GRE®
RESEARCH

# Development of a SIBTEST Bundle Methodology for Improving Test Equity With Applications for GRE Test Development

William Stout
Dan Bolt
Amy Goodwin Froelich
Brian Habing
Sarah Hartz
Louis Roussos

**February 2003**

GRE Board Professional Report No. 98-15P

ETS Research Report 03-06

ETS® Educational Testing Service
Princeton, NJ 08541

# Development of a SIBTEST Bundle Methodology for Improving Test Equity, With Applications for GRE Test Development

William Stout (Principal Investigator)
Educational Testing Service, Princeton, New Jersey
University of Illinois at Urbana-Champaign

Dan Bolt
University of Wisconsin, Madison, Wisconsin

Amy Goodwin Froelich
Iowa State University, Ames, Iowa

Brian Habing
University of South Carolina, Columbia, South Carolina

Sarah Hartz and Louis Roussos
University of Illinois at Urbana-Champaign

GRE Board Report No. 98-15P

February 2003

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

Educational Testing Service, Princeton, NJ 08541

**********

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

**********

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.



Educational Testing Service
Princeton, NJ 08541

**Abstract**

The purpose of this study was to develop a differential-item-functioning (DIF)/impact methodology capable of accurately isolating underlying, conceptually-based causes of DIF (differing item performances across examinee groups that are matched on ability) and impact (differing item performances across examinee groups that are not matched on ability) using data from the quantitative section of two administrations of the Graduate Record Examinations (GRE®) General Test. Analyses indicate that the "SIBTEST bundle methodology" that was developed in the study for GRE quantitative forms is effective and, as such, is exportable to various Educational Testing Service (ETS®) settings. This methodology should help improve test equity of future GRE quantitative forms, as well as test development and standardized tests in general. The developed methodology elevates statistically based DIF analyses from the mere screening of already manufactured items to the modification of test specifications and subsequent test construction processes.

Key words: Differential item functioning, DIF, differential bundle functioning, DBF, impact, test equity, test development, GRE General Test, SIBTEST, SIBTEST bundle method, multidimensionality-based DIF paradigm

**Acknowledgements**

The research reported here was collaborative in every respect and, except for acknowledging that William Stout was the principal investigator, the listing of authors on the cover page is alphabetical.

Dan Bolt and Mi-Ok Kim, who was a research assistant at the time the study was conducted, coded items from two administrations of the quantitative portion of the Graduate Record Examinations (GRE®) General Test using 15 operational categories developed for this study. Throughout this process, Sarah Hartz served as our GRE-quantitative category expert. No initial special expertise is claimed for these three individuals in regard to these two roles they carried out.

Ann Gallagher, Judith Levin, and Mary Morley — all Educational Testing Service (ETS®) research staff — provided categorizations of potential DBF-producing characteristics of items for use in the creation of categories for the present research.

## Table of Contents

## List of Tables

Page



## List of Figures

Page

## Introduction

A central goal of this study was to develop a methodology capable of improving test equity in general and then, in particular, to apply this methodology to the quantitative portion of the Graduate Record Examinations (GRE®) General Test and the proposed GRE Mathematical Reasoning Test. Indeed, an essential aim of this study was to provide useful information concerning the equity of these tests. This goal and the intended application were accomplished by further developing the multidimensionality-based differential item functioning (DIF) paradigm of Roussos and Stout (1996).

The core of the new methodology described here is to carefully bundle test items into conceptually meaningful and statistically dimensionally-distinct categories, some of which may then statistically be shown to be associated with gender- or ethnicity-based differences in test performance. The specific results of the application of this methodology, which are presented in this report, can be used to (a) evaluate current GRE quantitative test forms and (b) improve test specifications and item development of future GRE quantitative, Mathematical Reasoning, and other mathematics-oriented tests — with a goal of improving test equity. Moreover, our new DIF/impact methodology, applied in this report to GRE quantitative data, is completely general in its applicability to test equity. As such, this methodology, referred to as the "SIBTEST bundle method" throughout, can be applied to test equity evaluation and future test development in content domains other than mathematics (see Shealy & Stout, 1993, for a description and evaluation of the statistical DIF procedure SIBTEST).

The particular focus of the analyses presented in this report was the evaluation of differences in performance by gender- and ethnicity-based groups on the GRE quantitative test. Analyses between male and female examinees and between Black and White examinees were carried out using test items and examinee response data from two recent administrations of the GRE General Test (December 1995 and October 1996). Group differences in examinee performance were statistically evaluated by examining both impact and DIF for carefully defined, category-based bundles of items.

For the purposes of this study, "impact" was defined as the difference in performance across groups of examinees (based on ethnicity, gender, or other defining characteristics) on a test item or on a category-based bundle of test items. (A "category" is the conceptually-based organizing principle that was used to form a *dimensionally homogeneous* bundle of items — that

is, a bundle of items that all require broadly similar cognitive processing and, as such, are conceptually and dimensionally homogeneous.) For example, if 60% of White examinees but only 40% of Black examinees were to correctly answer a particular test item, this item would be judged as having "impact" against Black examinees. Impact, of course, combines construct-relevant and construct-irrelevant sources of examinee variation in providing the total difference in performance across groups.

"DIF," on the other hand, was defined as the difference in performance across groups of examinees on an item or category-based bundle of items *when examinees have been statistically matched on a particular cognitive or content construct that is central to the purpose of the test*. For example, if among a group of examinees statistically judged to have the same mathematics ability, 55% of White examinees but only 45% of Black examinees were to correctly answer a particular mathematics test item, this item would be judged to exhibit DIF against Blacks.

In summary, impact is the total difference in performance across groups of examinees, while DIF is the difference in performance across groups of examinees who have been statistically matched according to some relevant construct. Indeed, impact splits into a DIF portion and a non-DIF portion. It is simplistic, and in fact incorrect, to conclude that all instances of DIF constitute evidence of test inequity. In a DIF analysis, even though examinees are matched on a primary ability that is central to the test's purpose, examinees can still display group differences in performance due to the influence of other construct-relevant secondary abilities that are components of the test's overall measurement purpose.

To illustrate, consider a hypothetical test of high-school mathematics achievement consisting of 50% pure algebra and 50% pure geometry items. In addition, suppose that, on average, females perform slightly better on algebra items and males perform slightly better on geometry items. Then, by matching examinees on total test score, which measures an overall mathematics composite of algebra and geometry, a DIF analysis would simply result in many algebra items displaying DIF against males and many geometry items displaying DIF against females. This result could provide highly valuable and useful cognitive or test construction information, but would not constitute test inequity because no construct-irrelevant sources of examinee variation are present.

Analogously, the non-DIF portion of impact can contribute to test inequity in a quite subtle way. No matter how precisely one defines the target construct of a particular test, test

specifications designed to conform to this construct can vary widely. To illustrate, a high-school-level mathematics achievement test consisting of 40% geometry items and 60% algebra items, as dictated by test specifications, could display less impact against females than an apparently equally construct-valid test intentionally consisting of 50% geometry and 50% algebra items. The idea — a relatively new and pragmatic one — is that when multiple test specifications are possible and are all valid, test inequity occurs when a set of specifications is selected that produces larger impact than would other valid sets of specifications (Willingham & Cole, 1997; AERA, APA, NCME, & JCSEPT, 1999). The design and construction of the Second International Mathematics and Science Study (SIMSS) and Third International Mathematics and Science Study (TIMSS) assessments are important practical illustrations of the great variations in opinion that can exist among education experts concerning what items and what proportions of different content are appropriate for a test that is intended to measure achievement in a particular broad content area.

The main focus of the analysis of GRE quantitative data reported here was the amount of DIF and impact that resulted from the conceptual structure of categories developed for that test. A prerequisite for a successful category-level DIF and impact analysis is the careful construction of item categories. In particular, each category must be statistically judged to be (a) dimensionally homogeneous (that is, all items in the category depend similarly on the latent abilities controlling examinee performance), (b) conceptually meaningful, and (c) capable of producing high coding reliability (that is, independent coders of items into categories show a high rate of agreement). Moreover, different categories must be reasonably dimensionally distinct. Many, but not all, of our final set of 15 conceptually based categories were defined in terms of mathematical, or at least quantitative, constructs. This was expected due to the quantitative/mathematical measurement purpose of the GRE quantitative test.

## *Overview of SIBTEST Bundle Methodology*

### *Category Development and Coding*

There are three distinct stages of category development and coding in an actual SIBTEST bundle analysis. First, categories are developed that meet the criteria stated in the preceding paragraph. Second, in order to achieve high intercoder reliability, item coders are trained to correctly categorize items. Third, the actual coding of items into categories is carried out by multiple coders using whatever protocol has been developed. An overview of these stages is

provided below. (The step-by-step process of defining and selecting *preliminary* and *final operational* categories is explained more fully later in this report.)

*Category development.* In general, potential categories should be gleaned from multiple sources to produce a large preliminary set of categories. In any application of our SIBTEST bundle methodology, each application will likely draw from different sources of preliminary categories. However, in all applications, such sources should include:

1. categories substantively viewed as having the potential to produce DIF or impact
2. categories determined by test specifications and test structure
3. categories obtained using statistical techniques designed to discover dimensionally homogeneous bundles

In advising that categories be included from the third source, above, we rely on the obvious link between statistical dimensional homogeneity and substantive homogeneity of bundles. Namely, a statistically dimensionally homogeneous bundle will tend to be substantively homogeneous as well, and vice versa. Thus, the third source results in statistically bundled items that can be judged to possess similar latent multidimensional ability structures from the multidimensional item response theory (IRT) modeling perspective.

In the construction of GRE quantitative categories for the present study, we culled *preliminary categories* from three different sources:

- Educational Testing Service (ETS®) research staff proposed categories that they judged to have the potential to produce DIF or impact.
- We refined broad, GRE-quantitative, specification-based content categories that were also supplied by ETS.
- We derived categories using statistical tools that organize items into dimensionally homogeneous bundles and that rely on dimensionality analysis techniques that are specifically designed to be sensitive to the varying cognitive components influencing examinee performance on items.

Using these three different category sources, we selected 26 preliminary GRE quantitative categories for possible inclusion in the operational set of GRE quantitative categories for our DIF and impact analyses. In order to produce operational categories from the 26 preliminary categories, some categories were accepted in their original form, others were

dropped, and the definitions of still others were altered. This evolution was intended to make the operational categories more statistically dimensionally distinct from one another and/or to make the items within each category more dimensionally similar to one another.

The evolution from preliminary categories to statistically homogeneous operational categories is justified in part because items within each category should be as conceptually homogenous as possible. If items within a category are conceptually *heterogeneous*, the bundle as a whole may show little to no category-level DIF, and yet, homogeneous subbundles may display DIF. Further, even if category-level DIF is statistically detected, a single conceptual cause will either be lacking (due to multiple subbundle-based causes) or statistically hidden due to confounding (with a single cause for the observed bundle-DIF existing but its identification obstructed by the existence of several subcategory-based potential causes).

In addition to satisfying within-category homogeneity, operational categories should be relatively distinct in terms of both statistical dimensionality and substantive characteristics. This is accomplished in part by making them relatively independent or negatively associated. Here, the independence of two categories, A and B, means that the assignment of an item to category A appears probabilistically unrelated to its being assigned to category B. The negative association of two categories means that the assignment of an item to category A makes it appear less likely the item will be assigned to category B. These two probabilistically motivated concepts are made explicit later in the report.

Importantly, it is the combination of two categories being both relatively dimensionally homogeneous and being independent or negatively associated that makes them conceptually distinct from one another. By contrast, in the case of two strongly positively associated categories, a disproportionate number of items will be common to both category-based bundles. In this case, the two categories will be dimensionally similar and hence have rather similar conceptual definitions. It then becomes statistically difficult to accurately assign the relative DIF influence of each of the two associated categories using our SIBTEST analysis-of-variance (ANOVA) method. Moreover, it would be confusing and unnecessarily chaotic to the content or test design expert to have two categories with rather similar definitions. Importantly, coding reliability would also suffer.

*Training of item coders.* In the GRE quantitative project, the second stage of category development and coding — the training of item coders — began with the comprehensive training

of our two item/category coders by our GRE quantitative category expert, using our final 15 operational categories. Training was judged to be complete only when the category expert determined that the coders were each successfully (validly) and reliably (displaying high intercoder reliability) assigning items to categories. In categorizing items, the two coders used our carefully worded final category descriptions, provided later in Table 2, in addition to what they had informally learned about the final categories from the category expert through extensive discussion and performance feedback.

*Coding of items.* In order to correctly assign items to categories, our research staff solved at least two-thirds of the 900 GRE quantitative items that were to be coded (some items could be reliably and validly coded without being solved). Our trained coders achieved intercoder agreement on 80-85% of the item/category assignments. To assign the 15-20% of items for which disagreement occurred, the item coders conferred and reached consensus. Throughout the coding process, the item coders consulted the category expert when appropriate to be certain the categories were being appropriately interpreted.

The details of assigning items to categories vary from application to application, but across all applications, certain requirements must be met:

- At least two item coders must assign every item.
- Coders must have access to a careful and extensive definition of each category so that they can clearly determine how to categorize a given item. (Examples of prototypical items belonging to each category are also very useful for coders during this process; see the Appendix to review prototype items for our final categories.)
- Category experts must carefully train and evaluate the performance of the item coders.
- In order to obtain unique item-to-category assignments for all items, coders must achieve a high degree of intercoder reliability and must follow a protocol for resolving disagreements.

We emphasize that no matter how carefully the definitions of item categories are written, a category expert (or experts) must still carefully train coders to provide principles of correct and reliable coding that are not evident from the written definitions alone. In other words, the training process must go beyond merely training coders to thoroughly understand the written category definitions. Perhaps a useful analogy is that the careful and correct enforcement of American legal statutes requires both a thorough reading of written statutes as well as the study

of precedents that illustrate how the written statutes were applied — with the latter step clarifying aspects of the written statutes that may have been ambiguous.

***Category-Based DIF Estimation***

Once coders assigned items to operational categories, we estimated the amount of DIF that was present in each category-based bundle for each group comparison (Blacks vs. Whites or males vs. females) using the statistical SIBTEST procedure (Shealy & Stout, 1993). The SIBTEST bundle analysis was first conducted separately on GRE quantitative data from the December 1995 administration and then on GRE quantitative data from the October 1996 administration, thus allowing for cross-validation. The two analyses each resulted in a quantification of the amount of category-based differential bundle functioning (DBF) that was present, along with a statement of the statistical hypothesis-testing significance (strength of statistical evidence of category DBF) associated with each category for each administration and group comparison.

DBF is defined to be the expected difference in bundle score across the two groups of examinees, given that examinees have been statistically matched on an appropriate construct (which, ideally, has been judged to be construct-valid). The magnitude of this expected bundle score difference can be easily interpreted on the number-correct test-score scale. We consider DBF a more appropriate label than DIF for our purposes, because the SIBTEST bundle method assesses items collectively in bundles. Thus, when appropriate, we refer to DBF rather than DIF throughout the remainder of this report.

***Preliminary Description of Results***

As is discussed in detail later, the DBF analyses indicated the possibility of a few inconsistent results across the two administrations. As a result, further statistical methods were developed to quantify and test hypotheses in regard to these possible inconsistencies so as to better understand and explain them. In particular, category-bundle heterogeneity would be one likely suspect of the cause of such inconsistencies. Indeed, an important and, in the past, intractable DBF issue has been that items classified as being in one category are often also categorized as being in one or more other categories — a phenomenon we refer to as "overlapping categories." That is, the assignment of items to category-based bundles can, and

often does, result in the assignment of an item to two or more bundles.

The important issue of bundle heterogeneity needs a careful explanation. A category-organizing principle selects items that all share a common conceptual character. Statistically, this can be thought of as a dominant dimension on which all the items of the category "load." However, as indicated, items can and should be assignable to more than one category. More explicitly, items organized into a bundle by a particular category-organizing principle will also be influenced by dimensions associated with other categories to which they have been assigned, which play a secondary role for items in that bundle.

For example, an item put into the "QCD" category — which includes items that cannot be solved from the information given — could also be classified as Applied Geometry With Algebra (an applied geometry problem with an algebra component) and as a Speededness item (an item slower test takers are unlikely to solve due to its placement at the end of the test). That is, in addition to measuring the dominant dimension of the category-organizing principle (dominant because every item of the bundle is influenced by the dimension, which is QCD in this case), various items of a bundle will also often measure various secondary dimensions, each of which is the basis of another bundle-organizing principle (secondary because each such dimension will influence relatively few items of the bundle of interest). Thus, in general, the final categories developed using our method overlap rather than partition the set of items of an administration being studied.

Overlapping categories are a serious concern when carrying out DBF analyses because the secondary dimensions present in a bundle that was formed according to a particular organizing principle can influence the observed DBF for that bundle. In particular, if the secondary dimensions that are present vary widely across two administrations for two bundles resulting from a particular category-organizing principle, the amount of observed DBF for the same bundle-organizing principle could vary widely across the two administrations. As noted, the dominant or organizing dimension influences every item of a bundle, and the other dimensions play a secondary role because each such dimension influences only a limited subset of items in the bundle. Thus, the observed bundle DBF would mainly be the result of the dominant organizing dimension and would likely be statistically consistent across test administrations. Nonetheless, the discovery of DBF and the interpretation of the cause of observed DBF for a particular category is made much more difficult by the presence of

secondary dimensions.

In an effort to understand cross-administration inconsistencies that may arise from overlapping categories, we developed a SIBTEST-based ANOVA approach to disentangle the influence of overlapping secondary dimensional categories on a bundle's displayed DBF. As is related later, this approach enabled us to explain four out of the five inconsistencies across administrations that were found in our SIBTEST bundle analysis. It is interesting to note that, for the one inconsistency that our ANOVA approach did not fully explain, the composition of the (somewhat heterogeneous) category itself changed considerably between the two administrations. This was a new source of bundle heterogeneity for which we could not make statistical adjustments. (Our SIBTEST-based ANOVA methods and results are discussed in detail later in this report.)

