses. Of these five questions, three involved the responses of black students and two involved the responses of white students. It is important to note that fixing the slopes at 1.00 was not the cause of misfitting models. For all five questions, the data did not represent ordered responses at any slope.

For the 55 questions which did conform to the assumptions of a graded model, the second step in the data analysis involved using threshold parameters and their standard errors to calculate difficulty-shift coefficients. Because of the large number of comparisons being made (220), a conservative probability level ($p < .0001$) was used. The selection of this probability level for individual comparisons resulted in an overall probability levels of $p < .05$ for all comparisons.

Interpretation of difficulty-shift results also was a multi-step process. First, results for all questions were examined and the predominant patterns of differential item functioning were identified. Second, the subscore matrix for the COMP exam was used to identify particular subscores with particularly pronounced rates of differential item functioning. Finally, the divisions of subscores identified previously were used to identify particular types of questions with consistently high rates of differential item functioning. Because of the overlap in these divisions, analyses were restricted to the identification and explanation skills of the content subscores and the oral, written, and math skills related to the Communicating subscore.

Results

Patterns of Dif

Results of the difficulty-shift analyses indicate that a substantial number of the questions on the COMP exam function differently for blacks and whites. Table 1 presents the threshold (difficulty) parameters and their standard errors for the scores of black and white students on the 60 items of the COMP exam. Asterisks (*) are used to indicate those items for which it is impossible to calculate threshold values. In addition, difficulty-shift $(z)_w$ scores are presented for each response function. Positive *z* scores identify those response functions that favor blacks, and negative *z* scores identify the functions that favor whites. Asterisks adjacent to the difficulty-shift coefficients indicate which response functions have significantly different threshold parameters for blacks and whites.

Insert Table 1 about here

As previously noted, difficulty-shift coefficients could not be computed for five of the items on the COMP exam (3, 28, 29, 42, 58) because the data for these items do not conform to the assumptions of a graded model. Of the 55 questions analyzed, 32 significantly favor whites and none significantly favor blacks. Examination of these 32 questions reveals that 11 of the questions have substantial levels of difficulty-shift (significant differences for three or four threshold parameters) and 21 of the questions have moderate levels of difficulty-shift (significant differences for one or two threshold parameters).

Figure 2 presents graphs of the four response functions for blacks and whites on a COMP question with a substantial level of difficulty-shift (question 18). Each of the four graphs contrasts the response functions for blacks and whites on this question. The item depicted in Figure 1 uses the floor plan of a house as its stimulus and asks students to calculate building and energy costs for the house. Basic computational skills (multiplication and division) are required to answer this question.

Insert Figure 2 about here

An examination of the response functions depicted in Figure 2 clearly shows that the functions for blacks are shifted to the right. This shift indicates that question 18 is significantly more difficult for blacks than whites (i.e., black students with the same ability/achievement levels as white students are more likely to make lower scores on this question).

Figure 3 presents the response functions for a COMP item (question 55) with moderate levels of difficulty-shift. Again, each of the four graphs contrasts probabilities for a given score and all higher scores for whites with similar probabilities for blacks. An examination of the graphs in Figure 3 reveals that the response functions of blacks and whites are virtually identical for scores of one or greater and scores of two or greater. However, blacks and whites differ significantly for the response functions representing scores of three or more and scores of four.

Insert Figure 3 about here

It should be emphasized that the pattern identified in Figure 3 is replicated for all 18 instances of moderate difficulty-shift in which two response functions differ significantly. For the three questions for which only one response function is significantly different, that function always represents a score of four points.

COMP Subscores

A more detailed examination of the questions evidencing significant shifts in difficulty reveals that these questions are not evenly distributed across subscores. Table 2 presents the number and percentage of items for each content and process subscale with significant difficultv-shift coefficients. This table also presents the same data broken down by the nine cells of the content-by-process subscore matrix.

Insert Table 2 about here

An examination of the data in Table 2 indicates that the Functioning within Social Institutions (FSI) content subscale has a relatively low number of items with significant difficulty-shift values. Only six (30%) of the item comprising this subscale produce significant shifts in threshold parameters, and none of these shifts occur for more than two threshold parameters. Interestingly, four of the five questions that did not meet the assumptions of a graded model are contained in this subscale.

In contrast, the Using Science and Technology (US) subscale has a large number of items with significant difficulty-shift coefficients. Sixteen (80%) of the questions contained in this subscale produce significant difficulty-shift results, and six of the questions evidence shifts in three or more threshold parameters.

Rates of difficulty-shift for the Using the Arts (UA) subscale are more moderate. Ten (50%) of the items in this subscale produce significant results. For five of these ten items, significant difficulty-shift coefficients are present for three or more of the response functions.

Concerning the process subscales, both Communicating (COM) and Solving Problems (SP) a relatively large number of questions produced significant differences in threshold parameters for blacks and whites. Twelve (67%) of the Communicating questions and thirteen (54%) of the Solving Problems questions contain significant difficulty-shift coefficients. Furthermore, five of the questions comprising the Solving Problems subscale and two of the questions comprising the Communicating subscale evidence significant shifts in three or more threshold parameters.

The Clarifying Values (CV) subscale has the fewest instances of significant difficulty-shift results of any process subscale. Only seven (35%) of these questions produce significant difficulty-shift coefficients. However, four of these seven questions do evidence significant results for three or more response functions.

As one may surmise from the results for the individual content and process subscales, the incidence of shifts in difficulty levels is not evenly distributed over the nine cells of the content-by-process subscore matrix. All six items contained in the Using Science and Communicating cell evidence significant differences in the threshold parameters of blacks and whites. Similarly, six of the eight questions related to Using Science and Solving problems produce significant difficulty-shift results, as do four of the six Using Science and Clarifying Values questions.

