CollegeBoard
connect to college success™

Research Report
No. 2005-11

Identifying Content and Cognitive Dimensions on the SAT®

Mark J. Gierl, Xuan Tan, and
Changjiang Wang

www.collegeboard.com

College Board Research Report No. 2005-11

Identifying Content and Cognitive Dimensions on the SAT®

Mark J. Gierl, Xuan Tan, and Changjiang Wang

The College Board, New York, 2005

Mark J. Gierl is associate professor of educational psychology and Canada research chair in educational measurement in the Centre for Research in Applied Measurement and Evaluation at the University of Alberta.

Xuan Tan is a doctoral student in the Centre for Research in Applied Measurement and Evaluation at the University of Alberta.

Changjiang Wang is a doctoral student in the Centre for Research in Applied Measurement and Evaluation at the University of Alberta.

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

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,700 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three and a half million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #050481691) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

Our colleagues in the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta—Drs. Jacqueline P. Leighton, W. Todd Rogers, and Steve Hunka—provided helpful comments throughout this study. We appreciate their help. We are also thankful to the faculty and students who shared their ideas with us in the Dimensionality Research Group. The members include Dr. Jacqueline Leighton, Ying Cui, Rebecca J. Gokiert, Xuan Tan, and Changjiang Wang. We also thank Dr. Kristen Huff of the College Board, Dr. Cindy Walker of the University of Wisconsin-Milwaukee, and Dr. Richard Luecht of the University of North Carolina-Greensboro for their constructive feedback on this study.

Contents

Tables

Figures

Preface

In a policy document titled "An Important Message to Admissions Officers About the New SAT®" (May 15, 2004), the College Board states:

> A diagnostic feature will be added to the SAT to provide information to help improve a student's academic skills. In developing the SAT, the College Board works with committees of subject-matter experts, college faculty, and high school teachers to identify the critical thinking skills in reading, mathematics, and writing needed for academic success in college. Beginning with the new SAT, students will receive feedback on how well they performed on these college success skills. This information will help both students and secondary schools focus on those critical thinking skills important for college readiness. Research is currently under way to determine the most effective methodologies and reporting formats.

Our three-year program of research (Gierl, 2004) contains a series of studies for evaluating one methodology to promote diagnostic testing with the SAT. All three years are interdependent and, ultimately, related to improving the diagnostic utility of the SAT. In year 1, we will identify the dimensions that characterize student performance on the SAT. By assessing test dimensionality we can specify the "minimum number of dimensions or statistical abilities required to fully describe all test-related differences among the examinees in a population" (Tate, 2002) or "the number of detectable constructs the test is measuring" (Douglas, Kim, Roussos, Stout, and Zhang, 1999). We used both statistically based dimensionality analyses and content-based substantive analyses to identify and interpret the cognitive dimensions measured on the mathematics and critical reading sections of the SAT. The first set of results from these analyses is presented in this report. Once these dimensions are identified and interpreted, their stability across specific subgroups of examinees (e.g., gender, race, ethnicity) will be evaluated in both mathematics and critical reading. Hence, in year 2 of the proposed research, dimensionality-based subgroup analyses will be conducted using the differential item functioning (DIF) analysis paradigm (see Gierl, Bisanz, Bisanz, and Boughton, 2003; Gierl and Khaliq, 2001; Roussos and Stout, 1996). The studies conducted in year 2 will determine if the dimensions identified in year 1 are stable across different subgroups of examinees who take the SAT. Finally, in year 3 of the proposed research, the attribute hierarchy method (AHM) for cognitive diagnosis (Leighton, Gierl, and Hunka, 2004; see also Gierl, Leighton, and Hunka, 2000) will be applied to data from the mathematics and critical reading sections to identify students' cognitive strengths and weaknesses. The AHM is a psychometric method designed explicitly to link cognitive theory and measurement practice to facilitate the development and analysis of educational and psychological tests. The AHM will be applied to SAT data to extract diagnostic information about students' cognitive skills in mathematics and critical reading. In sum, our research program is designed to enhance the diagnostic value of the SAT by providing both a diagnostic framework and the empirical results to support the use of our framework in a large-scale testing program.