In addition to DBF, we completed an analysis of the impact associated with each of our 15 operational categories. For each category and group comparison (males vs. females and Blacks vs. Whites), the amount of impact per item was calculated for both test administrations. As is discussed later, the amount of impact per item was found to be statistically significant for all categories and both group comparisons. This is consistent with the difference in score distributions observed between male and female examinees and White and Black examinees on the GRE quantitative test.

In summary, the main result of our analysis is the set of 15 final item categories for the GRE quantitative test, together with the associated DIF and impact per item estimated for each category and a listing of which categories display significant hypothesis-test-based DBF. Using SIBTEST-based statistical hypothesis testing, a subset of these categories for which there is strong evidence of DBF *across administrations* was also obtained.

We strongly view the methodology developed for this project to have the potential to be usefully applied to standardized-test design and analysis settings other than the GRE quantitative test. At the very least, the results should be useful in the possible future development of a GRE Mathematical Reasoning Test. Further, we believe that our methodology, which grew out of the Roussos and Stout multidimensionality-based DIF paradigm (Roussos & Stout, 1996) and combines statistical and substantive considerations in forming and analyzing categories for DBF, can significantly improve test equity in any standardized test setting and thus, in particular, for any ETS-designed standardized test.

***Data***

The December 1995 administration of the GRE General Test, referred to in this paper as the first administration from which data for the present study was obtained, contained two operational forms of 30 items each and 20 pretest forms of 30 items each. A random sample of six of these pretest forms and the two operational forms (for a total of 240 items) was used to construct the operational categories. For the DBF and impact analyses, another random sample of six pretest forms was added to the original eight forms, producing a total of 420 items to be used in those analyses. Likewise, the October 1996 administration, referred to in this paper as the second administration from which data was obtained, contained two operational forms and 14 pretest forms, each of which had 30 items. All 14 pretest forms and the two operational forms were used in the DBF and impact analyses, for a total of 480 items.

For both administrations, each examinee was administered all 60 operational items, whereas each examinee was in effect randomly assigned to and administered only one pretest form. To be effective, a DBF analysis requires examinees to be matched according to a score that is appropriately construct-valid and administered to all examinees, so that the score matching provides a reliable and valid approximation for the dominant ability (dimension) being measured by the test. In order to obtain such a valid score on each administration, the corresponding 60 operational items were used as the matching subtest. Thus, the DBF and impact results reported here are based upon the analysis of 360 (420 - 60) pretest items from the first administration and 420 (480 - 60) pretest items from the second administration, for a total of 780 (360 + 420) items analyzed.

To complete the DBF and impact analyses, reference and focal groups ("reference" and "focal" being standard DIF terminology) were defined as males and females and Whites and Blacks, respectively. Except for obvious registration-form coding errors, or omissions that made male-female classifications impossible, all examinees were included in the DBF and impact analyses for the males versus females comparison. For the Blacks versus Whites comparison, examinees were split into three categories — White, Black, and other — based on information obtained from GRE registration forms. Only examinees who described themselves as "White, non-Hispanic" or "Black or African-American" were included in the analyses. In the first administration, approximately 650 females, 350 males, 700 Whites, and 100 Blacks were assigned to each pretest form. In the second administration, approximately 750 females, 350

males, 800 Whites, and 60 Blacks were assigned to each pretest form. From the conventional, one-item-at-a-time, DIF analysis perspective, these samples are medium to quite small in size.

Once the final set of operational categories was determined, all pretest items chosen from the first and second administrations were coded as belonging to one or more of the final item categories according to the process described earlier. The median number of pretest items per category was 24 in the first administration and 26 in the second administration. The Algebra category (problems that require any type of algebraic manipulation) and the Probability and Statistics category (problems having either a probability or statistics component and combinatorial problems) each had at least 75 items per category in all four cases. All categories had at least 17 items per category in both administrations, with the exception of the Line Graph category (problems that require the interpretation of a graph representing a trend with a line, segmented lines, or a curve) in the first administration, which included no Line Graph items. The resulting item-by-category incidence matrix for each administration became part of the basic data used in our analyses. Our category expert took special care to prevent drift of category definition between the first and second administrations.

## Method

### *Construction of Preliminary and Final Operational Categories*

#### *Specifications for Operational Categories*

The 15 operational categories were developed by being required to have four characteristics: Each was to be (a) conceptually coherent (based on cognitive content, item format, location on test, and so on), (b) exhaustive (every item must be coded in at least one category), (c) relatively homogeneous with respect to the statistically inferred latent multidimensional structure, and (d) approximately independent or negatively associated. Because dimensional homogeneity holds as well, this last characteristic insures dimensional distinctiveness of categories. We explain these four requirements in more detail below.

*Conceptual coherence.* To be "conceptually coherent," all items in a category must share a common conceptual basis. For example, as noted earlier, the QCD category — the group of quantitative comparison items with the correct answer D, "the relationship cannot be determined by the information given" — is such a conceptual category. Conceptual coherence is essential, because a category found to display DBF or impact must be clearly and easily interpretable to

those involved in writing and modifying text specifications, as well as to those involved in managing item writing and test production. Specifically, if a category is judged to have a deleterious effect on test equity, this category must be clearly understood from the content/cognitive perspective so that test developers can avoid including items belonging to the category in future test specifications, item writing, and test assembly.

*Exhaustiveness.* Because it is desirable to include all items in the analysis, exhaustiveness is usually essential. If the investigator does not require the analysis of all items, then this requirement can of course be dropped. We strongly discourage category-level analyses in which a sizeable proportion of items are not assigned to any category-based bundles. In general, if satisfying this criterion is difficult, a richer, more inclusive set of categories is probably needed.

*Relative homogeneity.* "Relatively dimensionally homogeneous" means that all of the items in a category are judged statistically to best measure approximately the same composite construct — a construct that combines the basic dimensions of the latent structure model. The basic idea of dimensional homogeneity is intuitive from a geometric viewpoint applied to the postulated latent abilities controlling examine test performance. Each item has a mathematically defined and intuitively understood direction of best measurement in the multidimensional, latent IRT model space of examinee abilities. This direction is loosely interpreted as the composite construct the item measures (or measures best). For example, an algebra item measures best in the algebra axis direction and measures much less well in the geometry axis direction. For a rigorous definition of the concept of the direction of best measurement of an item, see Zhang and Stout (1999).

Intuitively, the dimensional homogeneity of a category means that the directions of best measurement of all the items in a category-based bundle should be close, thereby forming a tight "cone" in the multidimensional latent space. For example, if the three dimensions of a hypothetical high-school-level mathematics test are algebra, geometry, and trigonometry, then an algebra item should measure best in a direction close to the algebra axis of the three-dimensional (algebra, geometry, trigonometry) IRT model's multidimensional, latent-ability coordinate system. Thus, an algebra bundle of items should form a fairly tight cone with a conic axis that lies close to the algebra coordinate axis of the three-dimensional coordinate system. By contrast, a trigonometry item with strong algebraic and geometric components would measure best in a composite direction that is roughly equidistant from all three axes. A category-based bundle

consisting of similar composite trigonometry items would form a fairly tight cone with a conic axis that is oriented in a direction that is roughly angularly equidistant to all three axes.

As detailed in a later section (*Creating Dimensionally Homogeneous Final Categories*), the dimensional homogeneity of a category is statistically assessed by appropriately combining item-pair homogeneity indices for all pairs of items in a category-based bundle into a bundle index of homogeneity, denoted by *h*. To obtain these needed item-pair homogeneity indices, appropriately defined item-pair conditional covariances (our basic building blocks for assessing latent multidimensional structure) are used to calculate, in a computationally complex but natural way, a dimensional homogeneity index for every item pair. This homogeneity index is scaled to lie between -1 and 1, where a value close to 1 indicates high item-pair dimensional homogeneity (that is, both items measure best in approximately the same direction of the latent ability space), and a negative value, or even a value close to 0, indicates low item-pair dimensional homogeneity (both items measure best in very different directions of the latent ability space).

Obtaining a very high degree of dimensional homogeneity for a category while also maintaining sufficient conceptual broadness for the category (needed to maintain content relevance) is difficult and often not achievable. Further, the broad nature of such categories precludes a high degree of dimensional homogeneity. However, obtaining a *reasonably* high degree of within-category dimensional homogeneity for sufficiently broad but conceptually coherent categories *is* achievable, and is also essential to successfully completing a statistically effective and useful category-level DBF and impact analysis using the SIBTEST bundle methodology.

As already emphasized, if a category-based bundle displaying DBF or impact has a dimensionally heterogeneous bundle-organizing principle, it is extremely difficult to determine the conceptual cause or causes, because in fact there are various homogeneous subcategories that are each possible sources of the DBF or impact. In particular, a category-organizing principle that is heterogeneous (i.e., allows for multiple subcategories) could easily produce a bundle with a different conceptual character when applied to another administration of the same test, thus leading to an apparent inconsistency in observed category-level DBF across administrations. Indeed, the internal dimensional heterogeneity of one of our operational categories, Probability and Statistics, exhibited this problem.

The bundle-based analysis is only useful when the conceptual organizing principle of the

bundle can be substantively articulated and unambiguously assigned as the cause of the observed bundle DBF or impact. Excessive heterogeneity of the bundle-organizing principle places a burden on the achievement of this goal. Further, an overly heterogeneous category can obscure or even hide important sources of DBF or impact. For example, a heterogeneous bundle may be viewed as having its items' directions of best measurement broadly dispersed geometrically about the bundle's direction of best measurement (the direction of best measurement of a bundle can be viewed as the average of the directions of best measurement of its member items). It is certainly possible that one direction within the bundle's cone of best item-measurement directions could tend to favor the focal group, while another direction could tend to favor the reference group. (Indeed, this was precisely the case for the quite dimensionally heterogeneous Probability and Statistics category.) One unfortunate possibility is that the overall category displays no statistically detectable DBF, while in fact there are strong and detectable DBF-producing subcategories with opposite group effects that could have been found by way of a finer and more dimensionally homogeneous bundling.

The key to the heterogeneity challenge is to define bundles that are relatively dimensionally homogeneous, and yet sufficiently conceptually broad and substantively well defined to be important to test equity efforts. Breadth of definition is important, because the bundles each need to contain a sufficient number of items to provide reasonable statistical power in the face of relatively small examinee group sizes for pretest items. Indeed, one important statistical observation is that decent hypothesis-testing power was obtained in this study in spite of relatively small sample sizes, because large bundles of items were available for DBF analysis. In particular, the unusually small number of Black examinees who took each GRE quantitative pretest item made having large numbers of items per bundle especially important to this project.

*Independence or negative association and distinctiveness of categories.* The necessary approximate independence or negative association of two categories needed to produce category distinctiveness is achieved by requiring an appropriately defined correlation (of item membership) between the two categories to be small or negative, thus allowing a probabilistic interpretation of category distinctiveness. For each pair of categories, this correlation was defined in a natural way: Each item of the set of items used to define our categories belonged to:

- neither category (0, 0)

- both categories (1, 1)
- the first but not the second category (1, 0)
- the second but not the first category (0, 1)

Since 240 items from the first administration were used to define categories, 240 couplets indicated possible membership in each of the two categories for each of the 240 items. Clearly, we could compute the ordinary Pearson-product-moment correlation using these 240 bivariate "observations" as a measure of category pair similarity. For example, consider two categories for which all the items had couplet (0, 0) or (1, 1). This would mean the two categories were identical and had a correlation of 1.

By contrast, having all (0, 1)s and (1, 0)s would produce two totally distinct (no overlap at all) categories and a correlation of -1. Consider two categories to which the 240 items were assigned independently (that is, the fact that an item was assigned to category 1 made it no more likely to have been assigned to category 2). Suppose also the two categories were each dimensionally homogeneous. Then, even though there was some category overlap, the categories would be forced to be quite distinct from one another. Our convention was to judge two categories to be approximately independent or negatively associated if the observed correlation between them was less than 0.3 (either near 0 or possibly negative, and hence producing conceptually distinct categories as required). This bound was chosen so that two acceptable categories from the set of final categories would indeed measure widely dissimilar cognitive constructs, as guaranteed by their having at most low overlap.

The importance of achieving approximate independence or negative association of categories for the SIBTEST bundle methodology was discussed earlier. A large positive correlation leads to two categories having a large overlap and hence measuring best in rather similar directions in the IRT model's multidimensional latent space. When two category-based bundles measure best in similar directions, it becomes statistically difficult to assign the relative bundle-DBF influence of each category (this is analogous to sorting out the influences of variables when high multicollinearity holds in multiple regression).

For test equity analysis and future test development purposes, we certainly want the categories used for these purposes to be widely separated (distinct) from one another in the latent ability space and, as a result, substantively distinct from one another. Thus, it would be

inappropriate to have two conceptually similar categories in the final operational set of categories. Additionally, having conceptually similar categories would make achieving high intercoder reliability very difficult. To illustrate, compare the confusion of interpretation that existed for our sometimes highly correlated preliminary categories (provided later in Table 1) with our conceptually and correlationally distinct operational categories (provided later in Table 2). The latter set of categories satisfied the convention that all correlations between two categories should be less than 0.3.

***Multistep Process for Achieving Operational Category Specifications***

To achieve the four basic category characteristics (conceptual coherence, exhaustiveness, relative dimensional homogeneity, and approximate independence or negative association) we used a multistep process. As the first step, 26 preliminary categories were compiled from the three sources described earlier, as detailed in the next section. Second, the 26 preliminary categories were ranked in order of their decreasing internal dimensional homogeneity, using the *h* index of category dimensional homogeneity (defined in a later section, *Creating Dimensionally Homogeneous Final Categories*). Third, a subset of the 26 categories was constructed by advancing through this ranking one by one and eliminating the category under consideration if it had a correlation greater than or equal to 0.3 with any of the previously selected, more homogeneous categories. In other words, only categories that were both internally relatively dimensionally homogeneous and approximately independent or negatively associated, and hence heterogeneous between categories, were placed in the final list of categories.

As the fourth and final step, the resulting set of categories was examined substantively from the mathematical content/cognitive perspective and then modified slightly to produce greater substantive homogeneity by the occasional addition or deletion of one or more items to a category-based bundle. The final result was a list of relatively dimensionally homogeneous and relatively distinct categories that were each substantively coherent and mutually exhaustive. This process of producing final operational categories is explained in more detail in the sections that follow.

***Development of Preliminary Categories***

The preliminary set of item categories was developed by combining categories obtained

from three different sources, as discussed previously. Cognitive categories from the first source — categories ETS research staff judged as having the potential to produce DIF or impact —had been recently categorized by ETS researchers as potential DBF-producing characteristics of items (Ann Gallagher, Judith Levin, & Mary Morley, personal communication, approximately February 1999). We used some of these cognitive categories — namely QCD, QCC (quantitative comparison items for which the correct answer is C, "the two quantities are equal"), and Word Conversion (word problems in which examinees must go through a nontrivial process of translating words to algebraic formulae or numerical analysis) —fairly directly as categories (although we altered the Word Conversion category slightly in order to increase intercoder reliability). In several cases, ETS categories were combined, split, or modified to create categories that obeyed our four criteria for inclusion. For example, there were clear connections between ETS spatial categories and our geometry categories, and our Calculation Intensive category (problems in which a straightforward solution requires nontrivial numerical calculations or appropriate estimations) absorbed the closely related Short-Cuts/Multiple Solution Paths category provided by ETS.

For the second source — broad GRE quantitative, specification-based content categories — the four categories provided by ETS (Algebra, Geometry, Pure Arithmetic, and Data) were refined to be more distinct and content specific. For example, the Data category (problems that require quantitative interpretation or translation of information, such as the synthesis of numerical data presented in charts, graphs, tables, or paragraphs) was split into several smaller categories — Probability and Statistics and four categories involving means of displaying data (Table, Bar Graph, Line Graph, Pie Graph). The formation of refined categories was accomplished through an interplay between statistical assessment of bundle homogeneity and content considerations. This process allowed final categories to overlap, which did not occur with the original four specification-based categories. For example, after refining, an item could be classified in both an algebra category and a geometry category.

Finally, the third source — the dimensionality-sensitive, item-bundle-producing, hierarchical cluster analysis program based on a conditional-covariance-based concept of proximity, HCA/CCPROX (Roussos, Stout, &Marden, 1998) — was used to form conceptually interpretable categories (bundles of items) based on the 60 operational items from the first administration.

The resulting categories from each source were studied, and preliminary category descriptions were formed based on the item contents of each bundle. Occasionally, an item was added or deleted in order to increase substantive homogeneity. Table 1 displays the set of 26 explicitly defined preliminary categories produced as a result of combining the categories obtained from our three sources. (For comparison purposes, the reader should be aware that final operational categories are displayed later in Table 2.)

Since categories gleaned from each source were developed relatively independently, several pairs of preliminary categories have the same or similar names and/or definitions. However, slight variations between definitions caused the classification of items into these categories to differ at least slightly. Since any high correlation between these categories would be completely removed in the final set of categories by the application of the correlation ≤0.3 criterion, having very similar categories in the preliminary set was not considered at all detrimental to the success of the method.

