DOCUMENT RESUME

ED 389 715 TM 024 195

AUTHOR	Freedle, Roy; Kostin, Irene
TITLE	Relationship between Item Characteristics and an Index of Differential Item Functioning (DIF) for the Four GRE Verbal Item Types. GRE Board Professional Report No. 85-3P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-88-29
PUB DATE	Jul 88
NOTE	58p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Black Students; *Difficulty Level; Ethnic Groups; *Item Bias; Performance; Prediction; *Racial Differences; Reading Comprehension; *Test Items; Verbal Tests; *White Students
IDENTIFIERS	Analogy Test Items; Antonyms; *Graduate Record Examinations; Scholastic Aptitude Test; Sentence Completion Method

ABSTRACT

The first of two studies reported examined the factors that predict differences in item responses for black and white matched examinees to analogies on the Graduate Record Examinations (GRE). Data were taken from 13 forms of the GRE Verbal Test, with a median sample size of 21,000 whites and a median sample size of 1,400 blacks for the purpose of computing Differential Item Functioning (DIF) values for each test form. Three factors were found independently to predict differential item responses to 234 GRE analogies: (1) item difficulty; (2) analogy stems with a part/whole relationship; and (3) analogy stems with an "attribute" relationship. A study of 220 Scholastic Aptitude Test (SAT) analogy items found very similar results. For both GRE and SAT analogies, black examinees performed differentially better than matched white examinees on the hard analogy items. Study II used the same forms to explore the importance of item difficulty as a predictor of differential item responses of blacks and whites for other verbal item types (antonyms, sentence completions, and reading comprehension). Item difficulty was found to be an important predictor of the observed differences. Several hypotheses were advanced to account for some of the observed ethnic differences. Appendix A gives sample means and standard deviations. Appendix B lists variable labels and intercorrelations. (Contains 11 tables and 36 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 715

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

□ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

GRE®
GRADUATE RECORD EXAMINATIONS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RELATIONSHIP BETWEEN ITEM CHARACTERISTICS AND AN INDEX OF DIFFERENTIAL ITEM FUNCTIONING (DIF) FOR THE FOUR GRE VERBAL ITEM TYPES

Roy Freedle

and

Irene Kostin

GRE Board Professional Report No. 85-3P
ETS Research Report 88-29

July 1988

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

ETS

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

BEST COPY AVAILABLE

TM024195

ERIC

Relationship between Item Characteristics and an Index of Differential Item Functioning (DIF) for the Four GRE Verbal Item Types

Roy Freedle and Irene Kostin

GRE Board Report No. 85-3P

July, 1988

Educational Testing Service, Princeton N.J. 08541

Copyright © 1988 by Educational Testing Service. All rights reserved.

Acknowledgements

We wish to thank Ed Kulick for conducting the DIF value analyses for us. We also want to thank Neil Kingston and Aster Tessema from the GRE program for making the GRE data tapes available to us.

This work was supported by funds made available from the GRE Research Committee. The results for the disclosed SAT verbal items reported herein were supported by PRPC Project 969-43.

Abstract

Two studies are reported. The first examined the factors that predict differences in item responses of two populations (Black and matched White examinees) to GRE analogies. Three factors were found to independently predict differential item responses of these two groups to 234 GRE analogy items: (1) item difficulty, (2) analogy stems with a part/whole relationship and (3) and analogy stems with an "attribute" relationship. An independent sample of 220 SAT analogy items was found to yield very similar results, especially for item difficulty and part/whole relationships. For both GRE and SAT analogies, Black examinees were found to do differentially better than matched White examinees on the hard analogy items.

Since item difficulty was found to be the most important predictor variable in Study I, Study II explored the possible importance of item difficulty as a predictor of differential item responses of Black versus matched White examinees for three other verbal item types of both the GRE and SAT tests (these types were antonyms, sentence completions, and reading comprehension items). Item difficulty was found to be an important predictor of the observed differences. An overview of the results suggests that the amount of verbal context might be an important determinant of the magnitude of the relationship between item difficulty and differential performance of Black versus matched White examinees. That is, analogies and antonyms (which have minimal verbal context) produced a stronger correlation of difficulty with differential ethnic performance than did sentence completion and reading comprehension items. This was found for both the GRE and SAT verbal item types. Several hypotheses are advanced to account for some of the observed ethnic differences.

General Introduction

Purpose. The main purpose of this paper is to determine some of the factors which contribute to Black and White examinee differences on GRE verbal analogies. A secondary purpose is to explore whether these significant predictors of analogy items also can be used to account for differential ethnic examinee performance for the remaining GRE verbal item types (i.e., antonyms, sentence completions, and reading comprehension items). Finally we seek to determine the generality of these predictors by examining the results of a related analysis of SAT analogy items.

A DIF analysis (which is an acronym for Differential Item Functioning) as used here (also see Dorans & Kulick, 1986) compares the performance of two groups who have been matched on their total verbal scores on the GRE (total verbal score includes performance on four verbal item types: antonyms, sentence completions, reading comprehension as well as analogies). One group is designated as the *base* group (usually this group comprises the larger group, typically this is the White examinee group) while the second group is called the *focal* group (typically the focal group consists of minority examinees). For each item in a verbal test form, a DIF value is computed. If the DIF value is negative, this means that the focal group had a lower percentage correct than the base group (matched on total verbal score). A positive DIF value would mean that the focal group performed better than the base group (matched on total verbal score) for some particular item. For further information regarding the calculation of DIF values see Dorans and Kulick (1983).

As a result of this analysis, about half the verbal items can be expected to yield a negative DIF value and the remaining half of the items can be expected to yield a positive DIF; the sum of all the verbal DIF values in a given test form will be close to *zero*. But for a given item type [there are four item types as described above] there is no particular reason why, say, all antonyms could not have negative DIF values or, say, all of the sentence completion items might have positive DIF values. This does not mean that such patterns will actually happen, only that there is no necessary constraint that the DIF values within a given test form will have to distribute themselves in any particular pattern--all that's required is that all DIF values sum to nearly zero after all the calculations have been made for the total set of verbal items (per test form).

We now present several possible patterns of results regarding a frequency distribution of DIF value magnitudes (for *n* verbal items) in order to generate hypotheses to help guide the interpretation of the many DIF results reported below. In what follows we attempt to relate possible patterns that a distribution of DIF values might have if the *n* items were grouped by item difficulty. While doing this does anticipate one of our main findings regarding the correlation of DIF magnitudes with item difficulty, it is necessary, for reasons of clarity, to be very specific at this early point in order to make explicit some possible frequency distributions to help guide discussion of our results.

1) Early work with DIF seemed to suggest (e.g., Dorans, 1982) that items only occasionally would show either very positive or very negative DIF magnitudes and, hence most of the remaining items would have DIF values close to zero. The reason this seemed to be a likely outcome of applying this DIF statistic to a set of _n_ items is that only a few items will have escaped the scrutiny of many item reviewers (who specifically are looking for items that may favor one group over another). The DIF procedure, by this reasoning, would be one way of finding those hypothetically few deviant items that have escaped earlier detection. Using this earlier type reasoning, one might also expect that such highly deviant items might occur _randomly_ with respect to, say, item difficulty. It might be an easy item or a more difficult item. There would be little reason _a priori_ to suppose that, say, just easy items might be more susceptible to large positive or negative DIF scores than, say, hard items.

2) Another possible pattern that might emerge when comparing minority examinees with matched White examinees (who typically form the great majority of test takers) is that perhaps the easy items for each of the four verbal item types might show slightly differentially better performance (i.e., positive DIF values) by the minority examinees. If such a pattern occurs, the consequence would be that all hard items for each of the verbal item types would have to yield negative DIF values (since all DIF values must sum to nearly zero when all four verbal item types are combined). If found, this would in turn imply that the minority test takers are experiencing differentially greater difficulty with just the hard items.

3) Another possibility that could emerge is that easy items for all verbal item types might be differentially more poorly responded to by the minority examinees while all hard items (because all DIF scores in a given verbal test form must sum approximately to zero) might be differentially _better_ responded to by the minority examinees.

4) Different combinations of these patterns might also occur. That is, one might find that particular items are occasionally found that are highly deviant (in either a positive or negative direction), but one might still find that they are embedded within a general trend effect such that easy items, say, have generally small but negative DIF values while harder items may have generally small but positive DIF values. Of course one can readily postulate other types of mixed patterns.

Only an empirical examination will show which of these several possible patterns is the actual one.

ERIC Full Text Provided by ERIC

Review of earlier studies. Several studies have recently been completed at Educational Testing Service (ETS) regarding the use of DIF values (and Mantel-Haenszel values)[1] for studying Black/White differences in test performance (Dorans, 1982; Dorans, Schmitt & Curley, 1988; Freedle, 1986a; Freedle, 1986b; Freedle & Kostin, 1987; Freedle, Kostin, & Schwartz, 1987; McPeek & Wild, 1986; Rogers & Kulick, 1987; Schmitt & Bleistein, 1987; Schmitt, Bleistein & Scheuneman, 1987; also see Schmitt & Dorans, 1987). Additionally other theoretical papers dealing with DIF and/or Mantel-Haenszel have appeared (Dorans, 1986; Holland, in press; Holland & Thayer, 1986).

Freedle and Kostin (1987) studied 220 SAT analogies and found three important variables helped to account for ethnic differences in performance. The most important variable was the item difficulty level: a correlation of .502 ($p < .0001$) between DIF value and item difficulty was found. Easy analogy items tended to yield negative DIF values while hard analogy items tended to yield positive DIF values. [This same pattern was also found in an experimental study involving "thinking out loud" while solving SAT verbal analogies by Freedle, Kostin and Schwartz (1987); in their study Black and White college students were matched on number of correct analogies obtained on their experimental analogy test.] A second predictor of importance was whether the items had "science" content or not. For example *galaxy:star* and *tadpole:amphibian* would be scored as having a "science" content whereas *gullible:credulous* would be scored as having no science content. The correlation here between science content and DIF value magnitude was -.328 ($p < .01$) which indicates that items with a "science" content tend to produce a negative DIF value. A third variable helpful in predicting DIF values indicated whether the semantic relationship described in the analogy stem was of the part/whole type or not ($r = -.217$, $p < .01$). For example, *letter:word*, *tree:bark*, *stanza:poem*, *star:galaxy* and *island:archipelago* all have a part/whole relationship in the item stem (note that some of these part/whole relationships would also be scored as having a science content). The significant correlation between DIF value magnitude and part/whole relationship indicates that the presence of a part/whole relationship tends to produce a negative DIF value. A separate stepwise regression analysis showed that these same three variables independently contribute to predicting DIF value magnitudes; that is, though the three variables are intercorrelated, the stepwise procedure showed that each still accounts for independent predictive variance with regard to DIF values.