Four of the six Using the Arts and Communicating questions also show significant shifts in threshold parameters, as do five of the eight Using the Arts and Solving Problems questions. Only one Using the Arts and Clarifying Values questions has a statistically significant difficulty-shift coefficient, and difficulty-shift rates are stable across the three Functioning within Social Institutions cells.

COMP Activities

The table of specifications for the 15 COMP activities provides additional information on the types of questions evidencing significant shifts in threshold parameters for blacks and whites. Table 3 presents the number and percentage of items with significant difficulty-shift coefficients broken down by content area and whether the activities for that content are require identification/description or explanation.

Insert Table 3 about here

Overall, activities requiring explanation are almost twice as likely to produce significant difficulty-shift results than are activities requiring identification/description (62% versus 33%). This tendency is most pronounced for the Functioning within Social Institutions content area where all six instances of difficulty shift occur in activities requiring explanation. Similarly, 9 of the 10 instance of difficulty-shift in the area of Using the Arts occur in activities requiring explanation. For the Using Science and Technology area, rates of difficulty-shift are extremely high both for activities requiring identification/description (83%) and for activities requiring explanation (79%).

Results for the identification and analysis skills required in the Solving problems and Clarifying Values areas are identical to the results presented above because identification in the process areas corresponds to identification/description in the content areas, and analysis corresponds to explanation. For the Communicating subscale, five (83%) of the six questions designed to evaluate students' abilities in sending and receiving numeric and graphic information produce significant difficulty-shift results.

An examination of the types of stimuli students respond to in the various COMP activities also provides information about differences in the ways items function for blacks and whites. Table 4 presents the number and percentage of questions producing significant difficulty-shift coefficients broken aown by each activity. In addition, descriptions of the stimulus materials used in these activities are included in the table.

Insert Table 4 about here

As would be expected from the high rates of difficulty-shift for the Using Science and Technology subscale generally, all five acti ities related to the Using Science area produce rates of difficulty-shift of 50% or more. Particularly high rates of difficulty-shift are observed for Activity 2 (a televisior program on plant genetics), Activity 5 (a blueprint of an energy-efficient home), and Activity 11 (a radio news broadcast on the Strategic Defense Initiative).

Several other activities in the areas of Functioning within Social Institutions and Using the Arts also produce very high rates of difficulty-shift. These activities include the blueprint of a church (Activity 6) designed to measures skills related to Using the Arts, a satirical article on United States foreign policy (Activity 9) also designed to measure skills related to Using the Arts, and a radio news broadcast on marriage (Activity 10) designed to measure skills related to Functioning within Social Institutions.

Discussion

The principal findings of the present research can be summarized as follows:

1. For 32 (58%) of the 55 questions on the COMP exam that were evaluated in this research, significant instances of differential item functioning (difficulty-shift) are present. For all of these questions, differences in threshold (difficulty) parameters favor white students, indicating that these COMP items tend to be more difficult for black students than whites students.

2. Instances of differential item functioning are not evenly distributed among the six subscales of the COMP exam. For the content subscales, Using Science and Technology questions have a very high rate of difficulty-shift, Using the Arts questions have a moderate rate of difficulty-shift, and Functioning with Social Institutions questions have a relatively low rate of difficulty-shift. For the process subscores both Communicating and Solving Problems questions have moderate to high rates of difficulty shift, while Clarifying Values questions have a relative low rate of difficulty-shift.

3. When questions are categorized on the basis of the content skills they assess (i.e., identification versus explanation), the results of this research clearly show that questions designed to assess explanation skills produce

much higher rates of differential item functioning than do questions designed to assess skills related to identification and description. In addition, questions designed to assess mathematics skills produce very high rates of difficulty-shift.

4. Setting aside the extremely high rates of differential item functioning (difficulty-shift) for the Using Science and Technology subscale, examination of the nature of the stimulus materials used in the 15 COMP activities shows that activities which rely on blueprints, require the interpretation of satire, or use a radio news format produce extremely high levels of difficulty-shift.

In reviewing these results it is important to note that differences in the functioning of COMP items do not automatically lead to the conclusion that the COMP exam is biased, in a legal sense, against blacks. To be sure, these results may be the produce of bias in test construction; however, they also may be the product of differences in the educational experiences of blacks and whites. To answer this question, studies of black and white students with similar educational profiles need to be conducted.

The results of the present research also are limited in their generalizability. The fact that the data are from one institution, coupled with the relatively small number of black students in the sample, makes generalizations beyond UTK impossible. However, these results are sufficiently compelling to warrant extensive research across all public colleges and universities in Tennessee.

Despite these limitations, the results of the present research clearly indicate that items on the COMP exam function differently for black and white students at the University of Tennessee, Knoxville. For whatever reason, a majority of the items on this test are more difficult for blacks than whites. Given black and white students of equal ability/achievement, the black student will not perform as well as the white student. Stated differently, black students must have higher levels of ability/achievement than white students to make the same scores as whites on the COMP exam.

If these results can be generalized to other institutions in Tennessee, historically black students and other colleges with large black enrollments are at a competitive disadvantage with regard to performance funding. Specifically, the dollars awarded under Standard III of the performance funding guidelines will be lower for these institutions than if they had large white student populations. Even if the results of the present research cannot be generalized beyond UTK, they indicate that efforts to increase the size of black enrollment (and black retention) at UTK are inherently in conflict with efforts to improve scores on the COMP exam.

References

American College Testing Program. (1987). College Outcome Measures Program: 1987-1988. Iowa City: Author.

Banta, T. W. (1988). Assessment as an instrument of state funding policy. In T. W. Banta (Ed.), Implementing outcomes assessment: Promise and Perils (New directions for institutional research, no. 59, pp. 81-94). San Francisco: Jossey Bass.