**Table 1**

***Preliminary Categories***

| Category | Definition |
|---|---|
| QCC[1] | All quantitative comparison items with answer C, "the two quantities are equal." |
| QCD[1] | All quantitative comparison items with answer D, "the relationship cannot be determined from the information given." |
| Word Conversion[1] | Word problems in which the essential formula or equation needed to arrive at the correct solution is not straightforwardly obtained from the wording of the problem. For problems in this category, examinees must go through a nontrivial process of translating words to algebraic formulae, or to numerical interpretation (when algebra is not necessary). |
| Easy Algebra[2] | Problems that require manipulation of algebraic expressions not involving powers, roots, or transcendental mathematical numbers such as π or e. This category includes all problems that require an easy algebra solution, including word problems that require an easy algebra solution. |
| Hard Algebra[2] | Problems that require complicated manipulation of an algebraic expression. Complicated algebra problems are defined as problems that involve nontrivial powers (larger integers than 3 and noninteger powers, including square roots), understanding of π or e, word problems in which the algebraic manipulation is challenging, factoring expressions, and so on. |

[1] cognitive categories; [2] refined general categories; [3] categories from bundle analysis (*Table continues*)

Table 1 (continued)

| Category | Definition |
|---|---|
| All Solutions[2] | In order to arrive at the correct solution for these problems, you *must* include at least one option that is not given directly in the problem. For instance, to correctly solve a problem you may need to recognize that both positive and negative numbers can occur in the problem, or that the numerical manipulation properties of positive numbers differ for numbers less than one versus numbers greater than one. If the list of options that must be included is clearly given in the problem, the item does not fall into this category. |
| Calculation Intensive[2] | Problems in which a straightforward solution would involve nontrivial numerical calculations, usually requiring a shortcut to estimate the calculations (examinees could easily have solved these problems using a calculator if calculators had been allowed on the GRE quantitative exam). This category does not include problems in which straightforward but at most moderately time-consuming calculations suffice. For instance, a problem that only requires the multiplication of two two-digit integers would not be included in this category. Instead, the problems in this category involve calculation (replaceable by appropriate estimations) of large powers, square roots of imperfect squares, π or e, and so on. |
| Data[2] | Problems in which the quantitative interpretation or translation of information is necessary. This includes traditional interpretation of graphs, long applied word problems, and rate problems. The essence of the conceptualization behind the category of items as a whole is that of quantitatively-oriented information filtering and interpretation. All problems in this category require the synthesis of numerical data presented in charts, graphs, tables, or paragraphs. |
| Applied Geometry[2] | Problems that require the interpretation of geometric figures and that are not straightforward in nature. These problems do not merely require the straightforward application of standard geometry formulae like that of the area of a triangle, circumference of a circle, perimeter of a square, and so on. Multiple steps are usually involved to complete them. All problems with geometric figures are not necessarily applied geometry problems — deductive interpretation of the shape is required. Here "applied" refers to the need for deductive reasoning rather than to the presence of a "real-world" setting. |
| Geometry With Memorized Formulae[2] | Geometry problems that require *only* memorized formulae to solve them (for example, straightforward calculations of areas or perimeters, calculation of the length of a hypotenuse of a right triangle, and so on). Therefore, a problem cannot belong to this category and to the above Applied Geometry category. Multi-step geometry problems are *not* considered to be in this category, even if one of the steps involves calculations using standard geometric formulas. |
| Geometry Without Algebra[2] | Geometry problems that do not involve the solution of an algebraic equation or manipulation of an algebraic expression to arrive at the correct answer (they may contain a variable, but they do not require algebraic manipulation). Note that problems in this category may be "applied" or "not applied." |
| Bar Graph[2] | Problems that require the interpretation of a bar graph. |
| Line Graph[2] | Problems that require the interpretation of a graph representing a trend with a line, segmented lines, or a curve. |
| Pie Graph[2] | Problems that require the interpretation of a pie-shaped graph. |
| Table[2] | Problems that require the interpretation of a table containing numerical information. |

[1] cognitive categories; [2] refined general categories; [3] categories from bundle analysis (*Table continues*)

Table 1 (continued)

| Category | Definition |
|---|---|
| Pure Arithmatic[2] | Arithmetic problems with no symbol manipulation (distinct from the Data categories above and below — just involves straightforward computation of numbers). |
| Probability and Statistics[2] | Problems having either a probability or statistics component (involving mean, standard deviation, median, range, flips of coin, drawing from hat, and so on). Combinatorial problems (e.g., using permutations and combinations) are also included in this category, since they require a kind of logic that is similar to the kind of logic required for certain kinds of probability problems. |
| Number Theory[2] | Problems that require examinees to understand general properties of numbers. This includes basic knowledge of the properties of integers, prime numbers, the number line, negative numbers, and numbers between zero and one. For instance, a question may require the knowledge that dividing a positive number by a number less than one gives a value larger than the original number. Or, a problem may give arbitrary numbers lying on specified intervals of the number line, and ask the examinee to order the value of the products or sums of these numbers. Problems from this category are often on the quantitative comparison portion of the GRE quantitative test. |
| Easy Algebra[3] | Problems in which equations or formulas involve only simple algebraic operations (problems also requiring geometric knowledge or reasoning are excluded), such as division, multiplication, addition, or subtraction by a constant (includes problems that require substitution or routine "plug-and-chug"). Problems in this category do *not* require the solution of an equation. However, the problems in this category can also involve simple manipulations of more than one equation (for example, two equations with two unknowns), provided the algebra required is easy. |
| Medium Algebra[3] | Problems that require the manipulation of one or more algebraic expressions with terms in the expressions involving more complicated manipulations than the above Easy Algebra category. This includes squaring, taking the square root, or applying some other simple function; and/or the expression involves a mixture of geometry with simple algebra. Problems can involve solving a simple equation for which the solution process involves at most one or two steps. |
| Hard Algebra[3] | Problems that require the solution of one or more equations with more than two steps needed. These problems usually involve expressions with $x^2$, $\sqrt{x}$, or some other higher-power function of x. The solution may also involve the nonroutine use of negative numbers or geometry. |
| Data[3] | Precisely the problems included in the graphical analysis section of the GRE quantitative exam. This section is usually comprised of problems 21-25 and includes five items that all relate to the same graph or figure. |
| Easy Word Problems[3] | Word problems that require neither complex algebraic manipulations nor complex translations from text to algebraic formulae. |
| Fractions With Numbers[3] | Problems that involve ratios of numbers or simple ratios of variables that require interpretation and not algebraic manipulation. Algebra problems that require many steps to arrive at a correct solution do not fall into this category, regardless of the presence of fractions with numbers in the problem. |

[1] cognitive categories; [2] refined general categories; [3] categories from bundle analysis (*Table continues*)

Table 1 (continued)

| Category | Definition |
|---|---|
| Pure Geometry[3] | Geometry problems that do not require algebraic manipulation of variables. There may be variables in the problem, but the solution does not require any algebraic manipulation of the variables. These problems may or may not involve a geometric shape. |
| Speededness[3] | The last several items of each section have been placed in this category. The number of items on each pretest was determined by looking at the bundle homogeneity index *h* for varying length (numbers of items) on each pretest form. |

[1] cognitive categories; [2] refined general categories; [3] categories from bundle analysis

***Creating Dimensionally Homogeneous Final Categories***

It is vital that each category included in the final set of categories be dimensionally homogeneous. As typically initially conceived by users of our bundle DBF methodology, the category-organizing principles intended to produce the preliminary categories usually result in item bundles that vary greatly in their dimensional homogeneity. In the present study, the four broad, specification-based content categories provided by ETS — Algebra, Geometry, Pure Arithmetic, and Data —were expected to be very internally heterogeneous. Hence, they could not be used as category-organizing principles until they were substantially refined. By contrast, because the software used for the third source of categories was designed to produce dimensionally homogeneous bundles, the resulting categories from that source needed little to no refining.

The dimensional homogeneity of the various category-organizing principles associated with the preliminary categories had to be assessed because these categories were candidates for inclusion in the final set of categories, which were required to be dimensionally homogeneous. In particular, it was essential to identify items that were dimensionally assigned to the wrong category-based bundle. Further, it was possible that a preliminary category-organizing principle produced bundles with so many inappropriately assigned items that the organizing principle itself needed refinement. This was indeed what happened with the four broad, specification-based, GRE quantitative categories supplied by ETS.

We developed a conditional covariance-based methodology for the two essential tasks needed to produce dimensionally homogeneous category-organizing principles that would result in all items being correctly assigned to bundles. The first of these two tasks was to assess the appropriateness of all of the category-based item assignments. In particular, we wished to

remove an item from a category-based bundle when the item was not sufficiently dimensionally homogeneous, relative to other items in the bundle. Closely related, we wanted the capacity to reassign poorly assigned items to different category-based bundles with which they were sufficiently dimensionally homogeneous. Finally, we wanted to refine a category-organizing principle if it caused too many items to be inappropriately assigned to a bundle, thus creating excessive heterogeneity within that bundle. In particular, such a bundle could be split to produce two distinct and more dimensionally homogeneous bundles. In this manner, a possible, preliminary category-organizing principle could be refined into two actual, preliminary category-organizing principles. To carry out these three closely related subtasks of the first task, we used a dimensional homogeneity index defined for all item pairs that was developed by Bolt, Roussos, and Stout (1998). Importantly, an item should never be shifted nor a category-organizing principle refined to achieve dimensional homogeneity unless it makes substantive sense to do so as well.

The second task was to measure the degree of dimensional homogeneity of the items in a bundle for all preliminary, category-based bundles, once the assignment of items to these bundles had been clarified and, in certain cases, the defining principles of the bundles had been refined. This measure was used for the important goal of comparing the relative dimensional homogeneity of the preliminary categories in order to retain those that were sufficiently dimensionally homogeneous for inclusion in the final set of categories. To accomplish this, the index *h* was defined based on the Bolt et al. (1998) item-pair dimensional-homogeneity index referred to above.

The foundation of the Bolt et al. (1998) index of item-pair dimensional homogeneity — as well as of the HCA/CCPROX procedure used as the third source of preliminary categories — is item-pair conditional covariances. To understand how conditional covariances can provide information about the latent multidimensional structure of a test, we again use the geometric viewpoint of the multidimensional latent space introduced earlier (see, Ackerman, 1996, for a summary of geometric modeling of a multidimensional latent space). This geometric conceptualization facilitates understanding of the assessment of the dimensional homogeneity of item pairs.

Recall the concept of the direction of best measurement of an item. Statistically, the idea is to select the direction in the latent space in which the item displays the highest average Fisher

information. Similarly, we can intuitively understand and rigorously define the direction of best measurement in the latent space of the number-correct score on a subset of test items (see Zhang & Stout, 1999), again using the concept of statistical information. A subset score that is often of special importance for the SIBTEST bundle methodology is the set of all operational items of a test. For our purposes, we refer to the direction of best measurement of the total score on the operational items as the unidimensional latent composite best measured by the test. Algebraically, this composite is a linear combination of the latent-space coordinate axes with all axis coefficients positive. Geometrically, this composite is simply a direction in the positive sector of the latent space (think of it as a vector pointing positively away from the origin of the latent-space coordinate system in the direction of the best measurement of the test score). For example, in the two-dimensional latent space of algebra and geometry of a hypothetical high-school-level mathematics test, the best-measured latent composite of a balanced algebra and geometry test would be the 45° line away from the origin, equidistant between the algebra and geometry axes. The intuitive idea is that the test score is best able to discriminate among examinees in this increasing (combined algebra and geometry) 45° direction in the latent-ability space.

Item-pair covariances — conditioned on the latent-ability composite best measured by the test and then averaged over all values of the composite — form the basic building blocks for latent dimensionality analyses of items and, hence, for the dimensional homogeneity tasks described above. A positive conditional covariance (understood through the remainder of the paper to have been averaged over the distribution of the best-measured composite) indicates a tendency toward dimensional homogeneity of an item pair, while a near-zero or negative conditional covariance usually indicates dimensional heterogeneity of an item pair.[1] For example, in a two-dimensional test with 50% algebra and 50% geometry items, two algebra items (thus dimensionally homogeneous) would have a positive conditional covariance, while an algebra/geometry item pair (thus dimensionally heterogeneous) would have a negative conditional covariance.

Note that this claim is intuitively clear when viewed properly. Consider an illustration. For a subpopulation of examinees defined as having 50th percentile (“average”) math ability (a composite of combined algebra and geometry ability), an examinee of the subpopulation answering a hard algebra item correctly constitutes evidence of high algebra ability (> 50th

percentile). This is the case because, although the examinee's math ability is fixed at the 50th percentile, his or her algebra ability is not. And because math ability is fixed at the 50th percentile, there is now inferential evidence that the examinee must have low (< 50th percentile) geometry ability. Thus, the conditional covariance for the algebra/ geometry item pair is negative. Similarly, if both items are algebra items, getting a hard algebra item right again provides evidence of high algebra ability (the inferred low geometry ability is irrelevant), which increases the probability of getting the second algebra item right. Thus the conditional covariance of the two algebra items is positive.

A general multidimensional-based theory of how conditional covariances are used to reveal the latent dimensional test structure is thoroughly presented in Zhang and Stout (1999). In particular, they show that the larger the conditional covariance is, the more dimensionality homogeneous the item pair is. If two items have a large, positive conditional covariance, then the two items are judged to be approximately dimensionally homogeneous and should be assigned to the same category. By contrast, the smaller (even becoming negative) the conditional covariance, the more dimensionally heterogeneous the items are (but refer again to Note 1 at the end of the report for a technical exception to this claim).

For a given test, the set of all item-pair conditional covariances can be put into a matrix for convenience. For example, if there are 60 test items, a 60-by-60 matrix in which the entry in the second row, fifth column denotes the conditional covariance between the second and fifth items would be used. Such matrices of conditional covariances form the basic input for solutions to the two essential dimensional homogeneity tasks described earlier, and in particular, are used to compute the Bolt et al. (1998) item-pair homogeneity index.

In the special case of GRE quantitative data used for the present study, the entire pool of examinees had taken the same set of operational items, and each examinee was, in effect, randomly assigned to one of several pretest item forms. For a given pretest item form of 30 items, a 90-by-60 conditional-covariance matrix could be constructed that gives the estimated conditional covariance for every pretest or operational item paired with one of the 60 operational items. We *could* define item-pair proximities for each pair appearing in one of the 90-by-60 conditional-covariance matrices using each item pair's conditional covariance. But we are unable to extend this to pretest item pairs *when each item is part of a different pretest form*. The advantage of using the Bolt et. al. (1998) homogeneity index instead is that it is definable for all

item pairs.

For the present study, each 90-by-60 conditional-covariance matrix was used to construct the Bolt et. al. (1998) index of the dimensional homogeneity of every item pair when both items belonged to the same conditional-covariance matrix. In particular, we obtained a 60-by-60 item-pair homogeneity matrix for the operational items and a 30-by-30 homogeneity matrix for each pretest item form (i.e., all item pairs were from the same pretest form), and we obtained a 30-by-60 homogeneity matrix for pretest/operational item pairs made up in part of pretest items from each pretest form.

As stated earlier, the Bolt et. al. (1998) item-pair homogeneities are defined not only within each pretest form, combined with the operational item set, but are also defined between arbitrary pairs of items, including pairs consisting of two pretest items from *different* pretest forms. This was done in the present study by combining different 90-by-60 pretest matrices. The interested reader can consult the Bolt et al. paper for details of exactly how the homogeneity index is computed for all item pairs.

The Bolt et al. (1998) index takes values close to one for approximately dimensionally homogeneous items. Thus, taking one minus the Bolt et al. homogeneity index converts it into a "proximity" index between item pairs, as needed. For this proximity index, the greater the proximity-based "distance" between two items, the greater the dissimilarity between the directions of best measurement of the two items — or in other words, the greater the dimensional heterogeneity between the two items.

We return to one aspect of the first task: the removal and possible reassignment of those few items that may have been improperly assigned to various categories. Given our item-pair proximity index, which was now defined for all item pairs of the test, we wished to determine for each particular category-based item bundle which items were close to it and which were not. The intuitive meaning of an item being close to a bundle is that its direction of best measurement is close to that of the specified bundle. Here, the direction of best measurement of a bundle (rigorously defined by Zhang & Stout, 1999) could be intuitively thought of as the average of the directions of best measurement of its member items. The basic input used for assessing the dimensional homogeneity of all items with respect to a specified bundle was a list of all items (both inside and outside the specified bundle) ranked in order of increasing distance from the specified bundle. An item's distance from a bundle was naturally defined as the average of the

item's Bolt et. al. proximities to all the items of the bundle.

Next, the actual distances of the items from the specified bundle were replaced by the closeness ranks (1, 2, ... 240) of the items obtained from these distances. Thus, test items that were relatively dimensionally homogeneous with the specified bundle (and hence had a small distance from the bundle) were ranked relatively close to one, and items that were dimensionally heterogeneous from the specified bundle were ranked far from one and perhaps relatively close to 240. Therefore, an item that was ranked relatively close to one but was not in the specified bundle had to be reexamined to determine if the item belonged in the specified bundle. Similarly, if an item in the specified bundle was ranked in the middle or closer to 240, it could have been improperly included in the bundle — or, even if it was correctly assigned, the item may have been heavily influenced by some other content/cognitive constructs captured by one or more other bundles. In this case, the direction of best measurement of the problematic item was rather different from that of the other items in the studied bundle and, hence, from the direction of best measurement of the bundle. Indeed, if the problematic item had a low rank for one of the other bundles, it may have been necessary to reassign it to this other bundle of items if it was conceptually homogeneous with it as well.

This method of ranking items in order of their decreasing dimensional homogeneity to the specified bundle was also found to be useful in terms of improving dimensional homogeneity of the preliminary category-defining principles. For example, analysis of each item's closeness rankings was used to help refine the four broad, GRE-quantitative, specification-based categories provided by ETS — Algebra, Geometry, Pure Arithmetic, and Data. The resulting, more narrowly defined and dimensionally homogeneous categories were then included in the list of 26 preliminary categories. For example, the Geometry category was split into several smaller categories, each requiring a different type of geometric knowledge. These refined categories then produced bundles with much smaller closeness rankings for their items, as desired, and thus produced dimensionally homogeneous bundles.

As intended, ranking items based on their dimensional homogeneity with their assigned bundle also produced better classification of items into category-based bundles by targeting certain items for reevaluation. The QCD category provides an example. Ironically and interestingly from the viewpoint of providing evidence of the effectiveness of the closeness rankings, these rankings exposed two QCD items that we had incorrectly solved, and thus had

incorrectly categorized. Additionally, the method correctly indicated that an overlooked QCD-like item that was not part of Quantitative Comparisons section of the test should have been included in the QCD category.