Other investigators who have also studied Black/White examinee differences in analogy performance (e.g., Boldt, 1983; Schmitt & Bleistein, 1987), while they have used smaller sample sizes (i.e., fewer items), have generally reached similar conclusions concerning "science" content. Additional studies comparing test performance of Black and White examinees have not always used the DIF statistic nor have they concentrated on just the verbal items. However, the findings of Shepard, Camilli, and Williams (1984) show that verbal *math* problems are systematically more difficult for Black examinees than White examinees. Rogers and Kulick (1987) used the DIF statistic and found that verbal math problems were differentially more difficult for Black examinees than White examinees. In addition, Bleistein and Wright (1987) found, using DIF values for comparing Asian-American examinees and White examinees, that for quantitative items, those items that had a purely mathematical content favored the Asian-American examinees (for whom English was not their best language) whereas the 'verbally loaded' mathematical items favored the White examinees.

These latter studies of quantitative items are relevant to our analysis of verbal items for the following reason: they point to the verbal dimension as critical to understanding ethnic examinee differences inasmuch as the presence of more language content even for quantitative items tends to interfere with minority examinee performance. Hence by focusing on a detailed analysis of the verbal content (e.g., science content) and structure (e.g., word class structure) of verbal items types such as analogies, we are more likely to be able to get at some basic sources of ethnic examinee difference that affect not only the verbal but the quantitative items as well.

Study I

Purpose

The purpose of Study I is to investigate whether signi-ficant predictors of DIF values between Black and matched White examinees can be found for the GRE analogy items.

Method

We selected 234 verbal analogies taken from 13 disclosed GRE test forms (forms 82-1, 82-2, 83-1, 83-2, 83-3, 84-1, 84-2, 84-3, 85-1, 85-2, 85-3, 86-1, 86-2). Each GRE form has two verbal sections each of which includes a set of nine analogies (2 x 9 x 13 = 234 items). Appendix A presents the means and standard deviations of the focal group, base group and total sample for the GRE verbal scores (GRE-V) for each of the 13 GRE test forms.

Each of the 13 GRE verbal forms was subjected to a DIF value analysis.[2] Hence each of the 234 analogy items was assigned a DIF value; a median sample size of 21,000 Whites and a median sample size of 1,400 Blacks were used to compute DIF values for each test form.

Also each of the 234 analogy items was coded[3] for the following variables:

1. Item difficulty - As our initial measure of item difficulty we used item position (as explained below). Each GRE form includes two sets of nine analogies each. The number '1' was assigned to the first analogy in each set, '2' to the second, and so on, with '9' being

assigned to the last member in each set. In each set of nine analogies, the first item is typically the easiest and the ninth item, the hardest. [Later in this paper we introduce another measure of item difficulty called "Actual Rank Difficulty"; the justification is presented in the Result section for Study I.]

2. Type of relation between the words in the stem - The relationship between the words in the stem was coded according to a thirteen category coding system, with some of the categories including a number of subcategories. Altogether (including categories and subcategories), there were 37 different codes used in this system. Examples of the types of categories used are part-whole (e.g., forest:tree) and class inclusion (e.g., flower:rose). Each item stem was coded as '1' or '0' if it satisfied or not a given analogical relationship. For example "forest:tree" would be assigned a value '1' in the appropriate column for part/whole and would be assigned a value of '0' for each of the remaining 36 additional columns. An earlier coding system (see Freedle, Kostin & Schwartz, 1987) which is very similar to the one used in the present study, yielded 72% agreement between two independent judges who coded 80 analogies from four SAT forms. [The reliabilities reported below for the other scoring variables were also based on SAT analogy items; see Freedle & Kostin, 1987.]

The following thirteen categories were used in coding each analogy item stem for its relationship:

a. Similarities: synonyms (e.g., car:auto, jump:leap).

b. Similarities: dimensional (e.g., smile:laugh, annoy:torment).

c. Opposites: antonyms (e.g., happy:sad, alive:dead).

d. Opposities: dimensional (e.g., hot:cool).

e. Modifier (e.g., green:leaf, food:tasty).

f. Functional (e.g., butcher:cleaver, patron:artist).

g. Causal (e.g., bacteria:disease, fungi:decay).

h. Conversion (e.g., grape:wine, colt:horse).

i. Action (e.g., knife:cuts, predator:hunts).

j. Class Inclusion (e.g., flower:rose, crime:theft)

k. Part-Whole (e.g., link:chain, forest:tree).

l. Class Membership (e.g., dog:bird, fork:tablespoon).

m. Quantitative (e.g., dime:dollar, inch:foot).

A description of these main categories and their sub-categories can be found in Freedle and Kostin (1987, see their Appendix A).

3. Parts of speech - Each word in the stem was coded according to whether it was a noun, a verb, or an adjective. Reliability was 100% between two judges for each of these categories as determined by coding 50 words taken from 25 analogy items.

In coding each item for parts of speech, each word in the pair of words in the item stem was coded separately. Three columns represented the three possible parts of speech for the first stem word and an additional three columns represented the three possible parts of speech for the second stem word. For example, if the first word of the stem was a noun, we coded a '1' for the noun column representing the first word and a '0' for the adjective and verb columns of the first stem word. If the second word of the stem was, for example, a verb, it was coded as '1' in a column representing verb use of the second stem word and was also coded as '0' for the remaining two columns for the second word.

4. Abstract versus concrete - Each word in the stem was coded as either abstract or concrete. A code of '1' was assigned for each concrete word ('0' otherwise). Using a 50 word sample, two judges agreed 96% of the time in coding each word as either abstract or concrete.

5. Presence or absence of science content - Each analogy stem was coded as to whether or not it contained science content. An example of an analogy stem with science content is tadpole;amphibian. A code of '1' was assigned if the pair of words in the item stem had science content ('0' otherwise). Two judges agreed 93% of the time in coding item stems as either science or nonscience: 300 items were coded.

6. Presence or absence of social/personality content -Each analogy stem was coded as to whether or not it contained social/personality type content. An example of an item with such content is gullible;creduous. A code of '1' was assigned if the pair of words in the stem had social/personality content ('0' otherwise). Two judges agreed 92% of the time for 50 item stems in classifying each stem as having social/personality content or not.

7. Frequency of occurrence - The frequency of occurrence of each word in the stem was obtained from the Francis-Kucera word frequency count (Francis & Kucera, 1982). Actually several derived variables were explored regarding word frequency: the mean frequency of the words in the stem, the log of the more frequent word, the log of the less frequent word, the log of the frequency of the first word in the stem, and the log of the frequency of the second word in the stem. The

variable that was the single best predictor was the log of the less frequent word; this is the variable reported below.

All the above variables were correlated with the DIF values using the product moment correlation.

Results and Discussion

The mean DIF value for the 234 analogy items was -.0058 (S.D. = .0495). A t-test (t = 1.8125, $p < .05$, 1-tail) was computed to determine whether this negative mean DIF value differed significantly from zero. The preponderance of negative DIF (which yields the negative mean DIF score above) is consistent with what other researchers (e.g., Dorans, 1982) have found: that Black examinees appear to have more difficulty in solving the analogy items than other types of verbal items. The next section examines which particular variables appear to contribute to this effect.

Correlations of variables with DIF value

Before we present the regression analyses, it is of interest to summarize which of the variables are significantly correlated ($p < .01$, 2-tailed) with DIF. These are shown in Table 1.

Insert Table 1 about here

The following observations can be made regarding the significant variables reported in Table 1. We first comment on those variables which yield a *positive* correlation with DIF value:

(1) There are nine item positions for analogies. [Item position is a generally good measure of item difficulty.] The correlation of item position with DIF values is .390. Generally, the early item positions (the easy items) yield negative DIF values while the later item positions (the more difficult items) yield positive DIF values. A positive DIF value for an item means that a higher percent of the Black examinees got that item correct as compared to their matched White examinees. [A negative DIF value for an item of course means that a higher percent of White examinees got that item correct as compared to their matched Black examinees.]

(2) The "attribute" variable also yields differentially better Black examinee performance vis-a-vis the matched White examinee population (that is, analogies such as *sage;wisdom* or *inexperience;neophyte*). This "attribute" variable is a sub-category of the larger Modifier category presented in the Method section above. "Attribute" in particular involves two words that function as nouns but where one of the nouns contains information that often can be seen to be a necessary attribute of the other noun; that is, while *wisdom* is a noun, one might speak of a *wise* sage (being 'wise' is here a necessary attribute of being a 'sage'). But note that in the item, 'wise' is transformed into a noun (wisdom) which is made to accompany the other noun (sage). This type of relation stands in contrast to another sub-category of the Modifier group which involves a true (i.e., adjectival) noun-attribute (e.g., *ductility:malleable* and *toxicity:poisonous* where

Table 1

Variables Most Strongly Correlated with DIF Values for GRE Analogies

Name of Significant Variable	r	Mean	S.D.
Item difficulty (item position)	.390 **	5.00	2.59
Part/whole relationship in stem	-.272 **	0.06	0.25
Attribute-nonmodifying-syntax (see text for definition)	.178 **	0.05	0.21
Science content	-.174 **	0.16	0.37
Concreteness (first stem word)	-.319 **	0.38	0.49
Concreteness (second stem word)	-.280 **	0.38	0.49
Adjective (first stem word)	.236 **	0.18	0.39
Noun (first stem word)	-.214 **	0.62	0.48
Social/personality content	.239 **	0.28	0.45
Log frequency (of less frequent word in stem)	-.275 **	0.50	0.52
Functional relationship in stem	-.206 **	0.19	0.39

** (p< .01, 2-tailed)

both examples involve two words, one of which can function as an adjective modifying a noun, as in saying "malleable ductility" or "poisonous toxicity"). Such distinctions do make a difference!

(3) We also get positive DIF values when the first of the two words in the stem is an adjective.

(4) We also get positive DIF values when the stem is coded as having social/personality content.

Regarding those variables associated with negative DIF values we have:

(1) If the stem involves a part/whole relationship Black examinees perform relatively less well than their matched Whites. (e.g., nose:face, forest:trees illustrate a part/whole relationship).

(2) Analogies with science content also contribute to differentially poorer performance by Black examinees relative to matched White examinees (e.g., tadpole:amphibian has a science content whereas sage:wisdom does not).

(3) Concreteness of both the first and/or the second word of the stem is associated with differentially poorer Black examinee performance as compared with matched White examinees. [Obviously, the converse of this finding is that when items are coded as abstract rather than concrete, this yields a positive DIF value where Black examinees perform differentially better than matched White examinees.]

(4) The occurrence of a noun as the first stem word is also associated with differentially poorer Black examinee performance as compared with matched White examinees.

(5) Regarding log frequency of the less frequent word in a stem, we also see that this yields a negative DIF value for Black examinees. This means that rare words yield better performance by Black examinees than by matched Whites while high frequency words yield worse performance!