Boyer, C. M., Ewell, P. T., Finney, J. E., & Mingle, J. R. (1987). Assessment and outcomes measurement - a view from the states: Highlights of a new ECS survey. AAHE Bulletin, 39, 8-12.

Burrill, L. E. (1982). Comparative studies of item bias methods. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 161-179). Baltimore: Johns Hopkins.

Camilli, G., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12, 87-99.

Council of Presidents and State Board for Community College Education. (1989, May). The validity and usefulness of three national standardized tests for measuring the communication, computation, and critical thinking skills of Washington state college sophomores: General report. Bellingham, WA: Western Washington University Office of Publications.

Forrest, A., & Steele, J. M. (1982). Defining and measuring general education knowledge and skills. Iowa City: American College Testing Program.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp.117-160). Baltimore: Johns Hopkins.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology (pp. 19-29). Amsterdam: Swets & Zeitlinger.

National Governors' Association. (1988). Results in education: State-level college assessment initiatives - 1987-1988. Washington, D.C.: Author.

Phillippi, R. H. (1989). An examination of item bias on the ACT-COMP Exam. Unpublished manuscript, University of Tennessee, Center for Assessment Research and Development, Knoxville.

Pike, G. R. (1989, August). A comparison of the College Outcome Measures Program (COMP) and the Collegiate Assessment of Academic Proficiency (CAAP) exams. University of Tennessee, performance funding report, Center for Assessment Research and Development, Knoxville.

Pike, G. R., & Banta, T. W. (1987). Assessing student educational outcomes: The process strengthens the product. VCCA Journal, 2(2), 24-35.

Pike, G. R., & Banta, T. W. (1989, March). Using construct validity to evaluate assessment instruments: A comparison of the ACT-COMP exam and the ETS Academic Profile. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No. 17.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 566-577.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Table 1

Threshold Parameters, Standard Errors, and Difficulty-Shift Statistics for the Response Functions for COMP Questions

ITEM		WHITE		BLACK		
		b	SE	b	SE	z2
(1)	1-	-3.376	0.053	-3.421	0.238	0.185
	2-	-2.958	0.045	-3.012	0.202	0.261
	3-	-0.376	0.025	-0.298	0.111	-0.686
	4-	0.374	0.026	0.421	0.112	-0.409
(2)	1-	-7.236	0.534	-6.721	1.501	-0.323
	2-	-5.322	0.173	-4.860	0.486	-0.896
	3-	-1.325	0.033	-0.884	0.115	-3.686
	4-	-0.750	0.030	-0.309	0.110	-3.868
(3)	1-	-7.003	0.313	*	*	*
	2-	-6.372	0.348	*	*	*
	3-	-2.165	0.035	*	*	*
	4-	-1.706	0.031	*	*	*
(4)	1-	-5.774	0.151	-5.039	0.512	-1.377
	2-	-3.890	0.064	-3.632	0.258	-0.971
	3-	-0.600	0.026	0.156	0.109	-6.747 *
	4-	0.348	0.026	1.057	0.121	-5.729 *
(5)	1-	-4.287	0.075	-3.012	0.194	-6.130 *
	2-	-2.979	0.044	-2.126	0.145	-5.629 *
	3-	0.994	0.029	1.761	0.125	-5.977 *
	4-	1.221	0.030	2.146	0.139	-6.505 *
(6)	1-	-4.511	0.084	-3.632	0.259	-3.228
	2-	-3.824	0.063	-2.911	0.195	-4.455 *
	3-	-1.080	0.027	-0.176	0.110	-7.961 *
	4-	-0.675	0.026	0.176	0.111	-7.465 *
(7)	1-	-5.227	0.144	-4.852	0.431	-0.825
	2-	-4.884	0.121	-4.431	0.357	-1.202
	3-	-2.321	0.038	-1.777	0.138	-3.801
	4-	-1.369	0.029	-0.895	0.117	-3.932 *
(8)	1-	-6.760	0.304	-8.255	70.851	0.021
	2-	-5.436	0.159	-5.225	5.071	-0.042
	3-	-1.459	0.035	-1.316	0.600	-0.238
	4-	-1.267	0.029	-1.079	0.182	-1.020
(9)	1-	-6.767	0.291	-6.036	0.921	-0.757
	2-	-5.995	0.196	-5.101	0.554	-1.521
	3-	-2.612	0.041	-2.107	0.147	-3.309
	4-	-1.848	0.034	-1.365	0.127	-3.674
(10)	1-	-3.946	0.066	-3.632	0.259	-1.175
	2-	-3.047	0.046	-2.786	0.186	-1.362
	3-	-0.516	0.026	-0.075	0.110	-3.902
	4-	-0.084	0.025	0.370	0.112	-3.956 *
(11)	1-	-2.683	0.038	-2.376	0.149	-1.996
	2-	-0.304	0.026	-0.217	0.111	-0.763
	3-	2.022	0.033	2.728	0.178	-3.900
	4-	2.701	0.040	3.421	0.232	-3.058
(12)	1-	-4.028	0.069	-3.328	0.224	-2.987
	2-	-3.144	0.048	-2.645	0.173	-2.779
	3-	-0.297	0.025	0.370	0.112	-5.812 *
	4-	1.573	0.032	2.146	0.139	-4.017 *
(13)	1-	-3.164	0.048	-2.847	0.191	-1.610
	2-	-2.756	0.042	-2.468	0.167	-1.672
	3-	-0.256	0.025	0.086	0.110	-3.032
	4-	0.059	0.026	0.473	0.112	-3.601
(14)	1-	-5.439	0.141	-5.046	0.466	-0.807
	2-	-4.326	0.080	-3.895	0.279	-1.485
	3-	-0.282	0.025	0.080	0.113	-3.128
	4-	0.074	0.025	0.486	0.113	-3.560
(15)	1-	-5.488	0.139	-4.430	0.356	-2.768
	2-	-4.821	0.101	-4.042	0.300	-2.461
	3-	-2.080	0.034	-1.254	0.124	-6.424 *
	4-	-1.189	0.027	-0.504	0.112	-5.946 *
(16)	1-	-5.129	0.113	-4.553	0.376	-1.467
	2-	-4.538	0.086	-3.753	0.266	-2.808
	3-	-1.946	0.032	-1.010	0.118	-7.656 *
	4-	-1.546	0.029	-0.401	0.113	-9.815 *