In addition to helping form categories that are more dimensionally homogeneous, this ranking index was also used to complete the second task: using the proximity-based ranks of a bundle's items to create a statistic *h* that indicates each bundle's degree of dimensional homogeneity. The creation of such a statistic to assess category homogeneity was necessary because the dimensionality assessment procedure we often use, DIMTEST (Stout, 1987), can only test whether a bundle of items is dimensionally distinct from any other specified bundle of items. DIMTEST cannot assess the dimensional homogeneity of the bundle itself.

Our bundle-homogeneity statistic *h* was calculated by using the above-described proximity-based ranks of all test items from the specific bundle for which we desired the *h* index to be calculated. Then, using these rankings, the sum of the ranks for those items in the bundle was calculated. The statistic *h* compared the observed value of these summed ranks of the items in the bundle with that of a hypothetical bundle of the same size, the item ranks of which were evenly distributed among all ranked items (that is, distances from the specified bundle were evenly spaced). One could intuitively think of randomly placing *n* items in this hypothetical bundle. For a category-based bundle with *n* items, the homogeneity statistic, *h*, is given by:

$$h = \frac{\sum \text{ item ranks of the bundle} - \sum_{i=1}^{n} i}{\sum \text{ evenly spaced } n \text{ ranks} - \sum_{i=1}^{n} i} \qquad (1)$$

Note that the centering term $\sum_{i=1}^{n} i$ represents an ideal, totally dimensionally homogeneous bundle consisting of the *n* closest items to the bundle actually being the bundle members. Hence, $\sum_{i=1}^{n} i$ is an appropriate centering quantity. From this defining equation, it is clear that a smaller *h* indicates a more dimensionally homogeneous bundle and that $h \geq 0$ always holds.

The empirical distribution of *h* under the assumption that the bundle was randomly generated (no unifying principle present among the items of the bundle) was obtained by calculating the *h* statistic for a simulation of 10,000 randomly generated bundles. Using this empirical distribution, we obtained the capacity needed to carry out a hypothesis test for assessing when a bundle is not sufficiently dimensionally homogenous to have resulted from a

coherent bundle-organizing principle. In the next section we will see that *h* is vital to the process of ensuring that the final bundle-organizing categories produce dimensionally homogeneous bundles.

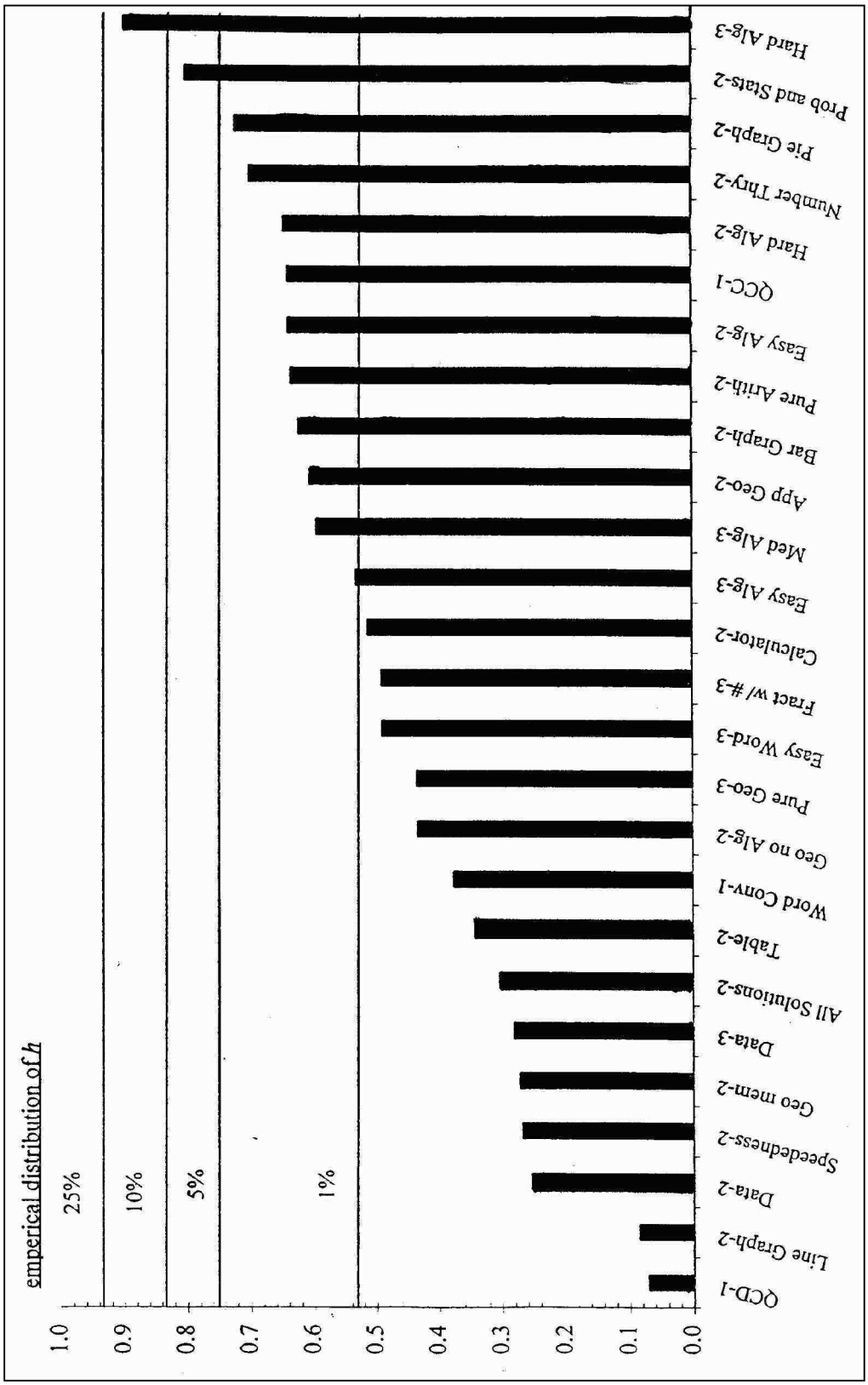
***The Reduction Process: Obtaining the Operational Categories***

The set of 240 items (six pretests plus two operational tests) from the first administration was classified into the 26 preliminary categories. Next, the homogeneity indices of these categories (using *h*) and degree of dimensional distinctiveness among the categories (as measured by the correlations between all pairs of category-based bundles) were evaluated in order to arrive at a relatively dimensionally homogeneous and approximately independent or negatively associated group of final categories. A description of how the preliminary set of 26 categories was then reduced and modified follows.

The preliminary categories were ordered by *h* from most to least dimensionally homogeneous, as shown in Figure 1. Each category was then considered for inclusion in the final set of categories in decreasing order of its internal dimensional homogeneity. At each stage, a category was accepted provided its correlation with other already selected members of the final set of categories was sufficiently low (less than 0.3; recall that a low correlation equals approximate independence and that a negative correlation equals negative association).

For example, QCD — the most dimensionally homogeneous category[2] — was chosen as the first member of the final set of categories. Then, the correlation between QCD and the next most homogeneous category, Line Graph, was evaluated. Since the magnitude of this category-pair correlation was less than 0.3, satisfying our criterion, the highly dimensionally homogeneous Line Graph category was judged to be sufficiently distinct from the first chosen category and was thus also admitted as an operational category. The process was continued in this manner.

Unfortunately, because of the high level of correlation between some of the preliminary categories, several categories were targeted for elimination, causing some items to become unclassified. To alleviate this problem, some categories that otherwise would have been eliminated were redefined to make them sufficiently more distinct. These redefined categories were then included in the final list of operational categories. For instance, Applied Geometry was




Note: 1 = cognitive categories; 2 = refined general categories; 3 = categories from cluster analysis

***Figure 1*. Preliminary categories in order of decreasing dimensional homogeneity or *h*.**

refined into a category more distinct from Geometry Without Algebra by extracting from it a new category, Applied Geometry With Algebra, which retained all Applied Geometry problems having an algebra component. The remaining items from the original Applied Geometry category were then moved to the Geometry Without Algebra category. The result is two new categories (Applied Geometry With Algebra and Geometry Without Algebra) that are less associated, and hence more distinct, than the original two, positively associated categories of Applied Geometry and Geometry Without Algebra.

One reason categories were not required to be nonoverlapping is that most GRE quantitative items integrate several mathematical concepts, and we thus wanted to code many of the items into overlapping categories that reflect the simultaneous influence of several major content/cognitive processes. For instance, a QCD problem is more accurately coded if we can acknowledge that it is also a Geometry Without Algebra problem. This overlapping allows items to be more accurately coded, instead of forcing each one into a single nonoverlapping category, which would be a drastic over-simplification.

During the categorization process, the Probability and Statistics category was extremely problematic. Probability and Statistics is a relatively large category, clearly defined, although not internally dimensionally homogeneous from either the content or the statistical dimensionality analysis perspective. In fact, it was the only category placed in the final set of 15 categories that had a homogeneity index above the 5% hypothesis-testing rejection level (thus not rejected) of the empirical, simulation-produced distribution of *h* under the null hypothesis of a randomly generated category. Thus, the *h* value for Probability and Statistics is consistent with the null hypothesis that items were assigned randomly to the category — random assignment of items to a category being the antithesis of a dimensionally homogeneous category.

Unfortunately, most questions in the Probability and Statistics category cannot be classified in any of the other categories. Additionally, parsing the category into more homogeneous subcategories would result in numerous microcategories from the content perspective. Thus, since classifying all items in at least one category was judged more important than improving homogeneity by dropping or dividing the Probability and Statistics category, this category was included as an operational category in its current and somewhat heterogeneous form. Table 2 displays the final operational categories. (To achieve a better understanding of these categories, see the Appendix for examples of prototypical items from each category —

except Speededness, which depends on test-form placement. The items exhibited in the Appendix are disclosed items from many different administrations and are distinct from items included on the two administrations analyzed in the paper, which were undisclosed items.)

**Table 2**

***Operational Categories***

| Category | Definition |
|---|---|
| Algebra | Problems that require any type of algebraic manipulation. Items in this category are often also classified as Geometry With Memorized Formulae, Applied Geometry With Algebra, Word Conversion, and so on. A problem with a straightforward application of a formula (e.g., plugging in numbers) is not considered Algebra if that is the only requirement of the problem. |
| Calculation Intensive | Problems in which a straightforward solution would involve nontrivial numerical calculations, usually requiring a shortcut to estimate the calculations (examinees could easily have solved these problems using a calculator if calculators had been allowed on the GRE quantitative exam). This category does not include problems in which straightforward but at most moderately time-consuming calculations suffice. For instance, a problem that only requires the multiplication of two two-digit integers would not be included in this category. Instead, the problems in this category involve calculation (replaceable by appropriate estimations) of large powers, square roots of imperfect squares, π or e, and so on. |
| Fractions With Numbers | Problems that involve ratios of numbers or simple ratios of variables that require interpretation and not manipulation. Algebra problems that require many steps to arrive at a correct solution do not fall into this category, regardless of the presence of fractions with numbers in the problem. |
| Bar Graph | Problems that require the interpretation of a bar graph. |
| Line Graph | Problems that require the interpretation of a graph representing a trend with a line, segmented lines, or a curve. |
| Pie Graph | Problems that require the interpretation of a pie-shaped graph. |
| Table | Problems that require the interpretation of a table containing numerical information. |
| Applied Geometry With Algebra | Problems that require the interpretation of geometric figures that are not straight-forward in nature and that require algebra. These problems do not merely require the straightforward application of standard geometry formulas like that of the area of a triangle, circumference of a circle, perimeter of a square, and so on. Multiple steps are usually involved to complete them. (All problems with geometric figures are not necessarily applied geometry problems — deductive interpretation of the shape is required. Here "applied" refers to the need for deductive reasoning rather than to the presence of a "real-world" setting.) |
| Geometry Without Algebra | Geometry problems that do not involve the solution of an algebraic equation or manipulation of an algebraic expression to arrive at the correct answer (they may contain a variable, but they do not require algebraic manipulation). Note that problems here may be "applied" or "not applied." |

(*Table continues*)

Table 2 (continued)

| Category | Definition |
|---|---|
| Geometry With Memorized Formulae | Geometry problems that require *only* memorized formulae to solve (for example, straightforward calculations of areas or perimeters, calculation of the length of a hypotenuse of a right triangle, and so on). Therefore, a problem cannot belong to this category and to the Applied Geometry With Algebra category. Multi-step geometry problems are *not* considered to be in this category, even if one of the steps involves calculations using standard geometric formulas. |
| Number Theory | Problems that require examinees to understand general properties of numbers. This includes basic knowledge of the properties of integers, prime numbers, the number line, negative numbers, and numbers between zero and one. For instance, a question may require the knowledge that dividing a positive number by a number less than one gives a value larger than the original number. Or, a problem may give arbitrary numbers lying on specified intervals of the number line, and ask the examinee to order the value of the products or sums of these numbers. Problems from this category are often on the quantitative comparison portion of the GRE quantitative test. |
| Probability and Statistics | Includes problems having a probability or statistics component (involving mean, standard deviation, median, range, flips of coin, drawing from hat, and so on). Combinatorial problems (e.g., using permutations and combinations) are also included in this category, since they require a kind of logic that is similar to the kind of logic required for some probability problems. |
| QCD | All quantitative comparison problems with answer D, "the relationship cannot be determined from the information given." |
| Speededness | The last four items of each section have been placed in this category. This number of items (four) was determined by looking at the bundle homogeneity index *h* for end-of-section item sequences of varying length on each pretest form. |
| Word Conversion | Word problems in which the essential formula or equation needed to arrive at the correct solution is not straightforwardly obtained from the wording of the problem. For problems in this category, examinees must go through a nontrivial process of translating words to algebraic formulae or to numerical analysis (when algebra is not necessary). |

## Analysis and Results

### *DBF Analysis of Final Operational Categories Using SIBTEST*

As a preliminary step in the analysis of the average amount of DBF produced by different categories of GRE quantitative items, the SIBTEST procedure (Shealy & Stout, 1993) was used to estimate the amount of DIF for each of the 780 pretest items. Indeed, this informational step is always useful as a part of the SIBTEST bundle method for background. For each pretest item, the SIBTEST DIF value, denoted as $\hat{\beta}$, was calculated using the 60 operational items from the corresponding test administration as the matching, or "valid," subset.

From the construct validity viewpoint, total score on the 60 operational items is an observable, empirical approximation of the unobservable, unidimensional, latent-ability

composite best measured by the GRE quantitative test (refer to the earlier discussion of the direction of best measurement of a test). For example, matching on total test score on a high-school-level mathematics test balanced between algebra and geometry amounts to attempting to match examinees on the algebra/geometry composite best measured by the test. Using total test score from the GRE quantitative operational test as an approximation for the latent-ability composite best measured by the exam, the DIF item analysis determined which items perform differentially across groups of examinees. Although the option was not needed in this study, the SIBTEST matching subset can, in general, be any subset of items of a test, not necessarily the operational items of the analyzed test.

For a single item on an arbitrary test, the SIBTEST $\hat{\beta}$ DIF value is the estimated average difference in the probabilities of correctly answering the item for two randomly chosen examinees from each of the two examinee groups (in our case, either Blacks vs. Whites or males vs. females), given that the two groups have been matched on a score that approximates the composite ability best measured by the test. A positive value of the SIBTEST $\hat{\beta}$ statistic indicates DIF against the focal group (in this case, females or Blacks), while a negative value of the SIBTEST $\hat{\beta}$ statistic indicates DIF against the reference group (in this case, males or Whites).[3] Using a numerical example, if $\hat{\beta} = 0.1$, then the estimated average probability (averaging over matching-score values) of answering the item correctly for a randomly chosen reference-group examinee is 0.1 greater than the average probability of answering the item correctly for a randomly chosen focal-group examinee. Since each pretest item is disjoint from the matching subset of operational items used to calculate the $\hat{\beta}$ DIF statistic, the values of the pretest item $\hat{\beta}$ s are not artificially required to sum to zero over all pretest items. By removing this artificial requirement that is present in most other DIF-analysis approaches, this SIBTEST DIF-analysis approach allows us to obtain an *unbiased* estimate of the amount of DIF for each item.

In addition to estimating the amount of DIF associated with each individual item, the SIBTEST procedure can estimate the amount of DBF associated with a set, or bundle, of items. In the GRE quantitative setting, the 15 operational categories were used to form separate bundles of pretest items to be analyzed for DBF. For each pretest form, the SIBTEST method was used to estimate the amount of DBF associated with each of the 15 operational categories using the 60

operational items from the corresponding test administration as the matching subset. Each resulting bundle $\hat{\beta}$ value is the estimated average score difference on the bundle between the two examinee groups, given that the two groups have been matched on GRE quantitative operational test score, which approximates the composite ability best measured by the test. This choice of scaling for $\hat{\beta}$ allowed us to directly assess the influence of the pretest bundle DBF on the total-test-score scale. For example, if a six-item bundle on a 50-item test yields $\hat{\beta} = 2.0$, the estimated DBF influence of the bundle, averaged across the entire latent composite ability range, will be a reduction of 2 in the test score of a randomly chosen focal-group examinee compared to a randomly chosen reference-group examinee. The amount of observed DBF is of practical importance if the estimated focal-group reduction in total test score will have serious societal consequences, as determined by the size of the estimated reduction in test score and the use of the test (e.g., part of the admission process to graduate school).

Each GRE-quantitative category bundle $\hat{\beta}$ is in fact the sum of its individual item $\hat{\beta}$s (this a general property of SIBTEST). As in the case of DIF for each GRE quantitative item, the pretest-item category bundle $\hat{\beta}$s are not artificially required to sum to zero over pretests for a fixed category or over all categories and pretests. Thus, we also have an *unbiased* estimate of the amount of DBF for each category-based pretest bundle relative to the scale best measured by the GRE-quantitative operational test items. For each of the 15 categories, 12 bundle $\hat{\beta}$s were calculated for each of the 12 pretest forms taken from the first administration, and 14 bundle $\hat{\beta}$s were calculated for each of the 14 pretest forms taken from the second administration (since each category was not present in every pretest form, a particular category could have fewer $\hat{\beta}$s).