(6) The presence of a Functional relationship in the stem yields a negative DIF value for Black examinees. An analogy is coded as functional if the words in the stem are semantically related such that one word has some function or use for the other word (e.g. butcher:cleaver, patron:artist).

Appendix B has the table of intercorrelations among many of the variables already mentioned.

Regression results using most significant predictors

In Table 2 we present the results of a stepwise multiple regression analysis (which is sometimes referred to as a hierarchical regression method) which predict the criterion variable (the DIF value magnitude) using 11 variables (presented in Table 1) as predictor variables.

Insert Table 2 about here

The results obtained thus far show that item position (which is a measure of item difficulty), part/whole relationship in stem, and the "attribute" relationship in stem were the only significant variables of the 11 selected for inclusion in the regression analysis. Two of these three predictors were also found to be significant predictors of DIF values for SAT analogies (see Freedle & Kostin, 1987; Freedle, Kostin, & Schwartz, 1987)--namely, item position (i.e., item difficulty) and part/whole relationships. [At the time that the SAT study was conducted the special sub-category which we call "attribute" had not yet been defined by us; therefore it is difficult to know whether "attribute" would have also been an important predictor for the SAT analogies.] Furthermore, the SAT results indicated that *science* content was independently contributing to the prediction of DIF values. By contrast, the current GRE analysis indicates that science content does not appear to independently contribute to the GRE analogy DIF values.

Regression results for DIF values using a different measure of item difficulty: the influence of actual rank difficulty on DIF

By virtue of test design, item position correlates quite highly with the percent of students who pass (i.e., correctly answer) a particular item within a set of analogy items. For example, using data from our SAT study (Freedle & Kostin, 1987) we find that the correlation between percent passing an item and its item position is -.935 (N=260). By contrast, for the GRE we find that the correlation between item position and percent passing is only -.805 (N=234). Because Freedle and Kostin (1987) showed that item difficulty is such a strong predictor of DIF for SAT analogies, we felt it important to explore a possibly better measure of actual item difficulty for the GRE than the one already presented (i.e., item position). To obtain this better measure of item difficulty for the GRE analogies we re-ranked each set of nine analogies so as to reflect their *actual* rank difficulty[4](based on the percent who passed each item). Below, in Table 3, we show that a substantial gain in the amount of variance accounted for among the DIF values results from introducing this new ranking.

Table 2

Multiple Regression of 234 GRE Analogies
Using Eleven Selected Variables as Predictors of DIF Value

Predictor Variable	F test df (1,222)	Multiple R	R Squared	R-sq Change	Simple R
Item position (item difficulty)	13.205 **	.390	.152	.152	.390
Part/whole	6.676 *	.442	.195	.044	-.272
Attribute	5.143 *	.468	.219	.023	.178
Science content	0.893	.477	.228	.009	-.174
Concreteness (first stem word)	0.556	.498	.248	.021	-.319
Concreteness (second stem wd.)	0.420	.504	.254	.006	-.280
Adjective (first stem wd.)	2.591	.515	.265	.011	.236
Noun (first stem wd.)	0.326	.515	.266	.000	-.214
Social/pers. content	0.796	.519	.270	.004	.239
Log freq. (the less freq. stem word)	2.262	.526	.276	.007	-.275
Functional relation	1.917	.532	.283	.006	-.206

* An F-score of F(1,222) = 3.89 yields a $p < .05$.
** An F-score of F(1,222) = 6.76 yields a $p < .01$.

The results reported here reflect a stepwise (i.e., hierarchical) regression method.

Insert Table 3 about here

In Table 3 we find that more variance can now be accounted for when we substitute actual rank difficulty for the earlier less accurate estimate of difficulty as reported in Table 2 earlier. We now account for 33% of the variance instead of only 28%. We also see that the same predictor variables are important across these two ways of measuring item difficulty. That is, item difficulty, part/whole relationship and attribute variables are still the best predictors of DIF values.

Because of the importance of the above reranked data as a truer reflection of item difficulty, we present in Table 4 below the mean and standard deviation of the DIF values associated with each of the nine rank difficulty positions.

Insert Table 4 about here

In the upper half of Table 4, t-tests are shown for the GRE analogy items for each of the nine item difficulty ranks to determine whether the mean DIF values for each position are significantly different from a mean of zero.[5] [A mean of zero would mean that Black and White examinees who are matched for verbal GRE-scores, did not differ from each other in their performance for each of these nine item ranks.] As one can see, GRE analogy items from ranks 1, 2, and 3 (namely, the easiest items) have negative mean DIF values which differ significantly from a mean of zero. That is, Black examinees do significantly worse than do White examinees with equal GRE verbal scores on the items which occur in these easiest ranks. In contrast, items in ranks 7, 8 and 9 (namely, the hardest items) have positive mean DIF values which differ significantly from a mean of zero. This means that Black examinees do significantly better on these harder analogy items than do Whites with equal GRE verbal scores.

Results exploring whether GRE analogy DIF values depend upon differences in omission rates for the two ethnic groups

There is the possibility that the calculation of DIF values could be sensitive to differential rates of omissions. While the GRE test instructions encourage students to guess (without penalty) there is some evidence that Black and White students differ in the degree to which they follow these instructions (Grandy, 1987) with the result that these two groups have different rates of omissions [Grandy's, 1987, definition of "omission" included items omitted (O) and items not reached (NR).]

Table 3

Modified Multiple Regression of 234 GRE Analogies

Predictor Variable	F test df(1,222)	Multiple R	R Squared	R-sq. Change	Simple R
Actual rank difficulty	29.297 **	.486	.236	.236	.486
Part/whole	5.841 *	.518	.269	.032	-.272
Attribute	4.712 *	.535	.286	.018	.178
Science-content	0.527	.539	.291	.004	-.174
Concreteness (first word)	0.347	.553	.306	.015	-.319
Concreteness (2nd word)	0.580	.558	.312	.006	-.280
Adjective (first word)	1.191	.562	.316	.005	.236
Noun (first word)	0.133	.563	.316	.000	-.214
Social/Pers. content	0.213	.564	.318	.001	.239
Log frequency (less freq. stem word)	1.914	.568	.323	.005	-.275
Functional relation	1.804	.573	.329	.005	-.206

* An F-score of F(1,222) = 3.89 yields a p < .05.
** An F-score of F(1,222) = 6.76 yields a p < .01.

The results for this table reflect a stepwise (i.e., hierarchical) regression method.

ERIC Full Text Provided by ERIC

Table 4

Significance of Departure of Mean DIF Values from a Mean of 0.0 for each Rank Difficulty Position for the GRE and SAT Analogies

The GRE Analogies

Actual Rank Difficulty	Mean DIF	S.D.	t-test	df
1	-.0319	.0358	4.54**	25
2	-.0540	.0511	5.40**	25
3	-.0341	.0562	3.10**	25
4	-.0108	.0575	0.96	25
5	.0038	.0450	0.43	25
6	.0161	.0476	1.73	25
7	.0228	.0264	4.38**	25
8	.0189	.0249	3.86**	25
9	.0178	.0178	5.08**	25

* p < .05, 2-tailed
** p < .01, 2-tailed

The SAT Analogies

Item Position	Mean DIF	S.D.	t-test	df
1	-.0360	.0255	6.67++	21
2	-.0482	.0403	5.60++[a]	21
3	-.0176	.0637	1.29	21
4	-.0193	.0406	2.24+	21
5	.0006	.0335	0.08	21
6	.0089	.0385	1.08	21
7	.0136	.0291	2.19+	21
8	.0171	.0172	4.62++	21
9	.0107	.0203	2.49+	21
10	.0189	.0170	5.25++	21

+ p < .05, 2-tailed
++ p < .01, 2-tailed

[a] One item at this position had the most extreme deviancy score of any of the 220 SAT analogy items; without this item the mean DIF value for the remaining 21 items at this position was M = -.0280, S.D. = .0419, t(20) = 3.08 (p < .01, 2-tailed).

In some early analyses of GRE analogies we explored whether the particular formula (see Dorans & Kulick, 1983) used for calculating DIF had a significant effect on the DIF value magnitudes. [Dorans and Kulick, 1983, define three such formulas: one which uses NR plus O plus rights (R) plus wrongs (W) in the calculations; a second which uses just O + R + W, and a third which uses just R + W.]

One GRE disclosed test form [82-2] used all three formulas for calculating DIF values for the 18 analogies. The mean DIFs (and their corresponding standard deviations) which resulted from applying each formula were as follows:

> 1. Using R + W; Mean DIF = -.0143 (S.D. = .0318). 2. Using R + W + O; Mean DIF = -.0149 (S.D. = .0326). 3. Using R + W + O + NR; Mean DIF = -.0149 (S.D. = .0327).

Furthermore the correlation of the DIFs resulting from the first two formulas was .9985; the correlation resulting from the first and third formulas was .9999; the correlation resulting from the second and third formulas was .9985.

Clearly, these results suggest that the GRE analogy DIF values are not very sensitive to whatever different omission patterns may exist between the two ethnic groups. [We should point out that Grandy's, 1987, study did not attempt directly to equate the two groups, whereas DIF analyses does directly equate them.]

Relationship of analogy DIF values for the GRE and the SAT

In an earlier study which analyzed the content of SAT analogies, Freedle and Kostin (1987) reported the nine most important correlates of SAT analogy DIF values. Since the GRE analogies were also scored for the same nine variables, we have an opportunity here to examine some similarities across two distinct sets of analogy items. The following nine variables which were found by Freedle and Kostin (1987) to be important correlates of SAT analogy DIF values are as follows:

1. Item difficulty (i.e., item position)
2. Science content
3. Social/personality content (e.g., gullible:credulous connotes personality attributes whereas bark:tree does not)
4. Syntactic content: Adjective (first stem word)
5. Syntactic content: Noun (first stem word)
6. Concreteness (first stem word)
7. Concreteness (second stem word)
8. Part/whole semantic relationship in stem
9. Log frequency of stem word with lower word frequency count.

In Table 5 we present the simple correlations for the GRE of the same nine variables which had been found to be important correlates of analogy DIF values for the SAT.

Insert Table 5 about here

We see that the nine variables which demonstrated a significant correlation between DIF value and a particular variable for the SAT analogies are also significant correlates ($p < .01$, 2-tailed) of DIF values for the GRE analogies. Hence, because the two sets of analogy data yield similar results in terms of which variables correlate significantly with DIF values, we can be reasonably confident in the stability of our findings.

In general comparison across these two large samples of verbal analogies shows us that the single most powerful correlate of DIF values is item difficulty--this is true for several ways of defining difficulty. In fact the upper and lower portions of Table 4 are remarkably similar in showing the degree to which item difficulty for the SAT and GRE analogies yield similar effects for mean DIF values for each position.