		WHITE		BLACK		
ITEM		b	SE	b	SE	z2
(17)	1-	-4.088	0.071	-3.471	0.235	-2.513
	2-	-2.970	0.045	-2.786	0.180	-0.992
	3-	0.267	0.025	1.130	0.125	-6.770 *
	4-	0.884	0.027	1.875	0.185	-5.301 *
(18)	1-	-3.687	0.061	-2.645	0.179	-5.510 *
	2-	-3.219	0.050	-2.331	0.161	-5.267 *
	3-	-1.114	0.027	-0.197	0.110	-8.096 *
	4-	-0.409	0.025	0.621	0.115	-8.752 *
(19)	1-	-2.918	0.045	-2.911	0.196	-0.035
	2-	-2.614	0.040	-2.541	0.172	-0.413
	3-	-0.772	0.026	-0.146	0.110	-5.538 *
	4-	-0.351	0.025	0.452	0.113	-6.938 *
(20)	1-	-5.439	0.147	-4.319	0.353	-2.929
	2-	-4.720	0.099	-3.576	0.253	-4.211 *
	3-	-0.757	0.026	0.005	0.109	-6.800 *
	4-	-0.019	0.025	0.827	0.117	-7.071 *
(21)	1-	-4.205	0.080	-3.048	0.202	-5.325 *
	2-	-3.187	0.050	-2.592	0.172	-3.322
	3-	-0.268	0.025	0.452	0.112	-6.274 *
	4-	0.516	0.026	1.398	0.133	-6.508 *
(22)	1-	-4.049	0.070	-3.328	0.225	-3.060
	2-	-3.285	0.051	-3.012	0.199	-1.329
	3-	-0.477	0.025	0.086	0.110	-4.991 *
	4-	0.842	0.027	1.522	0.139	-4.802 *
(23)	1-	-5.201	0.123	-4.219	0.328	-2.803
	2-	-4.066	0.073	-3.160	0.210	-4.075 *
	3-	-0.628	0.026	0.339	0.111	-8.482 *
	4-	0.159	0.025	1.229	0.127	-8.267 *
(24)	1-	-5.306	0.125	-4.692	0.400	-1.465
	2-	-4.778	0.097	-3.889	0.281	-2.991
	3-	-1.982	0.032	-1.117	0.121	-6.911 *
	4-	-1.572	0.029	-0.750	0.116	-6.875 *
(25)	1-	-4.447	0.086	-3.374	0.230	-4.370 *
	2-	-3.900	0.067	-3.048	0.202	-4.003 *
	3-	-0.394	0.025	0.452	0.112	-7.372 *
	4-	0.021	0.025	0.964	0.119	-7.755 *
(26)	1-	-2.900	0.044	-2.309	0.151	-3.758
	2-	-2.174	0.034	-1.438	0.120	-5.901 *
	3-	0.777	0.027	1.507	0.121	-5.888 *
	4-	1.398	0.031	2.107	0.141	-4.911 *
(27)	1-	-2.664	0.041	-2.353	0.160	-1.883
	2-	-2.150	0.034	-1.761	0.135	-2.794
	3-	0.452	0.026	0.685	0.116	-1.960
	4-	0.945	0.027	1.142	0.125	-1.540
(28)	1-	*	*	-5.564	0.676	*
	2-	*	*	-4.854	0.651	*
	3-	*	*	-0.696	0.112	*
	4-	*	*	-0.432	0.110	*
(29)	1-	-7.420	0.381	*	*	*
	2-	-6.345	0.320	*	*	*
	3-	-1.624	0.036	*	*	*
	4-	-1.390	0.033	*	*	*
(30)	1-	-6.302	0.282	-8.696	128.056	0.019
	2-	-5.690	0.185	-5.928	12.567	0.019
	3-	-2.371	0.037	-1.637	0.639	-1.147
	4-	-1.905	0.032	-1.252	0.274	-2.367
(31)	1-	-5.141	0.115	-4.553	0.372	-1.510
	2-	-4.256	0.076	-3.963	0.287	-0.987
	3-	-0.714	0.026	-0.390	0.114	-2.771
	4-	-0.230	0.025	0.086	0.111	-2.777
(32)	1-	-7.254	0.315	-5.269	0.679	-2.652
	2-	-6.300	0.193	-4.858	0.578	-2.366
	3-	-3.181	0.047	-2.337	0.222	-3.719
	4-	-3.017	0.044	-2.290	0.173	-4.073 *