SIBTEST bundle $\hat{\beta}$s are approximately normal (Shealy & Stout, 1993) with theoretical mean $\beta$ and estimated standard error (SE) denoted by $\hat{SE}(\hat{\beta})$, where the parameter $\beta$ represents the true amount of DBF present. Given the approximate normal distribution for $\hat{\beta}$, the value of a particular bundle $\hat{\beta}$ for a particular pretest form can have a considerable amount of variability depending on the magnitude of its SE. Thus, with probability one, $\hat{\beta}$ does *not* equal the true $\beta$. In fact, it is likely to be off from $\beta$ by as much as $\pm 1\ \hat{SE}(\hat{\beta})$. More precisely, only about two thirds of the time can we expect $\hat{\beta}$ to fall within $\pm 1\ \hat{SE}(\hat{\beta})$ of the true $\beta$, and about one third of the time we can expect $\hat{\beta}$ to fall outside $1\ \hat{SE}(\hat{\beta})$, in either direction.

Therefore, under the null hypothesis of no DBF, a single bundle $\hat{\beta}$ divided by its estimated standard error, $\hat{SE}(\hat{\beta})$, would be distributed approximately normally with mean zero and standard deviation one. Thus, for a fixed category, when the hypothesis of no DBF is true, it is very likely that at least one SIBTEST bundle $\hat{\beta}$ (out of 26 pretest bundles) would be significantly different from zero (using a Z test and a 5% level of significance), falsely indicating a significant amount of DBF present for that category-based pretest bundle. In other words, even if the category of interest did not display DBF, there would likely be at least one pretest form for which the category would demonstrate significant DBF. Thus, inconsistency of DBF results is likely to occur across pretest forms due to natural, random examinee-response variability. This is an important point: Each pretest bundle $\hat{\beta}$ is a random variable, the variability of which is driven by the randomness of examinee responses to test items. This fundamental assumption of randomness — namely, that the latent-ability value does not deterministically control an examinee's response to an item — lies at the heart of IRT modeling.

In order to determine the overall DBF of each category (i.e., across pretests), the first analysis, completed separately for each administration and group comparison, combined the pretest-item category bundle $\hat{\beta}$ values across pretest forms for each category to arrive at a single DBF index. In addition, by using the approximate normality of this index, we determined (separately for each administration) which of these category indices were significantly different from zero, thus indicating DBF for that particular category/administration/group-comparison combination. (Combining over all pretests for a particular administration controls for the problem of false positives that would occur if each category/pretest combination was analyzed separately for DBF. It also greatly increases DBF power by forming much larger category-based bundles to be analyzed for DBF. There are 15 categories, two administrations, and two population comparisons, for a total of 60 such possible $\hat{\beta}$ indices.)

The formation of the above-described bundle $\hat{\beta}$ index (for each category and group comparison combined over pretests within an administration) and a justification of its approximate normal distribution are outlined as follows. Fix the test administration and the group comparison. Let $\beta_{ij}$ denote the estimated bundle DBF for category $j$ items in Pretest $i$. Let $n_{ij}$ denote the number of items in bundle $(i, j)$. Let $n_j = \sum_i n_{ij}$. Thus, $n_j$ is the total number of

category $j$ items across all pretests of the administration. Let $\hat{\beta}_j = \sum_i \hat{\beta}_{ij}$. Then $\hat{\beta}_j / n_j$ is the average $\hat{\beta}$ per item in category $j$, and as such is an indicator of the influence per typical item of category $j$ in creating DBF, in the presence of whatever secondary category-organizing principles also influence the true bundle $\beta$ for the category $j$/administration/group-comparison combination.

In the present situation, since examinees were in effect randomly assigned to one and only one pretest, the $\beta_{ij}$ were independent over all $i$ for each fixed $j$. Thus, it was possible to calculate the $\hat{SE}(\hat{\beta}_j / n_j)$ using the standard approach to computing the standard error of the sum of independent observations. Since the $\beta_{ij}$ over all $i$ for a fixed $j$ were independent and approximately normally distributed, it follows from the central limit theorem that $\hat{\beta}_j / n_j$ was approximately normal with mean $\sum_i E\hat{\beta}_{ij} / n_j$ and standard error $\sqrt{\sum_i (SE(\hat{\beta}_{ij}))^2 / n_j}$. We defined the normalized statistic $B_j$ by dividing $\hat{\beta}_j / n_j$ by its estimated standard error $\sqrt{\sum_i (SE(\hat{\beta}_{ij}))^2 / n_j}$, with $\hat{SE}(\hat{\beta}_{ij})$ found in the usual way (see Shealy & Stout, 1993).

Under the null hypothesis of no DBF, each $B_j$ is distributed approximately standard normal (mean 0, standard deviation 1). Thus, we were able to conduct a hypothesis test of category DBF for each category/administration/group-comparison combination. Indeed, if DBF were present, then the statistic $\hat{\beta}_j / n_j$ (the numerator of $B_j$) would still be distributed approximately normal but with a mean (not equal to zero) indicating the true average amount of DIF per item for the bundle.

Table 3 presents the calculated $\hat{\beta}_j / n_j$ statistic and its observed level of significance found for male-versus-female DBF; Table 4 provides parallel information for Black-versus-White DBF. The observed level of significance (p-value) provided in Table 3 and Table 4 is the hypothesis-testing, observed level of significance corresponding to the rejection region of the null hypothesis of no DBF, for which the hypothesis-testing, null-hypothesis rejection-region boundary is formed using the observed value $\hat{\beta}_j / n_j$. This observed level of significance is thus a measure of the strength of evidence that the category displays DBF, with the usual interpretations for 0.05 and 0.01. For example, a p-value of 0.006 provides strong evidence of DBF, while a p-value of 0.169 provides very weak evidence of DBF. The number of items, $n_j$,

that were classified in each category for each administration is included in the tables as well. The groups favored by categories that demonstrated significant DBF (at the 0.05 level) are indicated in the tables.

**Table 3**

***DBF Analysis for Males-Versus-Females Comparison***

| Category | First administration | | | Second administration | | |
|---|---|---|---|---|---|---|
| | $\hat{\beta}_j / n_j$ | p-value | $n_j$ | $\hat{\beta}_j / n_j$ | p-value | $n_j$ |
| Algebra | 0.000 | 0.960 | 166 | -0.006**$^{F}$ | 0.008 | 185 |
| Calculation Intensive | -0.009 | 0.267 | 17 | 0.005 | 0.450 | 26 |
| Fractions With Numbers | -0.006 | 0.354 | 23 | 0.031**$^{M}$ | 0.000 | 21 |
| Bar Graph | 0.020**$^{M}$ | 0.006 | 23 | 0.020**$^{M}$ | 0.005 | 21 |
| Line Graph | — No items — | | | 0.003 | 0.729 | 21 |
| Pie Graph | 0.008 | 0.222 | 28 | 0.020*$^{M}$ | 0.015 | 17 |
| Table | 0.017*$^{M}$ | 0.024 | 21 | 0.014 | 0.053 | 22 |
| Applied Geometry With Algebra | 0.019**$^{M}$ | 0.007 | 24 | 0.001 | 0.930 | 26 |
| Geometry Without Algebra | 0.010 | 0.169 | 21 | 0.003 | 0.648 | 23 |
| Geometry With Memorized Formulae | 0.010 | 0.180 | 19 | 0.015*$^{M}$ | 0.041 | 19 |
| Number Theory | 0.001 | 0.918 | 43 | -0.006 | 0.127 | 69 |
| Probability and Statistics | 0.017**$^{M}$ | 0.000 | 75 | -0.002 | 0.583 | 100 |
| QCD | 0.027**$^{M}$ | 0.000 | 40 | -0.001 | 0.822 | 38 |
| Speededness | 0.004 | 0.485 | 45 | -0.007 | 0.121 | 56 |
| Word Conversion | 0.014*$^{M}$ | 0.012 | 33 | 0.001 | 0.778 | 75 |

* significant at $\alpha = 0.05$ level; ** significant at $\alpha = 0.01$ level; F = favors females; M = favors males

**Table 4**

***DBF Analysis for Blacks-Versus-Whites Comparison***

| Category | First administration | | | Second administration | | |
|---|---|---|---|---|---|---|
| | $\hat{\beta}_j / n_j$ | p-value | $n_j$ | $\hat{\beta}_j / n_j$ | p-value | $n_j$ |
| Algebra | -0.006 | 0.267 | 166 | -0.003 | 0.603 | 185 |
| Calculation Intensive | -0.026 | 0.113 | 17 | 0.002 | 0.897 | 26 |
| Fractions With Numbers | -0.009 | 0.508 | 23 | 0.019 | 0.331 | 21 |
| Bar Graph | 0.060**$^{W}$ | 0.000 | 23 | 0.034*$^{W}$ | 0.043 | 21 |
| Line Graph | — No items — | | | 0.054**$^{W}$ | 0.005 | 21 |
| Pie Graph | 0.027 | 0.070 | 28 | 0.034*$^{W}$ | 0.046 | 17 |
| Table | 0.030 | 0.052 | 21 | 0.042*$^{W}$ | 0.040 | 22 |
| Applied Geometry With Algebra | 0.026 | 0.066 | 24 | -0.001 | 0.960 | 26 |
| Geometry Without Algebra | 0.040**$^{W}$ | 0.005 | 21 | 0.026 | 0.137 | 23 |
| Geometry With Memorized Formulae | 0.033*$^{W}$ | 0.029 | 19 | -0.007 | 0.696 | 19 |
| Number Theory | 0.003 | 0.737 | 43 | -0.002 | 0.853 | 69 |
| Probability and Statistics | 0.032**$^{W}$ | 0.000 | 75 | 0.008 | 0.370 | 100 |
| QCD | -0.024*$^{B}$ | 0.019 | 40 | -0.028*$^{B}$ | 0.031 | 38 |
| Speededness | 0.033**$^{W}$ | 0.001 | 45 | 0.017 | 0.108 | 56 |
| Word Conversion | 0.036**$^{W}$ | 0.001 | 33 | 0.033**$^{W}$ | 0.000 | 75 |

* significant at $\alpha = 0.05$ level; ** significant at $\alpha = 0.01$ level; W = favors Whites; B = favors Blacks

Looking at the results from Tables 3 and 4, there appear to be many inconsistencies across administrations in the categories found to display significant DBF. However, recalling the natural random variability of DBF indices, a category could in truth be a consistent DBF-producing category, but due to the inherent random variability of examinee responses, would not always be flagged in both administrations as such. A true inconsistency can occur when the secondary dimensions present for a particular category change considerably across administrations, either due to changes from within the bundle-defining category itself (internal heterogeneity) or changes in the amount of overlap of the various secondary categories (external heterogeneity). (These two possibilities are illustrated later in the section entitled, *ANOVA-SIBTEST Approach to Evaluating DBF*.) Such differences can also occur when examinee populations vary in their latent-ability distributions across administrations. This possibility was explored but not judged to be a likely cause of inconsistent DBF results in this setting. Further, if the mean item difficulty for a category changes considerably across administrations, this can seriously influence the amount of true DBF present for that category.

In order to determine which categories produced true inconsistent DBF results across administrations (that is, results that could not be explained by random examinee response variation), the SIBTEST category-based bundle $\hat{\beta}_j / n_j$ pairs from the two administrations were tested to determine if there was a significant difference in their theoretical distributions (in essence, if there was a true difference in their mean values, E($\hat{\beta}_j / n_j$)). Since the SIBTEST administration-based bundle $\hat{\beta}_j / n_j$s divided by their estimated standard errors are distributed approximately normal with standard deviation 1 and are independent across administrations, testing for a difference in two E($\hat{\beta}_j / n_j$)s from the same category but from different administrations is easily accomplished by a standard normal distribution based Z-test. For the males-versus-females comparison, the hypothesis test identified only three inconsistent categories across the two administrations: Fractions With Numbers, QC D, and Probability and Statistics. For the Whites-versus-Blacks comparison, the hypothesis test identified only two inconsistent categories across the two administrations: Geometry With Memorized Formulae and Probability and Statistics. (A method to determine the causes of these inconsistencies is discussed later in the section entitled, *ANOVA-SIBTEST Approach to Evaluating DBF*.)

The above analysis shows that 11 categories out of 14 (for the males-versus-females

comparison) and 12 categories out of 14 (for the Blacks-versus-Whites comparison) display no evidence of inconsistency across administrations. In other words, in these 23 cases, no statistical evidence suggests that the distributions of the SIBTEST bundle $\hat{\beta}_j / n_j$ s are different for the two test administrations. Thus, confidence intervals for the mean amount of category-based DIF per item produced across administrations in these 23 consistent cases can be calculated by *combining* the original SIBTEST bundle $\hat{\beta}_j / n_j$ values from both administrations. Two confidence intervals also exist for the Line Graph category, but these are based only on the second administration. Thus, overall there are a total of 25 confidence intervals.

Table 5 displays the 23 combined-administration confidence intervals, plus the two confidence intervals for the Line Graph category. Each interval in Table 5 can be interpreted as an estimate for a particular category of the average difference (across items of the category) in the probabilities of getting an item correct between the reference and focal groups for matched examinees. Thus, the confidence intervals are on the scale of the amount of DIF per item, making them easy to interpret. A confidence interval that lies entirely to the right or entirely to the left of zero indicates significant DBF for a particular category. If a category displays significant DBF, the examinee group the category *favors* is shown in the table. For the categories for which hypothesis-testing significance holds, the midpoint of the confidence interval is an estimate of the effect size, which is an estimate of the average DIF per item.

As Table 5 shows for the males-versus-females comparison, seven categories show no evidence of producing DBF and five categories show evidence of producing DBF in favor of males. For the Blacks-versus-Whites comparison, five categories show no evidence of producing DBF, one category shows evidence of producing DBF in favor of Blacks, and seven categories show evidence of producing DBF in favor of Whites. It is fascinating — and potentially very useful for future test development efforts — to evaluate from the content/conceptual perspective which of these categories favor males or females and/or Blacks or Whites. The estimated amount of DIF per item (often called the effect size) for each DBF-producing category is of practical importance. This estimate, multiplied by the number of items of the category typically occurring for an administration, yields an estimate of the amount of DBF expected for the category on the total test-score scale.

**Table 5**

***Confidence Intervals for Amount of DIF per Item***

| Category | DIF for males-versus- females (Estimate ± $\hat{SE}$) | DIF for Blacks-versus-Whites (Estimate ± $\hat{SE}$) |
|---|---|---|
| Algebra | -0.0033 ± 0.0035 | -0.0046 ± 0.0083 |
| Calculation Intensive | -0.0008 ± 0.0094 | -0.0093 ± 0.0218 |
| Fractions With Numbers | Not consistent | 0.0040 ± 0.0229 |
| Bar Graph | 0.0196 ± 0.0097$^{M}$ | 0.0473 ± 0.0211$^{W}$ |
| Line Graph | 0.0026 ± 0.0148 | 0.0539 ± 0.0375$^{W}$ |
| Pie Graph | 0.0126 ± 0.0126$^{M}$ | 0.0294 ± 0.0220$^{W}$ |
| Table | 0.0152 ± 0.0101$^{M}$ | 0.0362 ± 0.0253$^{W}$ |
| Applied Geometry With Algebra | 0.0092 ± 0.0090$^{M}$ | 0.0122 ± 0.0211 |
| Geometry Without Algebra | 0.0062 ± 0.0095 | 0.0323 ± 0.0220$^{W}$ |
| Geometry With Memorized Formulae | 0.0126 ± 0.0104$^{M}$ | Not consistent |
| Number Theory | -0.0034 ± 0.0061 | 0.0002 ± 0.0140 |
| Probability and Statistics | Not consistent | Not consistent |
| QCD | Not consistent | -0.0264 ± 0.0164$^{B}$ |
| Speededness | -0.0022 ± 0.0066 | 0.0241 ± 0.0144$^{W}$ |
| Word Conversion | 0.0050 ± 0.0060 | 0.0339 ± 0.0132$^{W}$ |

M = significant DIF in favor of males; W = significant DIF in favor of whites; B = significant DIF in favor of blacks

These results should be highly useful in understanding DBF on the GRE quantitative test. One of the more interesting results from this analysis is that QCD items are highly DBF-producing in favor of blacks. While the content/cognitive reason behind this result warrants further study, from the quantitative perspective, Table 5 captures the core results of our study as it applies to the GRE quantitative and Mathematical Reasoning tests.

Although these categories are designed to be relatively distinct and homogeneous, the

estimated DIF per item for each category must be viewed in context. The average bundle $\hat{\beta}$ per item for a category is the result of averaging over all of the secondary dimensional influences on item performance for all the items of the category. Such possible influences include not only the other 14 defined categories, but also conceptual influences on item performance not explicitly identified as one of our operational categories. Influences on items from outside the operational category-constructs must be expected, since no reasonable item has only one single performance-controlling construct, and the other bundle-organizing categories surely do not include all such secondary dimensional influences. In the next section we attempt to explain the five inconsistencies found across test administrations that were mentioned earlier.

### *ANOVA-SIBTEST Approach to Evaluating DBF*

For relatively dimensionally homogeneous categories, the bundle SIBTEST method described earlier provides a complete and useful analysis of which operational categories produce DBF and their DBF effect sizes. Such an analysis allows investigators to reliably and validly assess the causes of DBF based on the conceptual nature of the categories. However, since our operational categories were constructed to allow for some overlap, the SIBTEST bundle analysis sometimes cannot provide a complete picture of either the DBF present in a category or its conceptual cause. Closely related, the influence of secondary overlapping categories is a likely cause of the statistically significant inconsistencies found between administrations for the five combinations of categories and group comparisons highlighted earlier. By contrast, we *have* been able to successfully analyze the 23 other category/group comparisons, most likely because their categories were sufficiently homogeneous.