An interesting difference between the two sets of data concerns the importance of science content in accounting for DIF values. While science content is an important predictor of DIF values for the SAT analogies, even after partialling out the effects of item difficulty (Freedle, Kostin & Schwartz, 1987), it is not a potent predictor of DIF values for the GRE analogies. To state this more accurately, for the GRE, science content and DIF value are significantly correlated ($r = -.174$, $p < .01$, see Table 5). To this extent the two sets of data are

Table 5

Correlational results for GRE analogies compared with SAT[a,b]

Name of Variable	GRE correlation with DIF score (N-234)	SAT correlation with DIF score (N-220)
Item difficulty (item position)	.390 **	.502 **
Science content	-.174 **	-.328 **
Social/personality	.239 **	.261 **
Adjective (first stem word)	.236 **	.230 **
Noun (first stem word	-.214 **	-.196 **
Concrete (first stem word)	-.319 **	-.197 **
Concrete (second stem word)	-.280 **	-.236 **
Part/whole relation in stem	-.272 **	-.214 **
Log freq. of stem word with lower frequency	-.275 **	-.260 **

** (p < .01, 2-tailed)

[a] The 13 disclosed GRE forms used for this analysis were: 82-1, 82-2, 83-1, 83-2, 83-3, 84-1, 84-2, 84-3, 85-1, 85-2, 85-3, 86-1 and 86-2. The 11 disclosed SAT forms used for this analysis were: 2F, 3E, 3H, 3I, 4E, 4H, 4I, 4W, 5D, 5E and 0B023.

[b] The second measure of GRE item difficulty (actual rank difficulty) and its correlation with DIF was .486 (note that this is similar in magnitude to .502 reported above for the SAT).

similar regarding science content and DIF values. However, in the full regression analysis, science content for the GRE does not contribute *independent* variance in accounting for DIF values. Instead the variables of attribute, item difficulty, and part/whole relationship are the only important independent predictors.

A further demonstration of the importance of item difficulty following extraction of several significant correlates of DIF values for GRE and SAT analogy items

The results reported above suggest that the single best predictor of DIF values for analogies is item difficulty. A further way to demonstrate the potent effect of item difficulty is first to extract in a stepwise manner the contribution of each of those variables that individually relate significantly ($p < .01$) to DIF values and finally to extract the relationship of item difficulty.

The regression results in Table 6 show that *item difficulty still accounts for a large percent of the variance even if it is the last variable to be assessed*. It predicts

----
Insert Table 6 about here
----

an additional 8 or 9 percent of the variance of DIF values over and above the 24 percent already accounted for by the previously regressed variables.

Effects on DIF of differential ethnic omission rates for SAT analogies in comparison with GRE

Earlier we described the possible effect of different ethnic omission rates on GRE analogy DIF analyses. We concluded that the effect was too small to have any significant impact on GRE DIF values for analogies. However, for SAT analogies we have some evidence that there may be more substantial ethnic differences in omission rates. [In fact the Schmitt and Bleistein, 1987, finding that ethnic differences in "speededness" affects DIF value magnitudes for the SAT does already implicate the existence of different omission patterns for Black versus matched White examinees.] While this is certainly true for our own SAT analogy data, we have conducted a preliminary examination of these DIF results for SAT analogies *using all three DIF formulas*; we find that the general distribution of DIF values still reveals the same general pattern no matter which formula is used: easy items yield generally small negative DIFs while hard items yield generally small positive DIFs. Thus, while omission differences should be explored more thoroughly in future work especially for the SAT data, we find that it will not nullify the general finding reported above--item difficulty still predicts DIF value magnitudes even when the effects of omitted items is excluded from the DIF analyses (i.e., the R + W

Table 6

Removal of Effects of Most Important Predictors from DIF Criterion Prior to Examining Residual Effect of an Item Difficulty Measure on the DIF Criterion

Variable Name	SAT Analogy Items[a,b] Sum of Percent Variance of DIF Accounted for	GRE Analogy Items Sum of Percent Variance of DIF Accounted for
Concreteness 1st stem word	.0388	.1016
Concreteness 2nd stem word	.0677	.1162
Adjective 1st stem word	.0866	.1298
Noun 1st stem word	.0866	.1308
Social/personality	.1288	.1501
Log Freq. lesser word freq.	.1567	.1781
Science-content	.1982	.1821
Functional	--	.1894
Attribute	--	.2131
Part/whole	.2182	.2401
Item difficulty	.3210	.3287

[a] Item difficulty for the SAT was measured by the item position which is a good estimate of rank difficulty; for the GRE however it was found that determining the actual rank difficulty of the items greatly improved the relationship between DIF value and item difficulty; hence, the actual rank difficulty measure was used for the GRE calculations.

[b] For the SAT the category called Functional did not reach the required level of significance to be included among this set of predictors, and, for the category called attribute, this category relationship had not been defined by us at the time of coding the SAT data, hence it could not be used in these calulations.

formula for DIF excludes the data on omitted items). Again, for the GRE analogies we have shown that there is little evidence to suggest that different ethnic omission rates are affecting DIF value magnitudes.

ERIC
Full Text Provided by ERIC

Evaluation of distribution of extreme DIF values. We stated in the introduction several hypotheses that would help guide our interpretation of DIF value distributions. We noted that one early use of DIF statistics (e.g., Dorans & Kulick, 1983) was to assume that it would help to detect "outlier" items that depart in some dramatic way from some expected value.

There are several versions that we can frame of this "outlier" hypothesis:

(a) That outliers are randomly distributed with respect to item position (i.e., random with respect to item difficulty). That means that if one first ignores the algebraic sign of the 10 most positive and 10 most negative DIF scores, that these 20 items would not be systematically distributed with respect to item position (e.g., *a priori*, one would not expect these extreme DIF values to occur among, say, just the easy or just the hard items).

(b) Independently of the assessment of how these 20 outliers may distribute themselves regarding item position (that is, item difficulty) one would still like to know whether the negative as opposed to the positive outliers may be differentially sensitive to different places along the item difficulty continuum (as assessed by item position). The more specific hypothesis to be tested here is whether the 10 most negative DIF values are randomly distributed with respect to item difficulty (as assessed by item position). A similar test is made for the 10 most positive DIF values.

Tables 7 and 8 evaluate these hypotheses regarding how extreme DIF values (both positive and negative) distribute themselves with respect to item difficulty (as assessed by item position).

Insert Tables 7 and 8 about here

In Table 7 we see that for the SAT analogies, column A versus column B shows that the random distribution assumption for the 20 extreme DIF values is rejected (Kolmogorov/ Smirnov test, $p < .01$, 2-tailed, N = 20). That is, these 20 DIF values tend to occur among the earlier item positions (i.e., the easier items). Similarly, for the GRE analogies, the random distribution of the 20 extreme DIF values is rejected (Kolmogorov/Smirnov test, $p < .05$, 2-tailed). Again, these 20 values tend to occur among the easier items.

For Table 8 a more detailed question has been raised regarding the separate distribution of the extreme negative versus the extreme positive DIF values. For columns A versus C (for the SAT analogies) a Kolmogorov/Smirnov test ($p < .01$, 2-tailed, N-10) reveals that the most negative items are not randomly distributed over item difficulty positions. Position 2 has a heavy concentration of these most negative

Table 7

Observed and Expected Frequencies for Item Positions for the Twenty Most Extreme DIF Values (the Ten Most Extreme Positive DIF plus the Ten Most Extreme Negative DIF Values) for SAT and GRE Analogies

	A	B	C	D
	SAT	SAT	GRE	GRE
	Freq. for	Theoretic	Freq. for	Theoretic
	20 most	curve	20 most	curve
	extreme	of expected	extreme	of expected
Item	pos & neg.	freq. for	pos & neg.	freq. for
Position	DIF values	extrema	DIF values	extrema
1	1	2	1	2.2
2	6	2	4	2.2
3	4	2	6	2.2
4	2	2	4	2.2
5	3	2	1	2.2
6	4	2	3	2.2
7	0	2	0	2.2
8	0	2	0	2.2
9	0	2	1	2.2
10	0	2	--	---
SUM	20	20	20	~20

p < .01[a]
2-tailed
(N=20)

p < .05
2-tailed
(N=20)

[a] The tests of significance applies to the Kolmogorov-Smirnoff test between an observed distribution and a theoretical distribution.

Table 8

Observed and Expected Frequencies of Occurrence of Ten Most Extreme Positive DIF Values and the Ten Most Extreme Negative DIF Values for the GRE and SAT Analogies

Item Position	A[a] (SAT) freq. for ten most negative	B (SAT) freq. for ten most positive	C (SAT) expected freq.for extrema	D (GRE) freq.for ten most negative	E (GRE) freq.for ten most positive	F (GRE) expected freq.for extrema
1	1	0	1	1	0	1.1
2	6 *	0	1	4 *	0	1.1
3	2	2	1	4 *	2	1.1
4	1	1	1	1	3 *	1.1
5	0	3	1	0	1	1.1
6	0	4 *	1	0	3 *	1.1
7	0	0	1	0	0	1.1
8	0	0	1	0	0	1.1
9	0	0	1	0	1	1.1
10	0	0	1	---	---	---
SUM	10	10	10	10	10	~10
	p<.01,2-t N=10	p<.07,2-t[b] N=10		p<.01,2-t N=10	p>.20,2-t N=10	

[a] The asterisk (*) indicates those places where the observed maximum frequency occurs in the distribution (ties are indicated by two asterisks per column).

[b] The p values are for the Kolmogorov-Smirnov one-sample test which compared an observed distribution against a theoretical distribution.

DIF values. For the GRE, somewhat similar results are obtained for the 10 most negative DIF values. For columns D versus F, the Kolmogorov/Smirnov test shows that these negative values are not randomly distributed over item difficulty positions ($p < .01$, 2-tailed, N=10). Positions 2 and 3 (the easier items) contain a heavy concentration of these negative values.

Also Table 8 shows that regarding the distribution of the most positive DIF values for the SAT analogies (columns B versus C), the random distribution assumption probably can be rejected (Kolmogorov/Smirnov test, $p < .07$, 2-tailed, N=10). Position 6 (medium difficulty items) has a heavy concentration of these positive values. For the GRE analogies, the random distribution hypothesis (columns E versus F) cannot be rejected (Kolmogorov/Smirnov test, $p > .20$, 2-tailed, N=10).

Another relevant consideration here is whether one can evaluate whether the distributions of extreme positive and extreme negative DIF values are significantly different from each other. In Table 8 we find, for the SAT (columns A versus B) that the distribution of extreme positive and negative DIF values come from different distributions (Kolmogorov/Smirnov test, $p < .01$, 2-tailed, N=10). For the GRE (columns D versus E) the same test shows that the distribution of extreme positive and negative DIF values come from different distributions since they differ significantly from each other (Kolmogorov/Smirnov test, $p < .01$, 2-tailed, N=10).