	WHITE		BLACK		
ITEM	b	SE	b	SE	z2
(33) 1-	-3.934	0.067	-3.632	0.257	-1.137
2-	-3.362	0.033	-3.048	0.204	-1.490
3-	-0.413	0.025	-0.126	0.110	-2.544
4-	0.520	0.026	0.906	0.119	-3.169
(34) 1-	-4.711	0.131	-4.319	0.339	-1.079
2-	-4.304	0.106	-3.522	0.242	-2.960
3-	-2.058	0.036	-1.057	0.119	-8.051 *
4-	-1.602	0.033	-0.728	0.115	-7.305 *
(35) 1-	-3.635	0.058	-3.374	0.227	-1.114
2-	-2.525	0.038	-2.468	0.163	-0.341
3-	0.954	0.027	1.117	0.126	-1.265
4-	1.381	0.030	1.623	0.147	-1.613
(36) 1-	-5.914	0.215	-4.553	0.406	-2.962
2-	-5.184	0.125	-3.889	0.298	-4.007 *
3-	-2.477	0.038	-1.493	0.128	-7.370 *
4-	-1.701	0.030	-0.685	0.113	-8.690 *
(37) 1-	-1.837	0.031	-1.564	0.171	-1.571
2-	-1.528	0.029	-1.093	0.133	-3.196
3-	1.056	0.029	1.714	0.132	-4.869 *
4-	1.389	0.032	2.225	0.152	-5.382 *
(38) 1-	-5.796	0.173	-4.430	0.356	-3.451
2-	-4.630	0.093	-3.753	0.267	-3.102
3-	-1.457	0.029	-0.850	0.117	-5.036 *
4-	-0.723	0.026	-0.065	0.110	-5.821 *
(39) 1-	-7.119	0.454	-6.824	1.430	-0.197
2-	-6.154	0.269	-6.116	0.983	-0.037
3-	-2.535	0.050	-2.124	0.156	-2.509
4-	-1.766	0.072	-1.407	0.135	-2.346
(40) 1-	-4.102	0.070	-3.819	0.273	-1.004
2-	-3.267	0.050	-2.786	0.182	-2.548
3-	0.280	0.026	0.952	0.120	-5.473 *
4-	0.664	0.027	1.493	0.136	-5.979 *

	WHITE		BLACK		
ITEM	b	SE	b	SE	z2
(41) 1-	-5.914	0.214	-4.692	0.387	-2.763
2-	-5.234	0.136	-4.219	0.312	-2.982
3-	-2.149	0.034	-1.216	0.121	-7.423 *
4-	-1.727	0.030	-0.794	0.115	-7.850 *
(42) 1-	*	*	-5.978	0.841	*
2-	*	*	-5.585	0.698	*
3-	*	*	-3.157	0.223	*
4-	*	*	-2.813	0.194	*
(43) 1-	-5.752	0.159	-5.564	0.599	-0.303
2-	-4.821	0.100	-4.555	0.375	-0.685
3-	-1.947	0.032	-1.411	0.128	-4.062 *
4-	-1.270	0.028	-0.674	0.114	-5.077 *
(44) 1-	-4.075	0.069	-3.962	0.288	-0.382
2-	-3.454	0.054	-3.374	0.227	-0.343
3-	-0.514	0.026	0.055	0.113	-4.907 *
4-	-0.198	0.026	0.463	0.113	-5.701 *
(45) 1-	-3.738	0.062	-3.200	0.218	-2.374
2-	-3.193	0.050	-2.757	0.184	-2.287
3-	-0.688	0.026	-0.096	0.110	-5.238 *
4-	-0.092	0.025	0.504	0.113	-5.150 *
(46) 1-	-4.417	0.080	-3.522	0.237	-3.578
2-	-3.276	0.049	-2.879	0.184	-2.085
3-	1.601	0.142	1.438	0.442	0.351
4-	2.409	0.034	2.398	0.150	0.072
(47) 1-	-3.008	0.046	-2.288	0.155	-4.453 *
2-	-2.548	0.039	-1.825	0.136	-5.110 *
3-	0.326	0.025	0.918	0.122	-4.754 *
4-	0.655	0.026	1.452	0.139	-5.636 *
(48) 1-	-1.322	0.029	-1.229	0.126	-0.719
2-	-1.095	0.028	-0.918	0.121	-1.425
3-	0.757	0.026	1.057	0.118	-2.483
4-	0.896	0.026	1.357	0.125	-3.611

		WHITE		BLACK		
ITEM		b	SE	b	SE	z2
(49)	1-	-4.251	0.074	-3.963	0.302	-0.926
	2-	-3.328	0.050	-3.241	0.222	-0.382
	3-	0.359	0.027	0.076	0.109	2.520
	4-	0.966	0.027	0.557	0.113	3.520
(50)	1-	-5.825	0.152	-5.564	0.604	-0.419
	2-	-4.888	0.097	-4.555	0.379	-0.851
	3-	-1.149	0.028	-1.536	0.131	2.889
	4-	-0.424	0.027	-0.653	0.113	1.971
(51)	1-	-4.314	0.079	-3.328	0.226	-4.118 *
	2-	-3.534	0.056	-2.944	0.194	-2.922
	3-	-0.223	0.025	0.589	0.113	-7.016 *
	4-	0.141	0.025	0.999	0.120	-7.000 *
(52)	1-	-4.435	0.082	-3.963	0.290	-1.566
	2-	-3.393	0.052	-2.847	0.185	-2.841
	3-	0.849	0.026	1.216	0.126	-2.853
	4-	1.043	0.027	1.536	0.137	-3.531
(53)	1-	-4.621	0.088	-4.553	0.376	-0.176
	2-	-4.007	0.067	-4.127	0.311	0.377
	3-	-1.002	0.027	-0.861	0.116	-1.184
	4-	0.227	0.025	0.136	0.112	0.793
(54)	1-	-6.266	0.322	-9.071	165.012	0.017
	2-	-5.099	0.168	-5.032	3.328	-0.020
	3-	-1.674	0.039	-1.240	0.325	-1.326
	4-	-1.015	0.027	-0.411	0.176	-3.392
(55)	1-	-4.511	0.149	-4.430	0.352	-0.212
	2-	-4.153	0.123	-4.042	0.297	-0.345
	3-	-1.393	0.039	0.176	0.116	-12.821 *
	4-	-1.252	0.033	0.525	0.113	-15.095 *
(56)	1-	-4.854	0.100	-4.430	0.352	-1.159
	2-	-4.468	0.084	-4.042	0.297	-1.380
	3-	-0.078	0.025	0.176	0.116	-2.141
	4-	0.213	0.025	0.525	0.113	-2.696