In general, inconsistency across administrations for a category in SIBTEST DBF results can come from two possible sources: a) the influence of other secondary overlapping categories from the operational list of categories, or b) heterogeneity within the inconsistent category even when there is no secondary category overlap. These two sources affect the DBF results in the same manner. For either source, the conceptual characteristics influencing differential examinee performance differ from item to item within the category in a way that can cause DBF inconsistencies across administrations. For example, if these secondary dimensions occur in different concentrations in same-category bundles from two different administrations, the DBF results across administrations for the bundles can be inconsistent, even though the bundles

represent the same category-defining principle.

To further illustrate and explain the problematic nature of within-category heterogeneity for DBF analyses, consider a hypothetical case of heterogeneity within a category called Geometry. Its items are dichotomized according to a split in one aspect of their cognitive nature: items requiring the interpretation of a geometric figure and items that do not. Further, suppose these two subcategory item types display equal amounts of DIF in favor of males and females, respectively. Finally, suppose the concentration of these item types is equal in the first administration, but in the second administration, the ratio of the item concentration of the first type to the second type is 3 to 1. Thus, due to the different concentration of items between administrations, a statistical analysis of the Geometry category will likely result in two inconsistent results: For the first administration, our conclusion would be that Geometry is a DBF-neutral category, and for the second administration, our conclusion would be that Geometry is a DBF-producing category in favor of males. This is an example of internal category heterogeneity.

The same example can be modified to illustrate the influence of secondary overlapping categories on the DBF results for a particular bundle-defining category. Suppose we have two final categories called Geometry and Algebra and that we are analyzing the Geometry category for DBF. Suppose pure geometry items are DIF-producing in favor of males, and pure algebra items are DIF-producing in favor of females. Further, suppose that on the first administration, the Algebra category intersects half of the Geometry category items (items in the intersection have both a geometry and an algebra component and hence are assigned to both categories), but on the second administration the Algebra category intersects only 20% of the Geometry items. Due to the change in the concentration of the overlapping Algebra category, the DBF analysis of the Geometry category will likely result in inconsistent results, with the second administration showing significantly more DBF in favor of males than the first administration. Here the category heterogeneity is caused by a change in secondary overlapping categories across administrations.

Since, excluding Probability and Statistics, all 15 final categories were relatively dimensionally homogeneous according to our *h*-index-based hypothesis-testing approach, we conjectured that the bundle SIBTEST inconsistencies across administrations for Fractions With Numbers and QCD for the males-versus-females comparison and Geometry With Memorized

Formulae for the Blacks-versus-Whites comparison were likely due to changes in the concentration of the secondary overlapping categories across administrations. However, since Probability and Statistics was judged statistically to be internally heterogeneous, another possible cause for the two inconsistencies involving this category was the heterogeneity of the category itself.

To explore whether variations in the influence of secondary overlapping categories were the cause of the inconsistencies involving the five inconsistent bundle/group-comparison combinations, an ANOVA procedure was developed. The procedure was applied separately to each category to adjust the DIF statistic $\hat{\beta}$ of each item in the category for the possible DBF influence of its secondary overlapping categories. The ANOVA model was used to form a hypothesis test of whether the inconsistencies between administrations found in the SIBTEST DBF results were due to the influence of variations in secondary overlapping categories or to some other cause — for example, the internal heterogeneity of an individual category, such as Probability and Statistics.

One potential statistical problem with this particular ANOVA analysis is that ANOVA analysis presumes independence of all dependent-variable observations (the item $\hat{\beta}$s). Since DIF $\hat{\beta}$ statistics for items from the same pretest form (all such pretest items being taken by the same subset of the examinee population) are potentially probabilistically dependent on one another, independence of all observations could seriously fail in our data set. However, a comparison of the calculated sample $\hat{\beta}$ variance for the items in the same pretest with the expected variance under the assumption of item $\hat{\beta}$ independence revealed only a slight discrepancy for the various pretests. This indicated that the correlations of the DIF item $\hat{\beta}$s for different item pairs from the same pretest were on average close to zero. Since the SIBTEST item $\hat{\beta}$ statistics were distributed approximately normal (Shealy & Stout, 1993), a near-zero correlation between two such statistics indicates approximate independence. Thus, although strict theoretical independence between item $\hat{\beta}$s was not logically obtained, based on empirical evidence the item $\hat{\beta}$s could be considered to be approximately independent, and hence, the usual ANOVA analysis could be carried out.

***Determining the Causes of Inconsistencies Across Administrations***

Since our main goal in using the ANOVA model was to determine the causes for the inconsistencies observed in the SIBTEST DBF results, our first ANOVA analysis was limited to the study of the five inconsistent category/group-comparison combinations mentioned earlier — Fractions With Numbers, QCD, and Probability and Statistics for the males-versus-females comparison and Geometry With Memorized Formulae and Probability and Statistics for the Blacks-versus-Whites comparison. To illustrate the ANOVA model that was developed for each of the five inconsistent categories, we use the QCD category for the males-versus-females comparison as an example. The other four category/group-comparison combinations were studied in a similar manner.

As Table 3 and Table 4 show, 40 items from the first administration and 38 items from the second administration were classified in the QCD category. A single QCD ANOVA model was defined to include both administrations. This ANOVA model used the individual, item SIBTEST $\hat{\beta}$s for each of the 78 QCD items as the dependent variable for the model. As such, the dependent variable in the ANOVA model served as a measure of the gender-based DIF associated with each of the QCD items.

The independent variables in the ANOVA model were a series of indicator variables for the secondary overlapping categories — coded 1 when the item was influenced by a particular secondary overlapping category and coded 0 when it was not. Overlapping categories were included in an ANOVA model if they overlapped with at least 10% of the items in the studied category from either the first or second administration. In other words, all other categories present in four or more QCD items either from the first or second administration were included as indicator variables in the ANOVA model. In this manner, the Word Conversion, Algebra, Number Theory, and Probability and Statistics categories were established as indicator variables. Finally, an additional indicator variable, Administration — representing whether an item belonged to the first or second administration — was added to the model.

Each factor of the ANOVA model occurs at two levels. The QCD ANOVA model for each QCD item can be written structurally in the following way:

$$\begin{aligned}\hat{\beta} = b_0 + b_1(\text{Word conversion}) + b_2(\text{Algebra}) + b_3(\text{Number theory}) \\ + b_4(\text{Probability \& statistics}) + b_5(\text{Administration}) + e\end{aligned} \quad (2)$$

Here, *e* denotes random error, the variables in parentheses are the indicator variables, and $b_0$, $b_1$ ... $b_5$ are the unknown regression coefficients of the model.

In developing this ANOVA model, the SIBTEST item statistic $\hat{\beta}$ was decomposed into three main sources of DIF:

- the amount of DIF due purely to the bundle-defining QCD category itself and captured by the intercept of the model ($b_0$)
- the amount of DIF due to each secondary overlapping category and captured by each category coefficient in the model ($b_1$ through $b_4$)
- the possible change in the amount of DIF between test administrations and captured by the Administration coefficient in the model ($b_5$)

If the intercept of this model was found to be significantly different from zero, then items in the QCD category were determined to display a significant amount of DIF after adjusting for the influence of overlapping categories and the influence of Administration. Thus, the intercept denotes the amount of DBF associated with the pure, defining aspect of the category, as expressed by the definition of the category found in Table 2. If one of the coefficients corresponding to the four overlapping categories was significantly different from zero, then the amount of DIF for QCD items that also belong to this overlapping category was determined to be significantly different than the amount of DIF occurring in QCD items not belonging to the overlapping category. Finally, if the coefficient corresponding to the Administration variable was significantly different from zero, then the amount of DIF for the QCD items between administrations was determined to be significantly different, even after adjusting for the effects of the various overlapping categories. In this case, secondary categories could not explain the observed inconsistency across administrations. When our focus was on whether there was an Administration effect (an inconsistency across administrations), the role of the other parameters of the model was to adjust for various possible covariates. Including these covariates in the model produced a model with good power to detect an Administration effect.

Using the statistical software package, SAS® (SAS Institute), and the method described above, the ANOVA models for each of the five inconsistent categories were estimated. Since the main focus was to determine whether the observed differences in the amount of DIF between administrations were statistically significant, the significance of the coefficient (that is, the

coefficient was statistically judged to be nonzero) corresponding to the Administration variable was tested at the $\alpha = 0.05$ hypothesis-testing level of significance for each of the five ANOVA models. For three category/group-comparison combinations — Fractions with Numbers for males versus females and Probability and Statistics and Geometry With Memorized Formula for Blacks-versus-Whites — the coefficient corresponding to the Administration variable was not found to be significantly different from zero. This SIBTEST ANOVA finding implies that, once the influence of the other secondary overlapping categories was accounted for, there was no evidence of a statistically significant difference in the amount of DIF between administrations for these three category/group-comparison combinations. In this manner, the ANOVA method explained the cause of the inconsistent SIBTEST DBF results for three out of the five inconsistent category/group-comparison combinations, namely secondary overlapping categories. Interestingly, for the Probability and Statistics/Blacks-versus-Whites combination, the internal heterogeneity of the category was thus not the identified source of the observed inconsistency (although it could also be contributing).

However, for the other two category/group-comparison combinations — QCD and Probability and Statistics for males versus females — the coefficient corresponding to the Administration variable was significantly different from zero in both cases. This SIBTEST ANOVA finding implies that the difference in the amount of DBF between administrations for these two category/group-comparison combinations was statistically significant, after adjusting for the influence of overlapping categories and allowing for natural, random examinee response fluctuations. Thus, there must be other unidentified influences that account for the inconsistencies in the observed $\hat{\beta}$ s across administrations for the Probability and Statistics and QCD categories for the males-versus-females comparison.

Throughout our analyses, the Probability and Statistics category was determined to lack internal dimensional homogeneity. (We examine the heterogeneous nature of this category more closely in a later section, *Impact Analysis of Final Categories*.) Thus, we have always suspected the DBF results for this category might not be consistent across test administrations in part due to this within-category heterogeneity. The QCD category, by its definition (as noted earlier, items are never defined as just belonging in the QCD category alone), can be highly influenced by other secondary categories. Probability and Statistics was included in the QCD ANOVA model because it occurs in more than 10% of the QCD items in both administrations. Thus, a possible

cause of the inconsistency of the DBF results for the QCD category for males versus females could be internal changes between administrations in the composition of the heterogeneous overlapping category, Probability and Statistics. This possible explanation is somewhat subtle: The possible cause is not variation in the amount of Probability and Statistics item overlap with QCD across administrations (such variation in overlap of secondary categories as was found to be the cause for the three cases above), but could be variation in the internal composition across administrations of the overlapping and heterogeneous Probability and Statistics category.

Using the ANOVA model for the QCD category described above, we were able to test this theory by splitting the Probability and Statistics indicator variable into two independent indicator variables according to the test administration of the Probability and Statistics items. In effect, using two variables for the Probability and Statistics category allowed the category to literally function as a different category in the each of the two administrations, even though it had the same name. In this manner, the modified QCD ANOVA model was able to determine whether the differences across administrations in the internal composition of the Probability and Statistics category affected the amount of DIF that was present in the QCD items. Analyzing this new ANOVA model for the QCD category, the coefficient corresponding to the Administration variable was no longer statistically significantly different from zero. This implies that the difference in the amount of DBF between administrations for the QCD category was explainable by the internally heterogeneous nature of the secondary overlapping Probability and Statistics category. Thus, analysis of our modified ANOVA model explained the cause of the inconsistent SIBTEST DBF results for the QCD category.

By using the ANOVA approach to study category-based item DIF, causes for four out of the five observed category/group-comparison inconsistencies were found — three due to variation in secondary overlapping categories across administrations and one due to variation in the internal composition of a secondary overlapping category across administrations. The fact that 23 of the 28 category/group-comparison combinations were consistent according to our hypothesis-testing approach, described earlier, illustrates that the SIBTEST bundle analysis usually produces consistent and useful results for carefully constructed and relatively homogenous categories, even when some secondary category overlap is present.

***SIBTEST ANOVA Analysis of All 15 Operational Categories***

While our primary goal in developing the ANOVA method described above was to determine whether changes in overlapping secondary categories were responsible for the five inconsistent category-based results across administrations found in the SIBTEST DBF analysis, the method can be slightly modified and applied separately to every one of the 15 operational categories to form an overlapping, category-based, decompositional analysis of the amount of DBF present in each category, for each group comparison and for either each administration separately or both administrations combined. While a thorough category-based DBF analysis has already been conducted using the SIBTEST bundle method, the ANOVA method can sometimes be of further value because it enables us to study the DIF for an item in a particular category decomposed into the influence of that category itself and the influence of other secondary overlapping categories. By thus controlling for the influence of the secondary overlapping categories, the ANOVA method can then assess the "true" DBF associated with each category-organizing principle for each group-comparison/administration combination.

If the combined influence of the overlapping categories on item-level DIF is small, then the categories flagged using the ANOVA method (categories with intercepts that were judged to be significantly different from zero) should be similar to those found by the SIBTEST bundle method (provided both methods are equally statistically powerful!). However, if the combined influence of secondary overlapping categories on DBF is sizeable, the categories flagged using the ANOVA method will likely often vary from the categories flagged using the SIBTEST bundle analysis. By combining the two analyses, the DBF associated with each category can be analyzed both in the context of other secondary dimensional category influences (by way of the SIBTEST bundle analyses) and with other category influences separated out (by way of the SIBTEST ANOVA method). We stress that for relatively statistically homogeneous categories, this context issue is largely inconsequential because the influence of overlapping categories at the bundle level is minor.

To complete the analyses, we used the SIBTEST ANOVA method to assess the presence of DBF for each final category (judged by whether the category intercept was significantly different from zero) for each administration separately. Since there were two administrations, 15 categories, and two types of DIF studied (Blacks-versus-Whites and males-versus-females), and the two ANOVA models for the category Line Graph from the first administration were not

calculated because of the lack of such items, 58 ANOVA models were each statistically analyzed. The components of the ANOVA models were exactly the same as those described for the SIBTEST bundle method, except for the lack of an Administration indicator variable. (Since only items from one administration were used for each analysis, the Administration variable was superfluous.) Thus, each ANOVA model included an intercept and indicator variables for all categories that overlapped with at least 10% of the items for the category modeled. If the intercept of a particular ANOVA model was found to be significantly different from zero, then items in this category were determined to display a significant amount of DBF even after adjusting for the influence of overlapping categories and the natural randomness of the item $\hat{\beta}$s. If one of the coefficients corresponding to the overlapping categories was found to be significantly different from zero, then the amount of DIF for items in this category that also belonged to the overlapping category was determined to be significantly different than the amount of DIF occurring for items in this category that do not belong to the overlapping category.

For the first and second administrations, respectively, Table 6 and Table 7 display categories with significant intercepts in their ANOVA models for each group comparison. In order to provide a comparison of the two approaches, the categories found to produce significant DBF from the earlier SIBTEST DBF bundle analyses are also included in the tables. (See Tables 3 and 4 for the original SIBTEST DBF results.) Inconsistencies between ANOVA and SIBTEST bundle results are denoted by an asterisk. Each table includes a list of any statistically significant overlapping categories (as determined from the ANOVA models) for each of the categories included in that table. After each such category, the sign of its coefficient is given, with a plus-sign indicating a contribution to DBF against the focal group (females or Blacks) and a minus-sign indicating a contribution to DBF against the reference group (males or Whites). When an inconsistency between the ANOVA and SIBTEST bundle results is explainable by the presence of one or more significant overlapping categories, as shown by the ANOVA analysis, this is also indicated in the tables.

**Table 6**

***ANOVA Results From the First Administration****

| — DBF for males versus females — | |
|---|---|
| ANOVA results | SIBTEST DBF results |
| Applied Geometry With Algebra $^{M}$<br>QCD $^{M}$<br>Bar Graph $^{M}$<br>Table $^{M}$<br>Probability and Statistics $^{M}$ | Applied Geometry With Algebra $^{M}$<br>QCD $^{M}$<br>Bar Graph $^{M}$<br>Table $^{M}$<br>Probability and Statistics $^{M}$<br>Word Conversion $^{MD}$ |
| Significant overlapping categories ($\alpha$ = .05): | None |
| **— DBF for Blacks versus Whites —** | |
| ANOVA results | SIBTEST DBF results |
| Geometry Without Algebra $^{W}$<br>Bar Graph $^{W}$<br>Probability and Statistics $^{W}$<br>Table $^{W}$<br>Algebra $^{BDE}$<br>Number Theory $^{WDE}$ | Geometry Without Algebra $^{W}$<br>Bar Graph $^{W}$<br>Probability and Statistics $^{W}$<br>Table $^{W}$<br>QCD $^{BDE}$<br>Speededness $^{WD}$<br>Word Conversion $^{WD}$<br>Geometry With Memorized Formulae $^{WD}$ |
| Significant overlapping categories ($\alpha$ = .05): | Algebra (-) with Word Conversion (+)<br>QCD (-) with Algebra (+)<br>Number Theory (+) with QCD (-) |

* Categories with significant DBF ($\alpha$ = .05) only
M = favors males, F = favors females
B = favors Blacks, W = favors Whites
D = discrepancy between ANOVA and SIBTEST results
E = Inconsistency explained by overlapping categories

**Table 7**

***ANOVA Results From the Second Administration****

| — DBF for males versus females — | |
|---|---|
| ANOVA results | SIBTEST DBF results |
| Fractions With Numbers [M]<br>Table [MDE] | Fractions With Numbers [M]<br>Algebra [FDE]<br>Geometry With Memorized Formulae [MD]<br>Bar Graph [MD]<br>Pie Graph [MD] |
| Significant overlapping categories ($\alpha$ = .05): | Algebra (-) with Bar Graph (+)<br>Table (+) with Probability and Statistics (-) and Algebra (-) |
| **— DBF for Blacks versus Whites —** | |
| ANOVA results | SIBTEST DBF results |
| Algebra [BDE]<br>Word Conversion [W] | QCD [BDE]<br>Word Conversion [W]<br>Line Graph [WD]<br>Bar Graph [WD]<br>Pie Graph [WD]<br>Table [WD] |
| Significant overlapping categories ($\alpha$ = .05): | Algebra (-) with Speededness (+) and Word Conversion (+)<br>QCD (-) with Word Conversion (+) |

* Categories with significant DBF ($\alpha$ = .05) only
M = favors males, F = favors females
B = favors Blacks, W = favors Whites
D = discrepancy between ANOVA and SIBTEST results
E = Inconsistency explained by overlapping categories

Over both test administrations, a total of six inconsistencies were found between the SIBTEST ANOVA and SIBTEST DBF bundle results for the males-versus-females comparison, and a total of 12 inconsistencies were found for the Blacks-versus-Whites comparison. Two out of the six inconsistencies for the males-versus-females comparison are explained by the presence of significant overlapping categories (with the coefficients having the appropriate signs), while five out of the 12 inconsistencies for the Blacks-versus-Whites comparison are similarly explained. Thus, the tables point to a total of seven explained inconsistencies.