Conclusion regarding extreme DIF values. Our results imply that extreme DIF values are not randomly distributed with respect to item difficulty (as assessed by item position). Also the extreme negative DIF values tend to occur primarily for the easier items. The positive DIF values may occur primarily for the medium difficulty items.

Conclusion regarding overall model of best fit for both extreme DIF values and the prevailing pattern for the remaining DIF values.

In the main, our report thus far has dealt with discovering the predictors that best account for the two samples of DIF values for GRE and SAT analogies. We have reported that the predictor dealing with item difficulty (i.e., by item position and/or by re-ranked item difficulty) is overall the single best predictor of analogy DIF values. In particular we found that most positive DIFs are associated with hard items and most negative DIFs are associated with easy items.

In light of these earlier analyses which apply equally well to SAT and GRE analogies, we have also presented evidence that a focus on just the extreme DIF values for analogies are distributed somewhat differently than the total set of DIF values. That is, while easy items tended to yield small negative values and while hard items tended to yield small positive values, one occasionally encountered large negative DIF values among the easy items and one also occasionally encountered large positive DIF values for the moderately difficult items. Hence, in terms of the hypotheses stated in our introduction, a mixed model appears to best capture our many statistical analyses of analogy DIF values. That is, the following mixed model appears to be correct: two distributions are superimposed as a result of carrying out a DIF analysis--a distribution of extreme outliers does occur and is superimposed on a second distribution of prevailing smaller DIF values such that most of the smaller negative DIF values occur for the easiest items while most of the smaller positive DIF values occur in the region of the hardest items.

Conclusions

Our analyses of 234 GRE analogy items have shown that differential ethnic performance comparing Black and White examinees matched on GRE-verbal scores can be accounted for by three variables: (1) item rank difficulty, (2) presence of a part/whole relationship in the item stem, and (3) presence of an "attribute" relationship in the stem. Black examinees perform differentially better, in general, on the harder analogy items and differentially better on those particular items which have an "attribute" relationship in the stem. Black examinees perform less well on those particular items which have a part/whole relationship in the stem.

It was noted that two of these predictors (item difficulty and part/whole) had also been found to be important predictors of DIF values for 220 SAT analogy items (see Freedle & Kostin, 1987). The differences between the two sets of data involve the significant "science" variable for the SAT DIF values and the significant "attribute" variable for the GRE DIF values (however, the reader should note that the "attribute" variable was not yet defined at the time the SAT analogies had been analyzed). These differences, especially the science variable, while they raised some questions about the generalizability of our predictors across the two analogy studies were in subsequent analyses shown to be actually quite similar in the sense that the nine significant correlates of the SAT analogy data were also found to be significant for the GRE analogy data.

We also found that only one statistical model appears to capture the details of the distribution of analogy DIF values with respect to the item difficulty variable--the mixed model (stated in the introductory section) which assumes that two distributions are imposed upon each other such that the extreme DIF values (positive and/or negative) combine with a distribution of smaller DIF values (where the easy items tend to yield small negative DIF values while the hard items tend to yield small positive DIF values). We also presented results showing that this mixed model appears to apply to both the SAT analogy DIF values as well as the GRE DIF values.

Study II

Introduction

Purpose. Our purpose in this study is to explore whether the most significant predictor of DIF value magnitudes for analogies (i.e., item difficulty) can also be implicated as an important predictor of DIF value magnitudes for the other verbal item types used in the GRE and the SAT--these other item types are antonyms, sentence completions, and reading comprehension items.

Background. The results reported above and earlier reports (see Freedle, Kostin & Schwartz, 1987; Freedle & Kostin, 1987) have shown that the single best predictor of DIF values for analogy items is item difficulty (as measured either by actual rank difficulty or by item position). A consequence of calculating DIF values for any particular verbal test form is that DIF values are assigned not only to the analogy items but to all the verbal items (e.g., antonyms, sentence completions and reading comprehension items as well as the analogy items) in a given test form. Because of this, it is immediately possible to explore the relationship of DIF values and item difficulty (as a predictor of DIF) _for each of the three remaining verbal item types: antonyms, sentence completions, and reading comprehension items_.

Method

For each of the 13 disclosed GRE verbal test forms (same as used in Study I) DIF values were calculated for each of the verbal items using the Dorans and Kulick (1986) method.

For each of the 11 disclosed SAT verbal test forms (as reported in Freedle, Kostin & Schwartz, 1987) DIF values were also calculated using the Dorans and Kulick (1986) method.

Coding categories for the antonyms, sentence completions, and reading comprehension items.

1. For the GRE test forms the actual rank difficulty was again used for the antonym, sentence completion and the reading comprehension items.

2. For the SAT test forms the actual rank difficulty was used only for the reading comprehension items (this was necessary because the items for this item type are not typically presented in order from easy to hard as is true for the remaining verbal item types). For antonyms and sentence completion items, however, we retained item position as the index of item difficulty to be consistent with the index used for SAT analogy item analyses.

3. As a check on the adequacy of rank order as a measure of relative difficulty across item types, we have also computed for each item type the average DIF value (for the GRE) for each of the following intervals of percent correct: 100 - 90, 89 - 80, 79 - 70, 69 - 60, 59 -50, 49 -40, 39 -30, 29 - 20, 19 - 10. That is, it is not clear that correlation of DIF with item position (or reranked item difficulty) can be directly compared across the four item types, especially since the number of item positions (or rankings) varies from one item type to another. Hence, a direct measure of average DIF value with percent pass was sought to help answer this question.

Results and Discussion

In Tables 9 and 10 one can see a very interesting pattern: the correlation of item difficulty with DIF value is quite strong for the item types which have virtually no context (i.e., the analogy and antonym items appear in very limited verbal contexts) and the correlation of item difficulty with DIF value is much less powerful for the remaining two item types (sentence completion and reading comprehension) both of which involve the presence of larger amounts of context in the item. While all four of these correlations show a significant relationship between DIF value and difficulty, nevertheless the varying magnitudes of the correlational effect across the four item types requires further investigation regarding what underlying variables may be accounting for this.

Insert Tables 9 and 10 about here

Upon examining Table 10 we can immediately dismiss the possibility that these different magnitudes of correlations are due to different average ranges of difficulty across the four item types (e.g., for any given test form typically the analogies and antonyms have items that extend into the very difficult range, while reading comprehension and sentence completion items have very few of these harder items represented). In particular Table 10 shows, moving down the rows, that for each decrement in proportion correct the average DIF value moves from generally negative DIF values to generally positive DIF values for each of the four item types. It is also clear from Table 10 that the strongest effects occur for antonyms and antonyms in the sense that the largest average negative DIF values occur here for the easy items while the largest average positive DIF values tends to occur again for these same two item types for the hardest items. Table 10 also shows (see footnote b, Table 10) that when the correlations between percent pass and DIF value is recomputed, the general pattern still suggests that the item types with little context produce the larger correlations between DIF and difficulty while the two item types with more context produce the smaller correlations.

Table 9

Correlation of DIF Values with Item Difficulty for Four Verbal Item Types for the GRE

Item Type	Correlation between DIF[a] Value & Actual Rank Difficulty	Sample Size
Analogy	.486 **	234
Antonym	.410 **	286
Sentence Completion	.159 *	182
Reading Comprehension	.179 **	286

* p < .05, 2-tailed
** p < .01, 2-tailed

[a] A given verbal test consists of two sections. For the set of items of a given item type within each section, items are ranked for relative difficulty. Thus there are two sets of ranked items for each item type per test form.

Table 10

Average DIF Values for Narrowly Defined Levels of Difficulty for Each of Four GRE Verbal Item Types

Verbal Item Type[a]

Probability Interval[b]	Analogy			Antonym			Sentence Completion			Reading Compre.		
Item correct	Mean DIF	SD	N	Mean DIF	SD	N	Mean DIF	SD	N	Mean DIF	SD	N
1.00 - .90	-.04	.05	21	-.04	.03	21	-.01	.03	20	-.02	.03	11
.89 - .80	-.05	.06	16	-.02	.05	23	-.00	.03	15	-.02	.03	26
.79 - .70	-.02	.07	07	.01	.07	11	.01	.04	15	-.00	.04	32
.69 - .60	-.03	.05	12	.02	.03	07	.01	.04	15	-.00	.04	28
.59 - .50	.01	.06	11	.03	.03	17	.01	.05	12	-.01	.03	23
.49 - .40	.02	.04	25	.02	.02	22	.00	.03	11	.01	.03	19
.39 - .30	.02	.03	18	.03	.02	20	.02	.02	11	.01	.02	11
.29 - .20	.02	.02	13	.02	.02	24	.01	.02	08	.01	.03	08
.19 - .10	.01	.01	07	.02	.01	12	.01	.00	02	.00	--	01

[a] Seven GRE disclosed test forms were used to obtain most of the data reported in this table; to obtain more stable estimates of Mean DIF for the hardest items, six additional GRE forms were used to increase the number of observations for hard items (defined as items between .10 and .29).

[b] The correlation between percent passing each item and DIF (for the data reported above) are: analogies -.46, antonyms -.46, sentence completions -.13 and reading comprehension -.24. These values are comparable to those reported in Table 9 of this report.

By way of explanation for the hypothesis concerning amount of context as influencing the pattern of correlations for the four item types, Freedle and Kostin (1987) showed that easy analogy items often employ high frequency words while hard analogy items often use low frequency words; they also suggested that high frequency words often have multiple meanings (i.e., have many different dictionary senses). In fact the Schmitt and Bleistein (1987) finding regarding what they called "homographs" supports our claim that words with multiple meanings are more likely to be found among the easy analogy items.

Freedle and Kostin (1987) further noted that subcultures might very well differ in how they interpret the meaning and significance of high frequency words (as compared with low frequency words). That is, since members of different subcultures have different backgrounds, they might experience differential difficulty in interpreting the intended senses of especially high frequency words (which are typically used in easy items); if so, then they should perform differentially more poorly on items involving such words.[6] That is, they will get negative DIF values on such items. [Relevant examples of high versus low frequency GRE analogy words revealing different levels of multiple meanings are presented in the extended footnote 6 cited above.]

However, items which involve more verbal context, such as the easy Sentence Completions items, can still use high frequency words (which in isolation have multiple meanings in the sense that they are open to several interpretations) with the result that their interpretation will be less ambiguous (because there is more context around these words to disambiguate them). Therefore DIF values for such easy Sentence Completion items might still be negative, but they will not be as extreme in magnitude as compared with DIF values for easy items taken from verbal types involving little verbal context (such as analogies).

To put it another way, since the amount of context surrounding words with multiple meanings is known to reduce the amount of uncertainty (Miller, Heise & Lichten, 1951), the fact that DIF values correlate more strongly with low context verbal item types (analogies and antonyms) than with high context verbal item types (sentence completions and reading comprehension items) suggests that it is the augmenting and diminution of multiplicity of meanings that may be operating differentially across these four item types.