		WHITE		BLACK		
ITEM		b	SE	b	SE	z2
(57)	1-	-7.235	0.461	-6.731	1.334	-0.357
	2-	-7.121	0.451	-5.585	0.729	-1.792
	3-	-3.687	0.069	-2.143	0.158	-8.955 *
	4-	-3.022	0.053	-1.505	0.134	-10.527 *
(58)	1-	-5.266	0.123	*	*	*
	2-	-4.961	0.106	*	*	*
	3-	-0.194	0.025	*	*	*
	4-	-0.057	0.025	*	*	*
(59)	1-	-5.095	0.112	-4.219	0.324	-2.555
	2-	-4.157	0.073	-3.576	0.247	-2.256
	3-	-1.078	0.027	-0.805	0.117	-2.274
	4-	-0.730	0.026	-0.298	0.112	-3.757
(60)	1-	-4.364	0.079	-4.319	0.349	-0.126
	2-	-3.526	0.055	-3.522	0.247	-0.016
	3-	-0.498	0.026	-0.390	0.111	-0.947
	4-	0.102	0.026	0.217	0.111	-1.009

Table 2
Rates of Differential Item Functioning Given Subscales on the COMP Exam

Process Subscales	CONTENT SUBSCALES			
	FSI	US	UA	TOTAL
COM	2 33%	6 100%	4 67%	12 67%
SP	2 25%	6 75%	5 62%	13 54%
CV	2 33%	4 67%	1 17%	7 39%
TOTAL	6 30%	16 80%	10 50%	

Table 3

Rates of Difficulty-Shift Given the Skills Required for Content Activities

	SKILL	
Content Subscore	Identification	Explanation
FSI	0 0%	6 43%
US	5 83%	11 79%
UA	1 17%	9 64%
TOTAL	6 33%	26 62%

Table 4
Rates of Difficulty-Shift Given the Type of Stimulus Material Used in Content Activities

		DIFFICULTY-SHIFT	
Activity	Description	Number	Percentage
1	Television Film - FSI	1	33%
2	Television Film - US	3	100%
3	Television Film - UA	1	33%
4	Print Article - FSI	3	50%
5	Print Blueprint - US	6	100%
6	Print Blueprint - UA	5	83%
7	Print Letter - FSI	0	0%
8	Print Advertisement - US	2	50%
9	Print Satirical Article - UA	3	75%
10	Radio News Show - FSl	3	75%
11	Radio News Show - US	3	75%
12	Radio Music Show - UA	1	25%
13	Printed Scenario - FSI	0	0%
14	Printed Scenario - US	2	67%
15	Printed Scenario and Slide - UA	0	0%

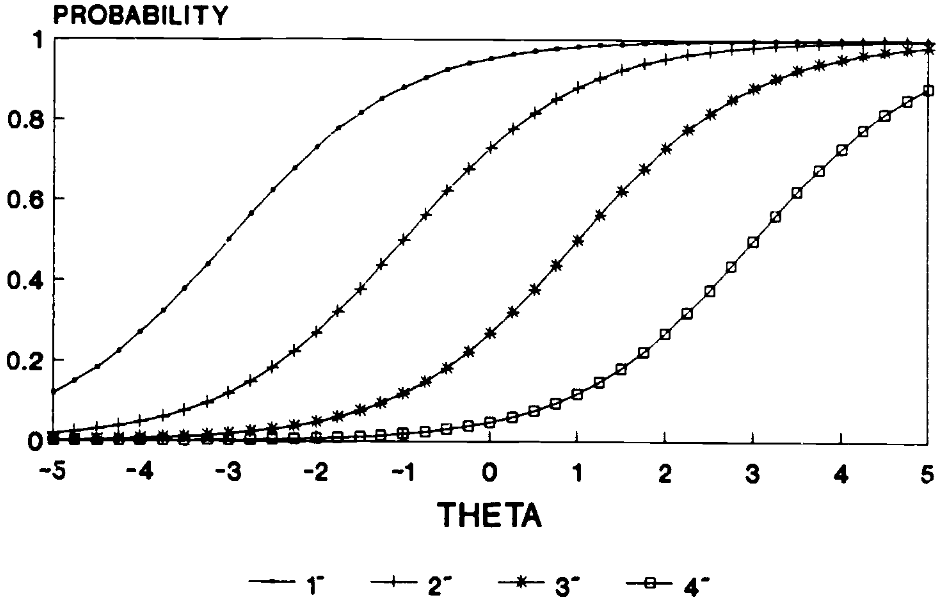
Figure Captions

Figure 1. Graded Response Functions for a Hypothetical COMP Item.

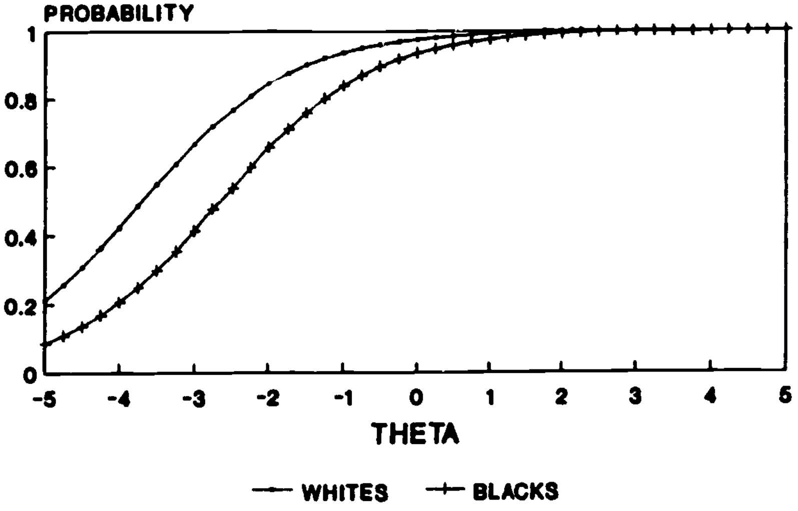
Figure 2. Response Functions for Question 18.