In each of the remaining 11 unexplained cases of inconsistency, the category was

determined to be significantly DBF-producing using the SIBTEST DBF bundle analysis, but was not determined to be significantly DBF-producing using the SIBTEST ANOVA method. For these 11 cases, the likely cause of inconsistency was a lack of statistical power on the part of the SIBTEST ANOVA method because category sizes were small. The power of a statistical test is defined as the probability of correctly rejecting a false null hypothesis. As applied to our analysis, the power of one of our statistical tests was the probability of correctly rejecting the null hypothesis of zero DBF for a category (correctly concluding the category is significantly DIF-producing).

To illustrate the difference in power between the SIBTEST DBF bundle and ANOVA procedures, consider a DBF analysis of a category-based bundle of 20 items. In the SIBTEST bundle-DBF analysis, all 20 items would be used to estimate the average amount of DIF per item for the category. However, in the SIBTEST ANOVA analysis, the same 20 items would not only be used to estimate the intercept, but also to estimate each of the coefficients corresponding to the secondary overlapping categories of the model. Furthermore, in our example, the estimation of the coefficients for the overlapping categories would be accomplished using a very small number of overlapping items per overlapping category — as few as, and likely not much more than, two items. Thus, when the category being modeled is small, our statistical ability to estimate and test the influence of an overlapping category on the data would be severally limited.

With only two exceptions (Speededness and Word Conversion), each of the categories determined to be significantly DBF-producing using the SIBTEST DBF method, but determined not to be significantly DBF-producing using the ANOVA method, contained fewer than 25 items. Therefore, the fairly sizeable number of inconsistencies found between the two category-based DBF methods appears to be almost entirely explained by a lack of power on the part of the SIBTEST ANOVA method for small categories.

### *The Heterogeneity of the Probability and Statistics Category*

Based on the SIBTEST bundle analysis, the Probability and Statistics category produced inconsistent results with regard to DBF across the two test administrations for both group comparisons. During the original formation of the categories, Probability and Statistics was chosen more to provide an exhaustive set of categories (recall its definition in Table 2) than for the homogeneous nature of its items. Hence, one likely explanation for the inconsistent DBF results is internal heterogeneity within the category, especially for the unexplained males-versus-

females inconsistency.

In order to examine the internal heterogeneity of the category and thereby help explain the inconsistent DBF results, the Probability and Statistics category was split into smaller, content-specific, and disjoint (i.e., mutually exclusive) microcategories. The microcategories were formed using internal ETS GRE-quantitative content classifications that, by design, are distinct and extremely content-specific. The resulting nine micro-categories that were classified as belonging to the Probability and Statistics category are:

- Combinatorics
- Probability
- Arithmetic Mean
- Weighted Mean
- Median
- Mean/Median Comparison
- Range
- Standard Deviation
- Percentile

Focusing on the males-versus-females comparison and using the SIBTEST bundle method, the overall amount of DBF present in each of the microcategories was then calculated for each administration. As a result, two of the nine microcategories were found to be significantly DBF-producing in favor of males: Probability and Standard Deviation (Probability in both administrations and Standard Deviation in the first administration — the only administration in which more than one item involved standard deviations). In addition, two microcategories were found to be significantly DBF-producing in favor of females: Median and Median/Mean Comparison (both categories in both administrations).

For the first test administration, out of a total of 80 items classified into the nine microcategories, 21 items were classified as Probability, eight items were classified as Standard Deviation, nine items were classified as Median, and seven items were classified as Median/Mean Comparison. For the second test administration, again out of 80 items, 24 items

were classified as Probability, one item was classified as Standard Deviation, nine items were classified as Median, and six items were classified as Median/Mean Comparison. The relative influence of three of these microcategories — Probability, Median, and Median/Mean Comparison — was approximately the same for both administrations. However, the relative influence of the Standard Deviation microcategory was considerably greater in the first administration. In addition, the SIBTEST bundle analysis of the Standard Deviation microcategory found this category to be extremely DBF-producing in favor of males, with an observed significance level (p-value) of 0.0001 in the first test administration. Thus, for the males-versus-females comparison, the DBF results for the Probability and Statistics category should have been significantly affected by the change in the composition of the items between the two administrations. We thus have an explanation for the category's only remaining unexplained inconsistency between administrations — namely, the internal heterogeneity of the Probability and Statistics category.

From our analysis, we have shown that the Probability and Statistics category included several, smaller and more specific cognitive constructs that influenced examinee performance differently. Four of these cognitive constructs were determined to have a significant amount of DBF potential associated with them — two in favor of males and two in favor of females. The difference in the DBF results for the males-versus-females comparison across administrations was very likely due to the relative influence of the Standard Deviation microcategory. The subtle way in which the Standard Deviation microcategory became an important determinant of DBF for the much broader Probability and Statistics category seems instructive.

### *Impact Analysis of Final Categories*

Impact was defined earlier as the average score difference on an item or group of items between two distinct groups of examinees without controlling for examinee ability. By contrast, SIBTEST measures the average score difference on an item or group of items between two distinct groups of examinees who are matched on some valid ability scale, such as their score on the operational items used for our GRE quantitative analysis. Using our working assumption that this test score provides a reasonably equitable means of assessing examinees on the GRE quantitative test, then we have a procedure for finding category-based bundles that display statistically significant DBF for examinees matched on a valid indicator (operational items test score) of the dominant construct the test is trying to measure (e.g., quantitative reasoning). By

contrast, category-based bundle *impact* measures the difference in score performance on that bundle between two groups with no attempt to match examinees in any way —particularly not in a way that is correlated with what the GRE quantitative measures. Since impact describes directly observable group-based score difference on category bundles, it is also vital to consider impact, in addition to DBF, when designing and evaluating a test for fairness and influence on different examinee groups.

To determine the amount of impact associated with each category for each administration, the impact of each individual item was calculated by taking the difference in the proportion of examinees that answered the item correctly from the two groups. Then, given each individual item-impact value, the total impact of a category bundle was calculated by taking the sum of the item-impact values within each category over all pretests. Finally, the mean of the item-impact values was calculated for each administration and category. Table 8 provides the final average impact per item for each category and each administration for the males-versus-females comparison, and in Table 9 shows parallel findings for the Whites-versus-Blacks comparison. In both tables, positive values indicate impact against the focal group (females or Blacks). In addition to Table 5, Table 8 and Table 9 are also potentially very useful for future efforts to improve the equity of the GRE quantitative and Mathematical Reasoning examinations.

The impact per item values for all categories, both administrations, and both examinee population group-comparisons are positive and sizeable. This reflects what has been widely observed about score distributions for these groups on the GRE quantitative test . It is certainly worth noting that no attempt was made in this study to investigate possible covariates — such as the number of engineering, science, and mathematics courses examinees had taken, grades they received in such courses, their planned career paths, and so on. For example, it is surely the case that the proportion of men in technical career paths is higher than that of women among those who took the GRE quantitative exam on these two administrations.

**Table 8**

***Impact Values for Males Versus Females***

| Category | First administration (impact per item) | Second administration (impact per item) |
|---|---|---|
| Algebra | 0.0888 | 0.0865 |
| Calculation Intensive | 0.1080 | 0.0819 |
| Fractions With Numbers | 0.0940 | 0.0749 |
| Line Graph | No Items | 0.0867 |
| Bar Graph | 0.1171 | 0.0935 |
| Pie Graph | 0.0804 | 0.1005 |
| Table | 0.0795 | 0.0799 |
| Applied Geometry With Algebra | 0.0813 | 0.0741 |
| Geometry Without Algebra | 0.1124 | 0.0804 |
| Geometry With Memorized Formulae | 0.1007 | 0.0850 |
| Number Theory | 0.0931 | 0.0834 |
| Probability and Statistics | 0.0831 | 0.0869 |
| QCD | 0.0928 | 0.0762 |
| Speededness | 0.0870 | 0.0932 |
| Word Conversion | 0.1010 | 0.0907 |

**Table 9**

***Impact Values for Blacks Versus Whites***

| Category | First administration (impact per item) | Second administration (impact per item) |
|---|---|---|
| Algebra | 0.1154 | 0.1202 |
| Calculation Intensive | 0.1369 | 0.0868 |
| Fractions With Numbers | 0.1336 | 0.0988 |
| Line Graph | No Items | 0.1156 |
| Bar Graph | 0.1245 | 0.1381 |
| Pie Graph | 0.0956 | 0.1464 |
| Table | 0.1207 | 0.1169 |
| Applied Geometry With Algebra | 0.1310 | 0.1184 |
| Geometry Without Algebra | 0.1319 | 0.1097 |
| Geometry With Memorized Formulae | 0.1299 | 0.1364 |
| Number Theory | 0.1194 | 0.1212 |
| Probability and Statistics | 0.1060 | 0.1191 |
| QCD | 0.1230 | 0.1186 |
| Speededness | 0.1208 | 0.1326 |
| Word Conversion | 0.1317 | 0.1309 |

### *Evaluating the DIF Effects of Studied Latent Dimensions: An Experimental Approach*

One advantage of the SIBTEST bundle detection method is its close connection to the Roussos/Stout multidimensional DBF model in which the contribution of latent conceptually organized dimensions to DBF is postulated (Roussos & Stout, 1996). Since the $\hat{\beta}$ bundle value for a category that is designed to be dimensionally homogeneous is easy to compute because it is the sum of the item $\hat{\beta}$ values for this bundle, SIBTEST bundle $\hat{\beta}$ values are well suited for quantifying the effects of dimensions contributing to DBF and are also easy to obtain. The ANOVA approach, which appears to have some capacity to assess the relative contributions of category-produced latent dimensions to DBF, should become more statistically powerful and informative when pretest items are intentionally designed to test the effects of various targeted dimensions, as defined by a variety of conceptual organizing principles (such as item content, item format, position of the item on the test, cognitive characteristics, and so on).

For example, in evaluating the relative contributions of two dimensions (say, dimensions

A and B) defined at two levels that indicate the absence or presence of the dimensions influencing examinee performance (say A0, A1 and B0, B1), we may construct a factorial design involving four item variations that are cognitively/conceptually the *same* in all respects except for their levels on A and B. The advantage of this approach is that by controlling for other sources of potential DBF by keeping the core of the item constant, we can more effectively isolate and evaluate the effects of A and B with better statistical power of detection. Using such an experimental approach, confirmation of dimensions suspected as contributing to DBF (along the lines of the Roussos & Stout latent dimensionality approach to DBF, 1996) would seemingly be more accurately tested in the experimental context, rather than depending on the observational data approach presented above.

Luckily, sets of items that appear to satisfy such an experimental study design actually did occur among the pretest items on the second administration. Such items are said to constitute an *item group*, which is defined as a group of items that all possess the same core, but are embellished differently by varying certain factors. For example, Figure 2 displays a core word problem involving the construction of an algebraic equation for determining the cost of various types of purchases. Variants of this item — which had been created for the GRE quantitative test by modifying the core with respect to a) the object being purchased, b) the gender of the purchaser, c) the concrete (numeric) versus abstract (algebraic) nature of the problem, and d) whether the item appeared in "long form" or "short form" — were present across pretest administrations (although all 16 combinations did not occur).

A SIBTEST ANOVA was conducted for this item group to determine the effect on DIF of changes in a) the type of product sold (whether a book or a drink), b) whether the person in the question was identified as a manager or not, c) the type of question form (long versus short), and d) the variable type (numeric or algebraic) for males and females, as measured by item $\hat{\beta}$. A Scheffe ANOVA test for the comparison of adjusted means indicated there is a significant difference in DIF at the $\alpha = .05$ level between questions with a numeric value (item $\hat{\beta} = -0.016$) and questions with an algebraic variable instead (item $\hat{\beta} = -0.050$). There were no other observed significant differences, although the power of this study may be low for small differences.

| Item Group 1 | |
|---|---|
| Short form | Long form |
| A refreshment stand sells small boxes of juice for $0.60 each and large boxes of juice of $0.90 each. Jane bought T large boxes of juice. If she spent the same amount of money on small boxes as on large boxes of juice, how many small boxes of juice did Jane buy? | A refreshment stand sold juice yesterday for $0.60 a box and sold 150 of the T boxes they had on hand. Today the manager wants the remaining boxes to bring in the same total as the juice sold yesterday. At what price, in dollars per box, must the manager sell the remaining boxes of juice? |
| The underlined words may be changed to any of the following to produce other item variants:<br>• juice: notebooks, textbooks, sodas<br>• Jane, manager: Marie (no status indicator)<br>• T: 8 | |

***Figure 2.*** **Short-form and long-form variants of a GRE quantitative item.**

Figure 3 presents a second item group that was analyzed (again, not all 16 combinations of possible item variations occurred). A SIBTEST ANOVA was performed to determine the effect on DIF (measured by $\hat{\beta}$) of changes in a) mode of transportation, b) male name versus female name, c) type of question form (long versus short), and d) the variable type (numeric or algebraic) for males and females. A Scheffe ANOVA test for the comparison of adjusted means indicated there is significant difference in DIF ($\alpha = .05$) between questions with male names (item $\hat{\beta} = 0.040$, as measured by the estimated intercept) and questions without male names (item $\hat{\beta} = -0.0110$, as measured by the estimated intercept) — an interesting finding. No other significant differences were observed, although the power of this study is likely low for small differences due to the limited number of items available for analysis.

While the specific results of the above 1995-1996 GRE-quantitative item groups do not seem widely applicable, the methodology used to isolate the influence of various conceptual factors on DBF can be applied to many settings. In fact, see Bolt (2000), for a much more extensive study of this experimental item-based approach using the SIBTEST bundle method. Interestingly, our analysis of the two item groups above suggests that relatively noncognitive dimensions (e.g., male versus female names) can influence DBF, as has been suggested by many researchers.

| Item Group 2 | |
|---|---|
| Short form | Long form |
| Joe bikes at a constant speed of 12 miles per hour and Pat bikes at a constant speed of 20 miles per hour. How many hours does it take Joe to travel the distance that Pat travels in 6 hours? | Joe bikes at a speed of 9 miles per hour on a certain bike trail for 4 hours and then turns around for the return trip along the same trail. He scheduled a total of 6 hours biking for the complete trip. If Joe is to complete his trip exactly on schedule, at what speed, in miles per hour, must he bike for the return trip? |
| The underlined words may be changed to any of the following to produce other item variants:<br>• Joe: John, Juanita, Joe<br>• Pat: Antonia, Marie<br>• bikes: takes a bus, drives a car, jogs | |

***Figure 3.*** **Short-form and long-form variants of a second GRE quantitative item.**

## Discussion and Conclusions

By further developing and using the multidimensional DBF paradigm of Roussos and Stout (1996), we achieved our central goal of discovering and assessing conceptually-caused differences in performance on the GRE quantitative test between males and females and also between Blacks and Whites. The DBF paradigm calls first for the development of conceptually meaningful and statistically discernable latent dimensions (conceptually based categories) that influence examinee performance on items. As detailed earlier, a preliminary set of 26 categories of GRE-quantitative items was first developed from three sources. These categories were then designed to be conceptually meaningful and exhaustive. By applying two additional criteria — (a) relative, within-category dimensional homogeneity and (b) approximate independence or negative association — to achieve relative distinctiveness between all pairs of categories, the 26 preliminary categories were refined to a final set of 15 operational categories. In the statistical analysis of internal homogeneity of these final categories, the QCD category was found to be highly dimensionally homogeneous, a finding that implies the existence of a cognitively unique, strongly influential, and possibly largely construct-irrelevant common dimension to all QCD items.

An analysis of the amount of DBF for each category suggested several possible inconsistencies in DBF results across the two administrations studied. However, using a

hypothesis-testing approach, only five out of 28 categories were found to produce statistically significant inconsistencies between administrations. For each of the 23 consistent category/group-comparison combinations, confidence intervals for the amount of category-based DIF per item — calculated for the combined administrations — showed that five categories were significantly DBF-producing in favor of males, seven categories were significantly DBF-producing in favor of Whites, and one category was significantly DBF-producing in favor of Blacks. Concerning the issue of the equity of the GRE quantitative test and the proposed GRE Mathematical Reasoning test, it would seem that the results of this analysis (presented earlier in Table 5) should be carefully studied, analyzed, and applied to the development of future GRE quantitative and Mathematical Reasoning tests.

Consider the five inconsistent categories that were found across the two test administrations. Since the operational categories developed for this study were not disjoint, inconsistencies in the amount of DBF observed across administrations could be caused by changes across administration in the concentration of secondary overlapping categories. As a result of an ANOVA analysis, changes across administration in secondary overlapping categories were found to account for three out of these five inconsistencies. Inconsistencies in DBF results across administrations in the QCD and Probability and Statistics categories for the males-versus-females comparison remained to be explained.