Some recent work by Schmitt and Bleistein (1987), briefly mentioned above, report that a measure called "homographs" (homographs are words which are spelled alike but have different meanings) helps to account for some analogy DIF values. That is, they found, for items which involve homographs in the stem or in the correct option, that Black examinees perform differentially worse on such items compared to matched White examinees.

Further work relating the number of dictionary definitions for each word in analogies to item difficulty (and its possible contribution to item DIF value) would be desirable to further specify the possibly different ways in which Black and matched White examinees interpret the meanings of particular words used in analogies. In the absence of such additional work, we believe that the Schmitt and Bleistein (1987) result regarding homographs lends general support to our suggestion that number of dictionary definitions (and hence cognitive uncertainty) might be related to item difficulty for a given verbal item type (e.g., within the set of analogies). While further work on homographs and number of dictionary definitions for analogy words might alter our conclusions, at present we believe that these ideas can further be used to explain the differences in the DIF value findings among the four verbal item types reported above (e.g., help to explain why number of dictionary definitions might play less of a role for item types that present more verbal context).

Effect of Omissions for the Four GRE Verbal Item Types.

Above we showed that the analogy DIF values alone were not sensitive to omission rates across the Black and White comparison groups. Here we inquire whether there is any evidence that the DIF values for all the verbal item types considered together might show evidence for overall sensitivity to omission rates. To test this we again used the three DIF formulas for the same GRE disclosed test [82-2] described above. The mean DIF and standard deviations for all four item types (n= 76) using the three formulas are as follows:

1. R + W; Mean DIF = .C003 (S.D. = .0306).
2. R + W + O; Mean DIF = -.0009 (S.D. = .0308).
3. R + W + O + NR; Mean DIF = -.0001 (S.D. = .0286).

We also correlated the DIF values for each of these formulas. The correlation of the first two formulas was .9956; the correlation of the first and third formulas was .9848; the correlation of the second and third formulas was .9875.

Clearly, these results suggest that for the GRE, the different rates of omission (and/or NR's) for the two ethnic groups does not have any significantly detectable effect on DIF values. Therefore the conclusions reached above concerning the effect of degree of item context on the relationship between item difficulty and DIF values does not seem to be subject to modification when possible differential rates of omission are also taken into account. It is possible of course that our analysis of just one GRE form (for all three DIF formulas) may have unestimated the possible effects of differential omission rates on DIF for all the verbal item types. Only further study can resolve this issue.

A comparison of GRE and SAT for the Four Verbal ltem Types

To see whether the pattern of correlations presented in Table 9 yields a stable result, we also examined the SAT DIF values for each of four item types. Thse results are presented in Table 11 below.

Insert Table 11 about here

In general the results presented in Table 11 appear to be very similar to those noted for Table 9. That is, the largest correlations for the SAT data are obtained for the analogies and antonyms, both of which again are item types that have very little contextual information presented. Also, for the two item types which do involve more contextual information, the correlations of DIF value with item difficulty is attentuated (especially for the reading comprehension items) in comparison with the analogy and antonym correlations. Again this result is consistent with our explanation of the intersection of ethnic differences with the precise interpretation of word meanings as a function of the amount of context present in the different item types.

We have briefly examined SAT data for possible effects of differential omission rates for the two ethnic groups. Our preliminary conclusion, especially for the SAT analogy items, is that there is more of a detectable effect due to omission rates on DIF value calculations (also see Schmitt & Bleistein, 1987). However, we still find that the general pattern of DIF values with respect to item difficulty are virtually the same regardless of the formula used to calculate DIF: that is, easy items still yield generally small negative DIF values while hard items still yield generally small but positive DIF values.

Because our results for omission rates are tentative, we leave open the possibility that it may be fruitful to study the effects of differential omission rates on DIF values in a more detailed way (for the SAT certainly) for each of the verbal item types.

Conclusions

Study II examined the importance of item difficulty in predicting DIF value magnitudes for three additional verbal item types (from both the GRE and SAT tests). This result agrees with those of Study I in that item difficulty indices were also found to be important predictors of analogy DIF values for the GRE and the SAT.

Study II also allowed us to conclude that the degree of context is probably an important modulator of the magnitude of the relationship between DIF and item difficulty in the sense that items such as analogies and antonyms (which have little context) yield stronger correlations between DIF and difficulty than do item types (such as

Table 11

Correlation of DIF Values with Item Difficulty for Four Verbal Item Types for the SAT Test of Verbal Ability

Item Type	Correlation of DIF Value with Item Difficulty[a,b]	Sample Size
Analogy	.502 **	220
Antonym	.332 **	200
Sentence completion	.257 **	165
Reading comprehension	.096 (*)	275

* p < .05, 2-tailed
** p < .01, 2-tailed
(*) p < .05, 1-tailed (approx.)

[a] The correlations reported for antonyms and reading comprehension were based on averaging for each item type two separate estimates of these correlations based on earlier analyses which used subsets of these items--all items, however, were used in these subsets.

[b] The measure of item difficulty used for the SAT test for analogies, antonyms and sentence completions was the item position information (which we have noted is highly correlated with item difficulty for the SAT); for the GRE however, as noted earlier in this paper, we have found that the actual (computed) rank difficulty is a better estimate of difficulty than is the item position information. The correlations for both the SAT and GRE Reading Comprehension items were always based on the actual (computed) rank difficulty; this is because the order of presentation in these test for reading comprehension is not intended to bear any relationship with item difficulty.

sentence completions and reading comprehension) which involve larger contexts. We argue for both the GRE and SAT tests that Black examinees may have differentially greater difficulty as compared with matched White examinees with those high frequency words which have multiple meanings--these being typically words which occur more often among the easy items.

While further study of differential ethnic omission rates is called for, we failed to find evidence, at least for the GRE verbal item types, of any significant contribution of differential ethnic omission rates on the DIF value magnitudes.

General Conclusions

Two studies have been reported that explore some of the properties which influence differential ethnic responses for the four verbal items types of the GRE.

In particular, for 234 GRE analogies, we find that item difficulty, a part/whole relationship and an attribute relationship independently predict DIF value magnitude. In a comparison study using 220 SAT analogies we find confirming evidence for item difficulty and a part/whole relationship. Overall, item difficulty indices prove to be the best predictor of DIF value magnitudes in both data sets.

We find evidence that the distribution of DIF score magnitudes for analogies favors a "mixed model" in the sense that there was evidence for a separate distribution of extreme DIF values (either positive or negative DIFs) which was superimposed upon another distribution of smaller DIF magnitudes--this latter distribution involves small but consistently negative DIFs fo: easy items and small but consistently positive DIFs for hard items. These distribu-tional effects exist for both the GRE and SAT analogies.

For Study II we examined the degree to which item difficulty is a good predictor of DIF magnitudes for the three remaining verbal item types: antonyms, sentence completions, and reading comprehension items. This was done for both GRE and SAT data sets. We find that item difficulty is a significant predictor of DIF values in both data sets; again, for each of these additional verbal item types, hard items are typically differentially better responded to by Black examinees while easy items are typically differentially better responded to by matched White examinees. Additionally we find that for item types with minimal verbal context (i.e., the analogies and antonyms) the magnitude of the correlation between DIf and difficulty is greater than for item types with more verbal context (i.e., the sentence completion and reading comprehension items). This relationship exists for both the GRE and SAT item types.

Several hypotheses were advanced about why item difficulty might be a predictor of differential item responding in the several data sets. We maintain that word frequency measures are an indirect index of the extent to which words may have multiple dictionary senses. Furthermore, developmental studies reported by Hall, Nagy and Linn (1984) suggest to us that some of the word types used to construct easy analogy items are used less frequently by young Black children (also see footnote 6), especially working class (as opposed to middle-class) children. We further argue that increased amounts of verbal context diminish the number of possible meanings that high frequency words can have.

While no significant effect of differential omission rates on each of the four verbal item types could be detected for the GRE data (as reported in both studies above), we suggest that further work may be needed to study its effect on DIF magnitudes at least for the SAT verbal item types since there is more evidence that significant differential ethnic omission rates occur for that test. However, preliminary analyses of the effects of omissions on SAT verbal items strongly suggests that the correction on DIF estimation when omits are eliminated frcm the calculations does not alter the general results reported above: item difficulty still appears to be a significant predictor of item difficulty for each of the four SAT verbal item types.

Practical implications. While it is true that the absolute magnitude of DIF values for any particular test item seldom is larger than about -.05 (where Blacks perform about 5 percentage points lower than matched Whites) or +.05 (where Blacks perform about 5 percentage points higher than matched White examinees), we must keep in mind that there might be a cumulative effect that these DIF values have, especially if one focuses on performance on all the harder items for all four verbal item types. It would seem to be the case that if one were to sum the percentage correct of each Black examinee on just the harder verbal items, evidence might be found to support the idea that they might be significantly outperforming their purportedly matched White counterparts for the entire verbal test section (since many hard verbal items yield positive DIF values for each of the four verbalitem types).

However, counterbalancing this idea is the other viewpoint: if one focused on individual Black examinee performance on just the easier verbal items for all four verbal items types, they might then appear to do significantly worse than their purportedly matched White counterparts (since many easy verbal items yield negative DIF values for each of the four verbal item types). One would have to ask which of these several ways of calculating verbal performance level is the most accurate indicator of performance in some criterial setting such as that provided by grades earned in graduate school.

References

Bejar, I., Embretson, S., Peirce, L. & Wild, C. (1984). Applying cognitive research results in GRE Test Development: Analogies. GRE Research Proposal, #84-19, Princeton, N.J., Educational Testing Service.

Bleistein, C. A. & Wright, D. (1987). Assessment of unexpected differential item difficulty for Asian-American examinees on the Scholastic Aptitude Test. In A. Schmitt & N. Dorans (Eds.), *Differential item functioning on the Scholastic Aptitude Test*. ETS Research Memorandum RM-87-1.

Boldt, R. F. (1983). *Status of research on item content and performance on tests used in higher education.* Research Report ETS RR-83-3. Princeton, N.J.: Educational Testing Service.

Chaffin, R. & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory and cognition*, *12*, 134-141.

Dawis, R. V., Soriano, L. V., Siojo, L. R. & Haynes J. (1974). Demographic factors in the education of relations in analogy word pairs (Technical Report 3). Minneapolis, Minn.: Univ. of Minnesota, Dept. of Psychology.

Dorans, N. (1982). *Technical review of SAT item fairness studies: 1975-1979*. Unpublished statistical report SR-82-90.

Dorans, N. (1986). *Two new approaches to assessing unexpected differential item performance*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, Ca.

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

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

Dorans, N., Schmitt, A. & Curley, W. E. (1988). Differential speededness: some items have DIF because of where they are, not because of what they are. Paper presented at AERA, New Orleans, La.