Figure 3. Response Functions for Question 55.

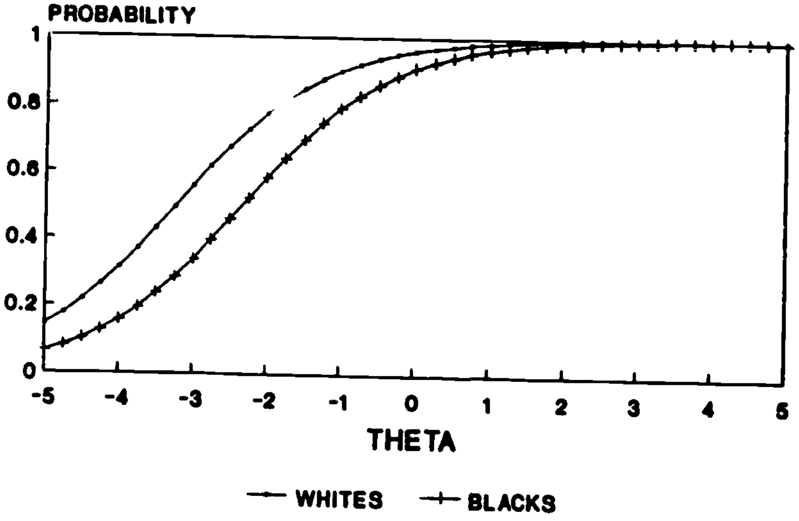
GRADED RESPONSE FUNCTIONS FOR A HYPOTHETICAL COMP ITEM



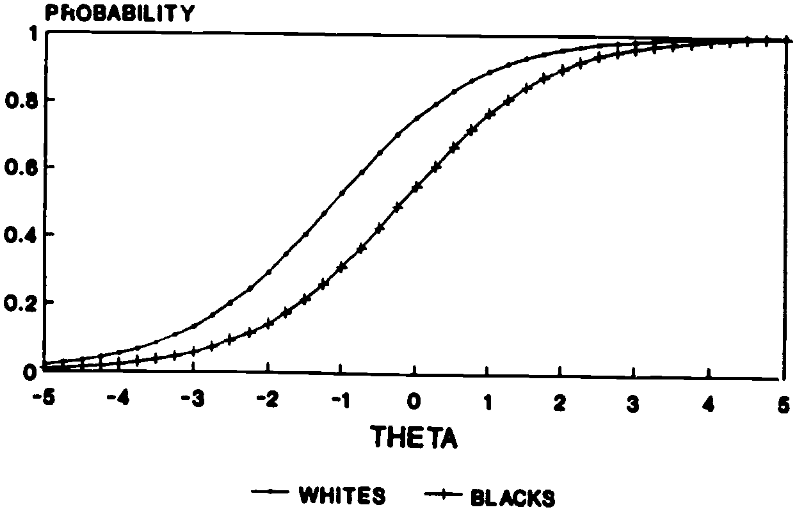
RESPONSE FUNCTIONS FOR QUESTION 18
(SCORES OF 1 OR GREATER)



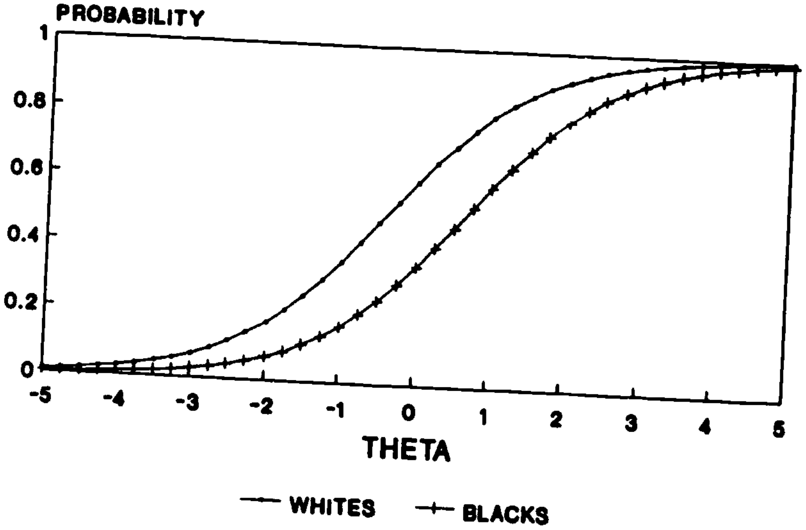
RESPONSE FUNCTIONS FOR QUESTION 18
(SCORES OF 2 OR GREATER)