A SIBTEST bundle DBF analysis of the types of items included in the Probability and Statistics category in each of the two administrations showed two homogeneous microcategories significantly favoring males (Probability and Standard Deviation) and two homogeneous microcategories significantly favoring females (Median and Median/Mean Comparison). While the relative influence of three out of four of these significant microcategories was the same in both administrations, the Standard Deviation microcategory (DBF observed level of significance p-value of 0.001) was heavily represented in the first administration and not in the second administration. Thus, the cause of the inconsistency for the Probability and Statistics category for the males-versus-females comparison between administrations appears to have been the change in the relative influence of one of the Probability and Statistics microcategories, Standard Deviation.

Since one of QCD category's overlapping categories was Probability and Statistics, the inconsistency between administrations in the QCD category could also have been due to the

changing character of the Probability and Statistics category across administrations. When the ANOVA approach was used to adjust for the different character of the Probability and Statistics category across administrations, the QCD category was no longer found to be inconsistent across administrations. In effect, the internal change in the nature of the Probability and Statistics category between administrations seems to have also caused the inconsistency in the DIF results for the QCD category. Thus all five DBF inconsistencies between administrations were explained.

The ANOVA approach was then slightly modified to decompose the amount of DBF for each particular category into two parts: the "true" amount of DBF for the category and the amount of DBF due to other overlapping categories. The results of this ANOVA approach as applied to the GRE quantitative test are presented in Tables 6 and 7. Although interesting because of the removal of confounding influences of secondary overlapping categories, these results lack the statistical power of SIBTEST DBF hypothesis testing approaches.

While our analysis of the average impact per item for each category showed that the amount of impact did vary somewhat from category to category, a positive and large impact was found for every category across both administrations and for both group comparisons. However, we must again emphasize that no covariates were analyzed in the impact studies, such as the number of technical courses examinees had taken. Table 8 and Table 9, which summarize the results of the impact analyses by group comparison, capture some of the project's central findings. Like Table 5, the results found in these tables may be potentially useful for future equity work on the GRE quantitative and Mathematical Reasoning tests.

An experimental ANOVA analysis — in which certain categories or dimensions of items were manipulated to determine how particular changes influenced DBF —suggested that relatively noncognitive and superficial item characteristics can sometimes contribute significantly to DBF, as has long been suspected by many DBF researchers. While this analysis is preliminary, the experimental method that was used is very promising and should be explored further (see Bolt, 2000, for a report of such an effort). It could become a valuable tool in future test equity methodology.

This study was designed to function on two levels: a) as a presentation of GRE-quantitative DBF and impact results that are relevant to future GRE quantitative and Mathematical Reasoning tests, and b) more generally, as a demonstration of a highly useful

methodology that combines statistical and content/cognitive considerations to produce a powerful tool for the study of DBF and impact at the category level, with the goal of improving test equity on all standardized tests. The methodology reported in this paper, which extended the DBF paradigm of Roussos and Stout (1996) and was facilitated by the use of SIBTEST, served to determine conceptual causes of DBF and impact in the GRE quantitative test. The analyses described in this paper can be applied to future test design and evaluation of the GRE quantitative and Mathematical Reasoning tests, as well as to other quantitatively oriented, standardized tests. In general, this research can also serve as the basis of further study in the field of test equity. Indeed, we would argue that the methodology developed for this study can greatly help the testing industry use statistical DBF and impact analyses as integral tools in the development and maintenance of equitable tests.

We again call practitioners' attention to the fact that all instances of DBF do not constitute inequity. Secondary DBF-causing dimensions can be central to the construct a test is designed to measure. By contrast, impact with no DBF present can sometimes cause inequity, because impact can vary over equally valid tests for the same target construct but for which test specifications differ.

In essence, the multidimensional DBF paradigm of Roussos and Stout (1996) and the related SIBTEST bundle methodology developed for this study elevate the statistical quality-control role of DIF (DBF) analyses from the mere removal of already manufactured, equity-defective items to a new, more central role: improving test equity by using the SIBTEST bundle methodology to modify the test specification and manufacturing process itself. The SIBTEST bundle method, of course, is just one specific example of the modern proactive statistical approach to quality control.

## References


Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychometrical Measurement, 20*, 311-330.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), & Joint Committee on Standards for Educational and Psychological Testing (JCSEPT). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*, 307-327.

Bolt, D., Roussos, L., & Stout, W. (1998). *Estimation of item dimensional measurement direction using conditional covariance patterns* (Law School Admission Council Computerized Testing Report No. 98-02). Newtown, PA: Law School Admission Council.

Douglas, J., Roussos, L., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 335-371.

Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical bundle analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1-30.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

Willingham, W. W., & Cole, N. S. (Eds.). (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152.

## Notes

1 As a technical point rendering the use of a conditional covariance as the measure of dimensional homogeneity of an item pair not totally effective, we note that a conditional covariance will be small for two dimensionally homogeneous items for which the directions of best measurement are both close to the test's direction of best measurement (see Zhang & Stout, 1999). There are ways of compensating for this that lie outside the focus of this report.

2 Note the apparently paradoxical result that the QCD category was found to be highly homogenous statistically, when in fact every QCD item always belongs to at least one other category. From the content/cognitive perspective, the existence of a strongly distinct and very influential cognitive processing aspect to QCD problems seems to transcend and dominate the also-influential mathematical content components and their associated cognitive processes. This was also the tentative conclusion of our research team, based on their experience with solving QCD problems.

From a multidimensional latent space IRT modeling perspective, this suggests that the discrimination coefficient corresponding to the cognitively unique aspect of QCD items is relatively large compared with the discrimination coefficients corresponding to the overlapping, category-based, latent secondary dimensions. This unique and highly influential cognitive aspect of QCD items could be problematic if seen by test development experts as peripheral, or even irrelevant, to the central measurement purpose of the GRE quantitative test. However, if the cognitive nature of the problem-solving process for QCD items is central, then this finding is not problematic, but rather indicates success in the construction of QCD items.

3 If one is accustomed to using the Mantel-Haenzsel delta scale for measuring DIF, then roughly, $\beta \approx -\Delta / 15$. The two scales have opposite sign conventions.

## Appendix

### Prototypical Questions for Operational Categories

All problems with a Column A and a Column B are quantitative comparison problems and have the following possible answers:

A if the quantity in Column A is greater;
B if the quantity in Column B is greater;
C if the two quantities are equal;
D if the relationship cannot be determined from the information given.

## Algebra

1. Column A Column B

$y = \frac{3x}{4}$, $x = \frac{2z}{3}$, and $z = 20$.

$y$ 11

Answer: A

2. If 25 percent of a certain number is 1,600, what is 10 percent of the number?

(A) 40
(B) 400
(C) 640
(D) 1,440
(E) 4,000

Answer: C

## Calculation Intensive

1. Column A Column B

$\frac{\sqrt{8}}{\sqrt{2}}$ $\frac{\sqrt{12}}{\sqrt{3}}$

Answer: C

2. What is the least integer value of ***n*** such that $\frac{1}{2^n} < 0.01$ ?

(A) 7
(B) 11
(C) 50
(D) 51
(E) There is no such least value

Answer: A

## Fractions With Numbers

1.

| Column A | Column B |
|---|---|
| $\frac{1}{12}$ of 17 | $\frac{1}{17}$ of 12 |

Answer: A

2. The value of $(1-\frac{5}{7})(1+\frac{3}{4})$ is

(A) $\frac{1}{28}$
(B) $\frac{3}{14}$
(C) $\frac{9}{28}$
(D) $\frac{13}{28}$
(E) $\frac{1}{2}$

Answer: E

## Pie Graph

PHYSICIANS CLASSIFIED BY CATEGORY IN 1977

Percent of Physicians by Category


100% = 421,300

1. Approximately what was the ratio of physicians in the surgical category to physicians in pathology?

(A) 10 to 1
(B) 8 to 1
(C) 7 to 1
(D) 5 to 6
(E) 4 to 5

Answer: B

2. Approximately how may more physicians were in psychiatry than in radiology?

(A) 3,000
(B) 6,300
(C) 12,600
(D) 24,800
(E) 37,000

Answer: C

## Table

CUSTOMER COMPLAINTS RECEIVED BY THE CIVIL AERONAUTICS BOARD

| Category | 1980 (percent) | 1981 (percent) |
|---|---|---|
| Flight Problems | 20.0% | 22.1% |
| Baggage | 18.3 | 21.8 |
| Customer Service | 13.1 | 11.3 |
| Oversales of Seats | 10.5 | 11.8 |
| Refund Problems | 10.1 | 8.1 |
| Fares | 6.4 | 6.0 |
| Reservations and Ticketing | 5.8 | 5.6 |
| Tours | 3.3 | 2.3 |
| Smoking | 3.2 | 2.9 |
| Advertising | 1.2 | 1.1 |
| Credit | 1.0 | 0.8 |
| Special Passengers | 0.9 | 0.9 |
| Other | 6.2 | 5.3 |
| | 100.0% | 100.0% |
| Total Number of Complaints ..... | 22,998 | 13,278 |

1. Approximately how many complaints concerning credit were received by the Civil Aeronautics Board in 1980?

   (A) 1
   (B) 23
   (C) 132
   (D) 230
   (E) 2,299

Answer: D

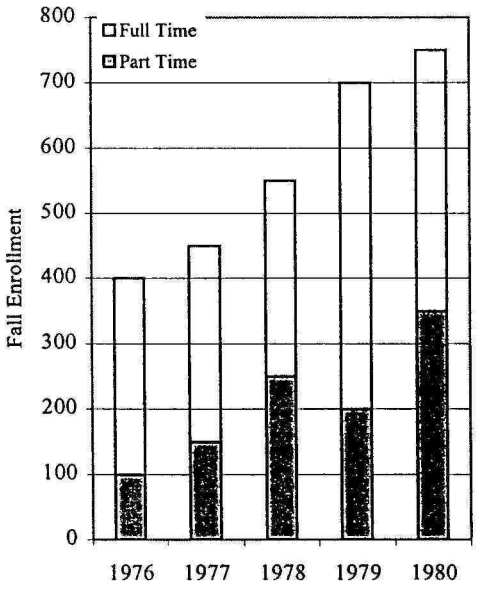
2. By approximately what percent did the total number of complaints decrease from 1980 to 1981?

   (A) 9%
   (B) 10%
   (C) 21%
   (D) 42%
   (E) 173%

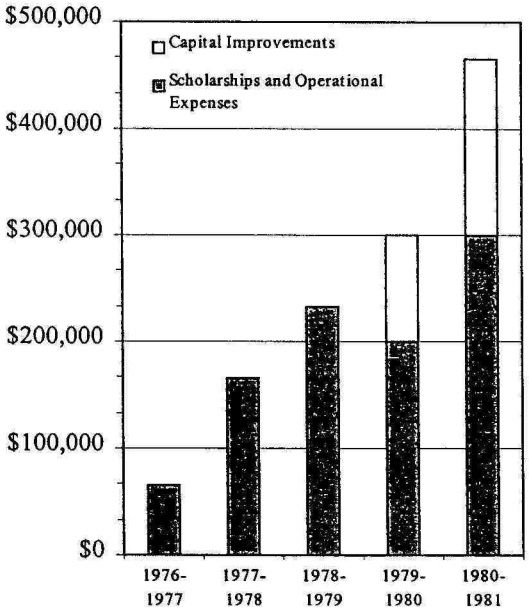
Answer: D

## Bar Graph

Total Student Enrollment




Allocation of All Contributions




1. What was the total number of students enrolled in College R in the fall of 1979?

(A) 200
(B) 250
(C) 500
(D) 650
(E) 700

Answer: E

2. By what percent did the number of part-time students enrolled increase from the fall of 1979 to the fall of 1980?

(A) 7%

(B) 42%

(C) $66\frac{2}{3}\%$

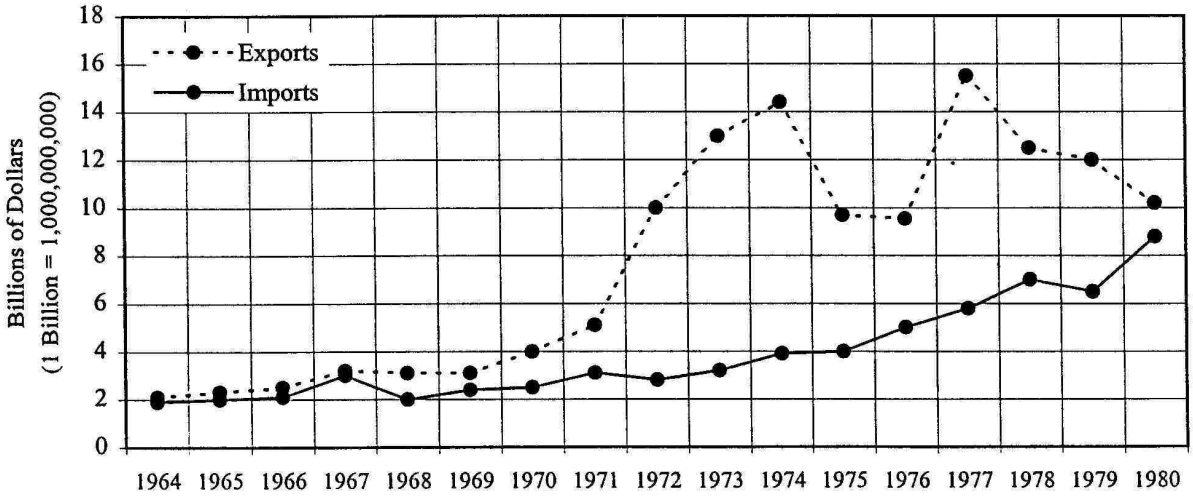
(D) 75%

(E) 80%

Answer: D

## Line Graph

FOREIGN TRADE OF COUNTRY X, 1964-1980
(in United States dollars)




1. For which year shown on the graph did exports exceed the previous year's exports by the greatest dollar amount?

   (A) 1972
   (B) 1973
   (C) 1975
   (D) 1977
   (E) 1980

Answer: D

2. Which of the following is closest to the amount, in billions of dollars by which the increase in exports from 1971 to 1972 exceeds the increase in exports from 1972 to 1973?

   (A) 1.9
   (B) 3.9
   (C) 5.0
   (D) 6.1
   (E) 8.0

Answer: A

## Number Theory

1. 

| Column A | Column B |
|---|---|
| The greatest prime factor of 15 | The greatest prime factor of 14 |

Answer: B

2. If $a$ and $b$ are both positive even integers, which of the following must be even?

I. $a^b$

II. $(a+1)^b$

III. $a^{(b+1)}$

(A) I only
(B) II only
(C) I and II only
(D) I and III only
(E) I, II, and III

Answer: D

## Probability and Statistics

1. 

| Column A | Column B |
|---|---|
| The average (arithmetic mean) of $x$, $y$, and 6 is 3. | |
| $\frac{x+y}{2}$ | $\frac{3}{2}$ |

Answer: C

2. The average (arithmetic mean) of five numbers is 25. After one of the numbers is removed, the average (arithmetic mean) of the remaining numbers is 31. What number has been removed?

(A) 1
(B) 6
(C) 11
(D) 24
(E) It cannot be determined from the information given.

Answer: A

## QCD

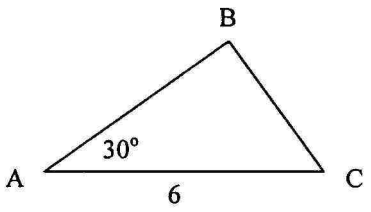
1. Column A Column B

$x^2 = 16$
$y^3 = 64$

$x$ $y$

Answer: D

2. Column A Column B




AB BC

Answer: D

## Word Conversion

1. Column A Column B

Working at constant rates, machine $R$ completely presses $x$ records in 0.5 hour and machine $S$ completely presses $x$ records in 0.75 hour ($x > 0$).

| The number of records completely pressed by $R$ in 3 hours | The number of records completely pressed by $S$ in 4 hours |
|---|---|

Answer: A

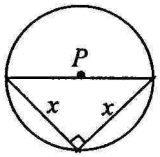
2. The price per pair of brand $X$ socks is \$2 and the price per pair of brand $Y$ socks is \$3. If there is no sales tax and a customer chooses only from among these two brands, what is the greatest number of pairs of socks that he can buy with exactly \$25?

(A) 9
(B) 10
(C) 11
(D) 12
(E) 20

Answer: D

## Applied Geometry With Algebra

1. Column A Column B




The area of the circular region with center $P$ is $16\pi$.

$x$ 4

Answer: A

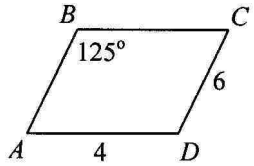
2. A rectangular floor 18 feet by 10 feet is to be completely covered with carpeting that costs $x$ dollars per square yard. In terms of $x$, how many dollars will the carpeting cost? (1 yard = 3 feet)

(A) $20x$
(B) $28x$
(C) $60x$
(D) $180x$
(E) $540x$

Answer: A

## Geometry Without Algebra

1. Column A Column B




$ABCD$ is a parallelogram.

The area of region $ABCD$ 24

Answer: D

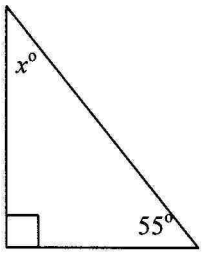
2. If a rectangular block that is 4 inches by 4 inches by 10 inches is placed inside a right circular cylinder of radius 3 inches and height 10 inches, the volume of the unoccupied portion of the cylinder is how many cubic inches?

(A) $6\pi - 16$
(B) $9\pi - 16$
(C) $160 - 30\pi$
(D) $60\pi - 160$
(E) $90\pi - 160$

Answer: E

## Geometry With Memorized Formulae

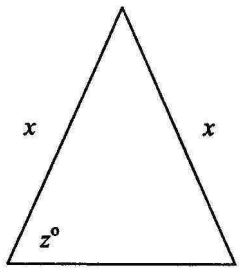
1. Column A Column B




X 35

Answer: C

2. Column A Column B




$x > y$

$z$ 60

Answer: A

04041-009051 • Y33E.2 • Printed in U.S.A.