Francis, W. N. & Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Houghton Mifflin, Boston: Mass.

ERIC
Full Text Provided by ERIC

Freedle, R. (1986). Second interrim report: Report on results of a semantic scoring of 4 analogy subtests. Princeton, N.J.: ETS, manuscript dated Jan. 10, 1986 (a).

Freedle, R. (1986). Preliminary findings of two projects dealing with application of a statistic for comparing Black and White examinees of SAT analogies, GRE analogies (with additional findings reported for SAT anatonym, SAT sentence completion, and SAT reading comprehension items). Princeton, N.J.: ETS unpublished report dated May 16, 1986 (b).

Freedle, R. & Kostin, I. (1987). Semantic and structural factors affecting the performance of matched Black and White examinees on analogies items from the Scholastic Aptitude Test. Princeton, N.J.: Educational Testing Service, Research Report Final Report, PRPC project, submitted August, 1987.

Freedle, R., Kostin, I. & Schwartz, L. (1987). A comparison of strategies used by Black and White students in solving SAT verbal analogies using a thinking aloud method and a matched percent-correct design. ETS Research Report RR-87-48. Princeton, N.J.: Educational Testing Service.

Grandy, J. (1987). Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring. Research Report 87-38. Princeton, N.J.: ETS.

Hall, W. S. & Freedle, R. (1975). Culture and language; The Black American Experience. Wash.,D.C.: Hemisphere/Halstead/Wiley Press.

Hall, W. S., Nagy, W., Linn, R. (1984). Spoken words; effects of situation and social group on oral word usage and frequency. Hillsdale, N.J.: L. Erlbaum & Assoc.

Holland, P. W. (in press). On the study of differential item performance without IRT. Proceedings of the Military Testing Association.

Holland, P. W. & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel statistic. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Ca.

Kok, F. G. & Mellenbergh, G. J. (1985). A mathematical model for item bias and a definition of bias effect size. Paper presented at the Fourth European Meeting of the Psychometric Society and the Classification Societies, Cambridge, England.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

McPeek, W. M. & Wild, C. L. (1986). Performance of the Mantel-Haenszel statistic in a variety of situations. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, Ca.

Miller, G. A. (1951). Language and communication. New York: McGraw-Hill Co.

Miller, G. A., Heise, G. A. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the text materials. Journal of Experimental Psychology, 41, 329-335.

Rogers, H. J., Dorans, N. & Schmitt, A. (1986). Assessing unexpected differential item performance of Black candidates on SAT form 3GSA0B and TSWE form E43. Report Report SR-86-22. Educational Testing Service, Princeton, N.J.

Rogers, H. J. & Kulick, E. (1987). An investigation of unexpected differences in item performance between Blacks and Whites taking the SAT. In A. Schmitt & N. Dorans (Eds.), Differential item functioning on the Scholastic Aptitude Test. ETS Research Memorandum, RM-87-1. Educational Testing Service, Princeton, N.J.

Schmitt, A. & Bleistein, C. (1987). Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test Analogy items. ETS Research Report RR-87-23. Princeton, N.J.: Educational Testing Service.

Schmitt, A., Bleistein, C. & Scheuneman, J. D. (1987). Determinants of differential item functioning for Black examinees on Schlastic Aptitude Test analogy items. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Washington, D. C., April, 1987.

Schmitt, A. & Dorans, N. (Eds.). (1987). Differential item functioning on the Scholastic Aptitude Test. ETS Research Memorandum, RM-87-1. Princeton, New Jersey.

Shepard, L., Camilli, G. & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Smith III, H. R. (1986). A summary of data collected from the Graduate Record Examination test takers during 1984-1985. GRE Data Summary Report No. 10, April, 1986. Educational Testing Service, Princeton, N.J.

Sternberg, R. J. (1977). *Intelligence, information processing and analogical reasoning: the componential analysis of human abilities*. Hillsdale, N.J.: Erlbaum & Assoc.

Thissen, D. & Wainer, H. (1985). *Studying item bias with item response theory*. Paper presented at the Fourth European Meeting of the Psychometric Society and the Classification Societies, Cambridge, England.

Whitely, S. (1977). Relationships in analogy items: a semantic component of a psychometric task. *Educational and psychological measurement*, *37*, 725-739.

Wright, D. (1986). *An empirical comparison of the Mantel-Haenszel with standardized methods of detecting differential item performance*. Paper presented at the American Educational Research Association meetings, San Francisco, Ca.

Footnotes

[1] The item analyses conducted regularly by ETS use the DIF statistic (as used in this study) as well as the Mantel-Haenszel statistic. Other work has shown (Wright, 1986) that these two statistics (the Mantel-Haenszel and the DIF statistic) are very highly intercorrelated at .99. Hence the practical import of our findings for DIF should apply as well to any calculations which use the Mantel-Haenszel procedure.

[2] In the version of the formula for computing DIF values used here, percent correct was calculated by taking the number of subjects who got a particular item correct and dividing by the sum of those who got it correct plus the number that either omitted the item or got it wrong. This was called the R + W + O formula (R- rights; W- wrongs; O - omits). These calculations were done separately for the two groups being compared.

While earlier work by Schmitt and Bleistein (1987) indicated that SAT analogy items which ended a test section were subject to "speededness" effects (where Blacks in particular were more likely than matched Whites to not complete all the items near the end of the section), there is no corresponding "speededness" effect for GRE analogy items because these items never end a test section. Nevertheless, we used a DIF formula that excludes NR (not reached) from the calculation of GRE DIF because we are interested later in comparing GRE DIF and SAT DIF for analogies.

[3] The selection of coding categories comes from variables which proved useful in earlier psycholinguistic studies (e.g., word frequency, concrete/abstract, semantic relationships, parts of speech, etc. [see Bejar, Embretson, Peirce & Wild, 1984; Chaffin & Herrmann, 1984; Dawis, Sioriano, Siojo & Haynes, 1974, Hall & Freedle, 1975; Miller, 1951; Sternberg, 1977; Whitely, 1977] . These categories were developed without reference to categories used by GRE or SAT Test Development staffs to code the content of their verbal item types.

[4] The correlation between our re-ranking of item position--referred to in this paper as "actual rank difficulty" which is the main measure which we use as an index of item difficulty in many of our regression analyses--and a standard ETS measure of item difficulty (called P+) is -.953 (based on a sample of 54 GRE analogy items taken from three test forms). The negative correlation is due to the fact that low ranks are associated with larger percents correct.

ERIC
Full Text Provided by ERIC

Another question that might be raised concerning our re-ranking of items (to get at a more accurate measure of item difficulty especially when we seek to combine items across different test forms in our large regression analysis of 234 GRE analogy items) is to what extent the results depend upon the fact that the re-ranking reflects how the total White sample (the base group) did on the test. That is, isn't it possible that had we conducted our ranking using just the percent of Blacks who passed, that this would have dramatically affected the conclusions that we reached? To deal with this possible criticism we calculated the correlation of percent Whites (base group) who passed each item with the percent Blacks who passed each item. The correlation is .9459 (N=54 analogy items; only three GRE test forms were available for this type of analysis based on exact percentage passing). The correlation of the matched White group (who were most similar to the Black sample in conducting the DIF analysis) with the percent Whites (base group) who passed was .9598 (N=54). Obviously the two latter correlations are very similar in magnitude. Hence the fact that we based on re-ranking of item difficulty on the larger base group (consisting of all White examinees) rather than the smaller Black sample should have no major impact on the general results reported here.

[5] A graph of rank difficulty against mean DIF actually suggests that a linear component is very strongly present relating rank with DIF value. However, one can also see that a somewhat better fit would involve adding possibly a second-and/or a third-order nonlinear component to our predictors. In this paper we deal only with the strong linear relationship.

[6] It is important to understand why high frequency words might cause the Black examinee population problems in interpreting the exact nuance of meaning for analogies. Hall, Nagy and Linn (1984) report an analysis of the growth of word use for Black and White children of two socioeconomic levels (middle-class and working class). In one analysis they found that there was a special class of words dealing with objects more commonly encountered in rural than urban areas (e.g., bee, butterfly, feed, grass, land and animal). The middle-class language sample commonly used these high-frequency words, whereas the working-class children used these less often. Furthermore words such as *bee*, *butterfly*, *grass*, *animal* ... often form the basis for *easy* analogy stem construction, especially for easy analogy items having a *science* content: these words are all high frequency words (i.e., they occur often in broadly sampled text counts; Francis & Kucera, 1982). Such words often have more multiple meanings than low frequency words. For example, "grass" and "animal" have more dictionary senses than many rare words. Delay in exposure to words of this type can conceivably have a long-term effect on behavior.

While one initially would think that the vocabulary for GRE easy analogy items would not involve very many of these high frequency words, a brief examination of three GRE forms yielded seven high frequency "rural" words (such as "horse" "snake", etc.); these words had a mean of 5.3 dictionary entries per word.

Clearly, although they do not occur very often, the GRE test does make use of "rural" type words, and these high frequency words typically have about 5 dictionary senses; the intended sense is more difficult (relatively speaking) to extract when little context is provided. Other high frequency non-rural words from the same three forms yielded eight words (e.g., "water" "game" "legs", etc.) which had a mean of 5.2 dictionary senses. Thus "rural" high frequency words seem to be similar in their relative number of dictionary senses to non-rural high frequency words. In general, these high frequency words for the GRE are quite similar to the words we have encountered in analyzing the SAT analogies.

In contrast, the mean number of dictionary entries for a sample of 10 GRE low frequency words (taken from same three GRE forms) was 2.0. This included such words as "vehemence" "anathema" "taxonomist" and "neophyte"). Clearly, the number of different dictionary senses is greatly reduced for low frequency words.

Appendix A[a]

The Means and Standard Deviations of the Verbal GRE Score for the Focal Group, Base Group and the Total Sample for each of the 13 GRE-V Test Forms

	Focal Group:			Base Group:					
	Black Examinees			White Examinees			Total Sample		
Test	Mean			Mean			Mean		
Form	GRE-V	S.D.	N	GRE-V	S.D.	N	GRE-V	S.D.	N
82-1	372.4	101.6	2680	517.5	109.6	32148	506.3	115.7	34828
82-2	372.7	101.1	1547	506.7	103.0	22154	497.9	108.1	23701
83-1	380.8	104.6	2257	521.2	109.6	28831	511.0	115.1	31088
83-2	367.2	097.4	1390	510.7	105.3	19188	501.0	110.8	20578
83-3	360.8	100.2	1117	502.7	111.2	15616	493.2	116.0	16733
84-1	388.8	106.8	2146	525.1	112.2	28942	515.7	117.0	31088
84-2	367.8	098.3	1327	510.5	106.3	19762	501.5	111.4	21089
84-3	367.8	093.6	1108	504.1	105.4	15036	494.7	110.2	16144
85-1	383.1	101.7	2320	521.0	109.0	33215	512.0	113.7	35535
85-2	356.4	093.9	1111	496.7	104.9	14166	486.5	110.3	15277
85-3	370.1	086.3	1402	497.4	108.8	19185	488.7	112.1	20587
86-1	394.6	105.6	2357	525.3	105.2	34694	516.9	109.9	37051
86-2	389.8	098.3	1184	510.0	101.8	20948	503.6	105.2	22132

[a] In calculating DIF (Differential Item Functioning), the distribution of GRE-Verbal scores for the White examinees is weighted at each score level by the frequency of Black examinees who obtained a particular score; hence the mean of this adjusted (weighted) distribution of White examinees will by this procedure have the same mean as the Focal Group. What is reported above is the mean and standard deviation of the unadjusted (unweighted) distribution of the Base (White) group. The total sample results represent the sum of the Focal and Base group examinees.