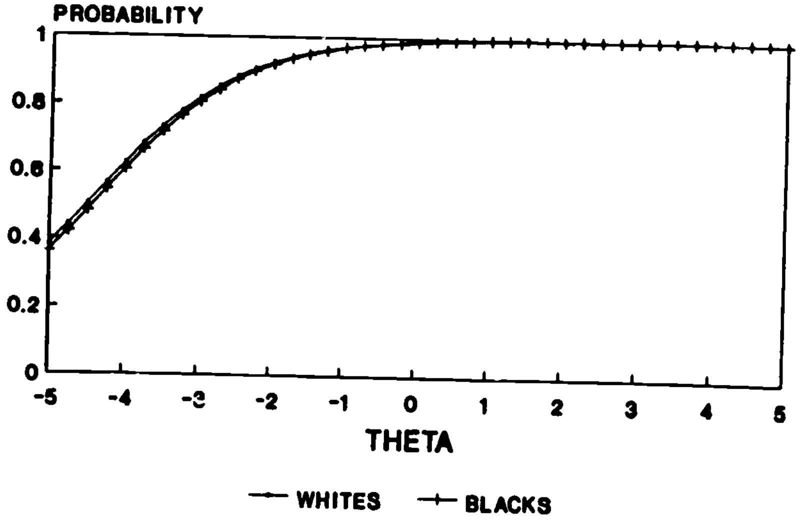
RESPONSE FUNCTIONS FOR QUESTION 18
(SCORES OF 3 OR GREATER)



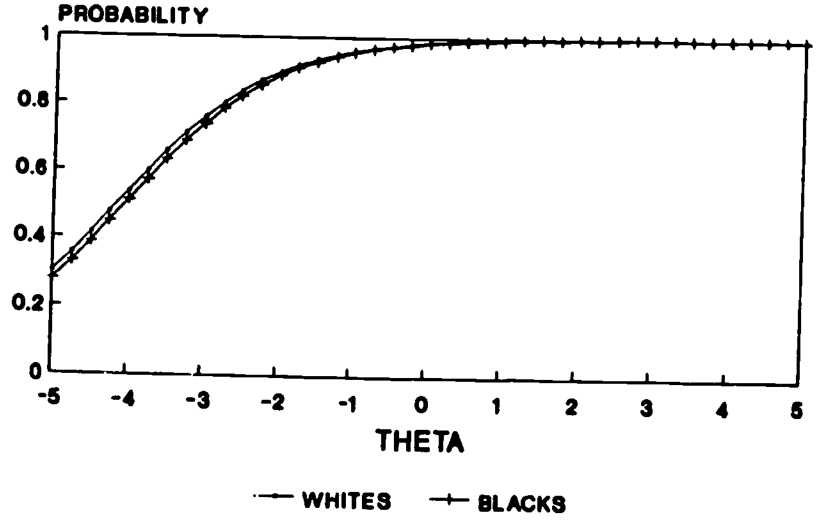
RESPONSE FUNCTIONS FOR QUESTION 18
(SCORES OF 4 OR GREATER)



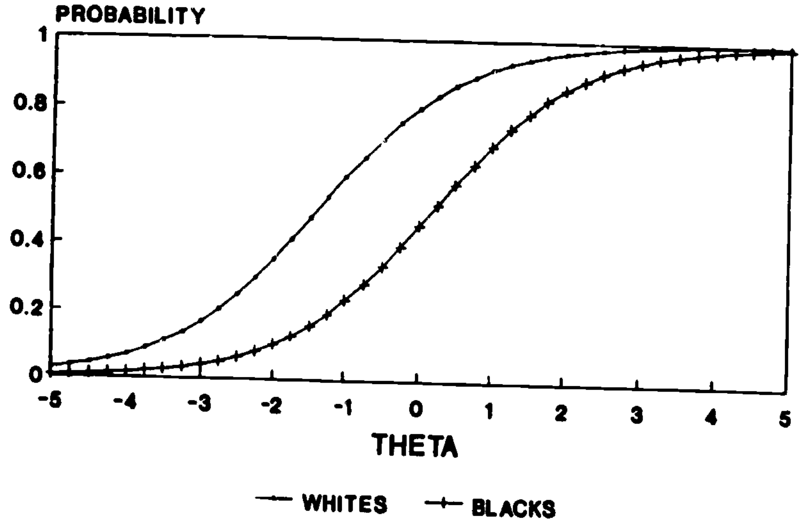
RESPONSE FUNCTIONS FOR QUESTION 55
(SCORES OF 1 OR GREATER)



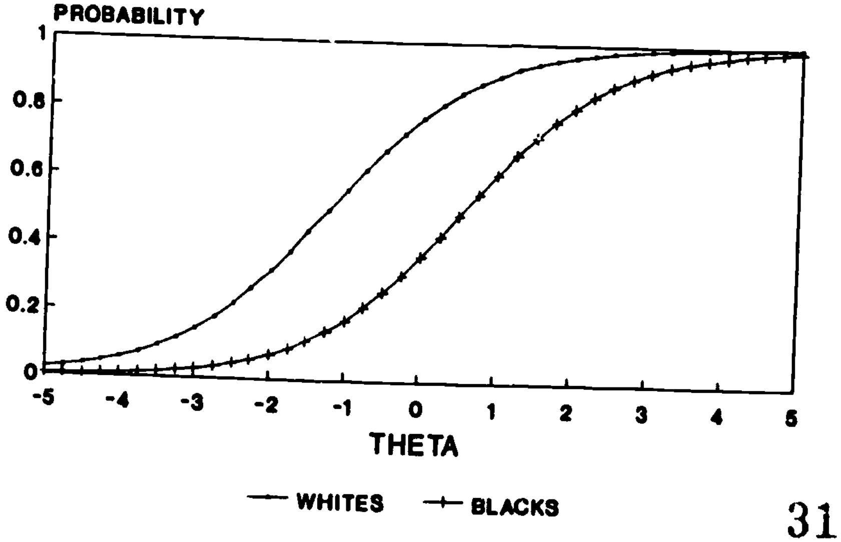
RESPONSE FUNCTIONS FOR QUESTION 55
(SCORES OF 2 OR GREATER)



RESPONSE FUNCTIONS FOR QUESTION 55
(SCORES OF 3 OR GREATER)



RESPONSE FUNCTIONS FOR QUESTION 55
(SCORES OF 4 OR GREATER)


PROBABILITY
THETA
WHITES
BLACKS