The results for forms 85-1, 85-2 and 85-3 presented in this table can be compared with the total population of examinees who took the GRE-V test in the 1984-1985 testing year. In a data summary report (Smith, 1986) for 5 test forms, 193,000 examinees had a mean GRE-verbal score of 486 and a standard deviation of 121. This compares favorably with the total sample reported for 85-1, 85-2 and 85-3 which combined have a weighted mean of 499.8

Appendix B

List of Variable Labels and Table of Intercorrelations of 26 Variables Using a Sample of 234 GRE Analogies

<u>Variable Labels</u>

v1 DIF value

v2 Item position (original rank within each set of nine analogies, two sets per test form; these original ranks were based on pretesting results)

v60 Actual rank difficulty (items ranked based on percent passing item from the base White group of examinees; items ranked within each set of nine analogies)

v3 Abstract/conrete: first stem word

v4 Abstract/concrete: second stem word

v5 Adjective: first stem word

v6 Noun: first stem word

v7 Verb: first stem word

v8 Adjective: second stem word

v9 Noun: second stem word

v10 Verb: second stem word

v11 Frequency of the first stem word

v12 Frequency of the second stem word

v13 Science content

v14 Social/personality content

v15 Part/whole relationship

v60 (see v60 listed above)

v61 v11+1 (done in order to calculate logarithms)

v62 v12+1 (done in order to claculate logarithms)

v63 Less frequent entry of v61 and v62

v64 More frequent entry of v61 and v62

v65 Log of v61

v66 Log of v62

v67 Log of v63

v68 Log of v64

v69 "Attribute" (attribute-nonmodifying-syntax; see text)

v70 Functional relationship in stem

The reader should note that the table of intercorrelations is significant only to two decimal places; we have indicated a third decimal place so that a rounding off operation could be carried out by the interested reader.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
V1	1.000	0.389	-0.318	-0.279	0.236	-0.214	0.031	0.081	-0.116	0.076	-0.130	-0.139
V2	0.389	1.000	-0.299	-0.280	0.145	-0.187	0.088	0.223	-0.164	0.030	-0.107	-0.150
V3	-0.318	-0.299	1.000	0.564	-0.329	0.541	-0.341	-0.195	0.230	-0.125	0.019	-0.025
V4	-0.279	-0.280	0.564	1.000	-0.345	0.384	-0.132	-0.224	0.404	-0.305	-0.008	0.092
V5	0.236	0.145	-0.329	-0.345	1.000	-0.611	-0.231	0.277	-0.289	0.136	-0.075	-0.053
V6	-0.214	-0.187	0.541	0.384	-0.611	1.000	-0.628	-0.173	0.453	-0.397	0.086	-0.035
V7	0.031	0.088	-0.341	-0.132	-0.231	-0.628	1.000	-0.059	-0.272	0.354	-0.031	0.096
V8	0.081	0.223	-0.195	-0.224	0.277	-0.173	-0.059	1.000	-0.502	-0.131	-0.040	-0.122
V9	-0.116	-0.164	0.230	0.404	-0.289	0.453	-0.272	-0.502	1.000	-0.790	0.014	0.033
V10	0.076	0.030	-0.125	-0.305	0.136	-0.397	0.354	-0.131	-0.790	1.000	0.012	0.048
V11	-0.130	-0.107	0.019	-0.008	-0.075	0.086	-0.031	-0.040	0.014	0.012	1.000	0.165
V12	-0.139	-0.150	-0.025	0.092	-0.053	-0.035	0.096	-0.122	0.033	0.048	0.165	1.000
V13	-0.173	-0.139	0.152	0.184	-0.149	0.198	-0.097	-0.040	0.172	-0.169	0.016	-0.053
V14	0.239	0.257	-0.300	-0.270	0.119	-0.219	0.151	0.139	-0.102	0.018	-0.090	-0.197
V15	-0.271	-0.168	0.295	0.301	-0.124	0.203	-0.127	-0.075	0.150	-0.118	0.089	0.035
V60	0.486	0.805	-0.309	-0.269	0.226	-0.208	0.033	0.273	-0.153	-0.017	-0.065	-0.163
V61	-0.130	-0.107	0.019	-0.008	-0.075	0.086	-0.031	-0.040	0.014	0.012	1.000	0.165
V62	-0.139	-0.150	-0.025	0.092	-0.053	-0.035	0.096	-0.122	0.033	0.048	0.165	1.000
V63	-0.228	-0.166	0.054	0.036	-0.101	0.047	0.041	-0.063	-0.018	0.066	0.692	0.328
V64	-0.148	-0.152	-0.013	0.048	-0.074	0.034	0.030	-0.101	0.038	0.027	0.765	0.742
V65	-0.290	-0.304	0.182	0.165	-0.193	0.140	0.016	-0.073	0.032	0.014	0.526	0.129
V66	-0.142	-0.224	0.099	0.074	-0.067	0.048	0.007	-0.245	0.051	0.114	0.120	0.748
V67	-0.274	-0.363	0.209	0.150	-0.199	0.142	0.020	-0.140	0.044	0.048	0.441	0.292
V68	-0.216	-0.253	0.119	0.120	-0.102	0.075	0.008	-0.229	0.052	0.102	0.284	0.712
V69	0.216	0.094	-0.112	0.074	0.019	0.063	-0.097	0.109	0.011	-0.090	-0.033	-0.050
V70	-0.205	-0.235	0.394	0.337	-0.203	0.356	-0.238	-0.140	0.255	-0.192	-0.012	0.087

	V13	V14	V15	V60	V61	V62	V63	V64	V65	V66	V67	V68
V1	-0.173	0.239	-0.271	0.486	-0.130	-0.139	-0.228	-0.148	-0.290	-0.142	-0.274	-0.216
V2	-0.139	0.257	-0.168	0.805	-0.107	-0.150	-0.166	-0.152	-0.304	-0.224	-0.363	-0.253
V3	0.152	-0.300	0.295	-0.309	0.019	-0.025	0.054	-0.013	0.182	0.099	0.209	0.119
V4	0.184	-0.270	0.301	-0.269	-0.008	0.092	0.036	0.048	0.165	0.074	0.150	0.120
V5	-0.149	0.119	-0.124	0.226	-0.075	-0.053	-0.101	-0.074	-0.193	-0.067	-0.199	-0.102
V6	0.198	-0.219	0.203	-0.208	0.086	-0.035	0.047	0.034	0.140	0.048	0.142	0.075
V7	-0.097	0.151	-0.127	0.033	-0.031	0.096	0.041	0.030	0.016	0.007	0.020	0.008
V8	-0.040	0.139	-0.075	0.273	-0.040	-0.122	-0.063	-0.101	-0.073	-0.245	-0.140	-0.229
V9	0.172	-0.102	0.150	-0.153	0.014	0.033	-0.018	0.038	0.032	0.051	0.044	0.052
V10	-0.169	0.018	-0.118	-0.017	0.012	0.048	0.066	0.027	0.014	0.114	0.048	0.102
V11	0.016	-0.090	0.089	-0.065	1.000	0.165	0.692	0.765	0.526	0.120	0.441	0.284
V12	-0.053	-0.197	0.035	-0.163	0.165	1.000	0.328	0.742	0.129	0.748	0.292	0.712

	V13	V14	V15	V60	V61	V62	V63	V64	V65	V66	V67	V68
V13	1.000	-0.275	0.121	-0.184	0.016	-0.053	0.032	-0.030	0.134	0.021	0.165	0.023
V14	-0.275	1.000	-0.164	0.301	-0.090	-0.197	-0.128	-0.179	-0.167	-0.343	-0.256	-0.336
V15	0.121	-0.164	1.000	-0.195	0.089	0.035	0.144	0.063	0.176	0.090	0.148	0.149
V60	-0.184	0.301	-0.195	1.000	-0.065	-0.163	-0.138	-0.133	-0.297	-0.250	-0.343	-0.268
V61	0.016	-0.090	0.089	-0.065	1.000	0.165	0.692	0.765	0.526	0.120	0.441	0.284
V62	-0.053	-0.197	0.035	-0.163	0.165	1.000	0.328	0.742	0.129	0.748	0.292	0.712
V63	0.032	-0.128	0.144	-0.138	0.692	0.328	1.000	0.559	0.560	0.267	0.666	0.308
V64	-0.030	-0.179	0.063	-0.133	0.765	0.742	0.559	1.000	0.387	0.547	0.406	0.646
V65	0.134	-0.167	0.176	-0.297	0.526	0.129	0.560	0.387	1.000	0.094	0.820	0.407
V66	0.021	-0.343	0.090	-0.250	0.120	0.748	0.267	0.547	0.094	1.000	0.393	0.878
V67	0.165	-0.256	0.148	-0.343	0.441	0.292	0.666	0.406	0.820	0.393	1.000	0.436
V68	0.023	-0.336	0.149	-0.288	0.284	0.712	0.308	0.646	0.407	0.878	0.436	1.000
V69	-0.088	0.072	-0.052	0.111	-0.033	-0.050	-0.046	-0.051	-0.059	-0.146	-0.107	-0.133
V70	-0.009	-0.137	-0.127	-0.205	-0.012	0.087	-0.027	0.056	0.037	0.134	0.051	0.141

	V69	V70
V1	0.216	-0.205
V2	0.094	-0.235
V3	-0.112	0.394
V4	0.074	0.337
V5	0.019	-0.203
V6	0.063	0.356
V7	-0.097	-0.238
V8	0.109	-0.140
V9	0.011	0.255
V10	-0.090	-0.192
V11	-0.033	-0.012
V12	-0.050	0.087
V13	-0.088	-0.009
V14	0.072	-0.137
V15	-0.052	-0.127
V60	0.111	-0.205
V61	-0.033	-0.012
V62	-0.050	0.087
V63	-0.046	-0.027
V64	-0.051	0.056
V65	-0.059	0.037
V66	-0.146	0.134
V67	-0.107	0.051
V68	-0.133	0.141
V69	1.000	-0.097
V70	-0.097	1.000

297980