Introduction

The starting point for this program of research is the identification and interpretation of cognitive dimensions that characterize student performance on the SAT. The outcomes from the dimensionality analyses will serve as the foundation for the research conducted in years 2 and 3. Unfortunately, dimensionality assessment is not a straightforward process. Rather, it entails complex analyses with numerous decisions and consequences resulting from these decisions that can lead to different answers to the question, what is the dimensional structure for this test?[1] In this report we highlight key and sometimes contentious issues related to dimensionality assessment when the goal is to *identify a dimensional structure that can promote diagnostic assessment*. We will also present the first set of results from our research on the dimensionality of the SAT.[2] The structure of this report is as follows: (a) We begin with a description of diagnostic assessment; (b) we review the outcomes from previous research on the dimensionality of the SAT; (c) we describe the data and methods used in the current study; (d) we present the results from both exploratory and confirmatory dimensionality analyses, (e) we summarize the main findings from our study and, finally, (f) we outline the next steps that will be taken to interpret the dimensional structure of the SAT.

Diagnostic Assessment and the SAT®: Preliminary Considerations

Diagnosis was defined by the influential philosopher and evaluator Michael Scriven (1991) as:

[1] Dorans and Lawrence (1999) characterize this complex decision-making process as the "relative dimensionality principle," which states that the dimensions extracted from the data depend on many factors including the number of each type of measure entered into the analysis, the metric of the analysis, the methods used for analysis, and the unit of analysis.

[2] The second set of dimensionality results will be presented in Technical Report #2. This report will be completed by June 30, 2005.

> The process of determining the nature of an affliction, of a putative symptom of disorder, or of poor performance, and/or the report resulting from the process. This may or may not happen to involve identification of the cause of the condition, but it always involves classifying the condition in terms of an accepted typology of afflictions or malfunctions, hence the terms it uses are evaluative. Diagnosis is not a primary type of evaluation; it presupposes that a true evaluation—such as the annual checkup—has already occurred, and has led to the conclusion that something is wrong. The task of diagnosis is classificatory. (p. 124)

Similarly, *cognitively* diagnostic assessment can be considered the process whereby test results are used to identify and classify examinees' cognitive skills. In fact, a defining characteristic of cognitively diagnostic assessment, according to Nichols (1994), is an explicit statement of the substantive assumptions used by test developers to construct test items and assign test scores. These substantive assumptions often specify the knowledge and skills required by examinees to solve test items. Thus, the diagnostic process is designed to identify and report the cognitively based symptoms associated with diverse test performance, thereby providing examinees with information about their problem-solving strengths and weaknesses.

Subscore analyses can guide the diagnostic process when these subscores have an interpretable dimensional structure. For example, if a test is developed to measure diverse cognitive skills across numerous content areas, then the test specifications imply a multidimensional structure with different cognitive skills and content areas reflecting different dimensions measured in the domain of interest. In this case, the total score could be viewed as a composite measure of the cognitive skills and content areas whereas the subscores could represent dimensionally homogeneous measures of the specific cognitive skill and content area combinations, as outlined in the test specifications.

Richard E. Snow,[3] for one, claimed that psychologically meaningful and useful subscores could be obtained from conventional achievement tests. Moreover, he argued these subscores represented important ability *dimensions* that would show different patterns of relationships among demographic, cognitive, and affective variables. To further this view, Snow developed and began applying a *multidimensional approach* to achievement test validation in his later work (e.g., Hamilton, Nussbaum, Kupermintz, Kerkhoven, and Snow, 1995; Kupermintz, Ennis, Hamilton, Talbert, and Snow, 1995; Kupermintz and Snow, 1997; Nussbaum, Hamilton, and Snow, 1997; see also special issue of *Educational Assessment*, Vol. 8, No. 2, 2002).

Whether used during the test development or analysis stage, a multidimensional approach to test score validation is compelling because it suggests that psychologically complex constructs can be found in educational and psychological tests. These constructs are identified using contemporary dimensionality procedures and reported as subscores. By implication, a multidimensional approach also suggests that educational and psychological tests can have *diagnostic value* because reliable and valid information about examinees' cognitive strengths and weaknesses can be obtained from their performance on well-defined, dimensionally distinct subtests (*Standards for Educational and Psychological Testing*, 1999; Tate, 2002, 2004). The results can then be reported as subscores (Goodman and Hambleton, 2004; *Standards for Educational and Psychological Testing*, 1999). These subscores can help direct remedial efforts, particularly when examinees do poorly on the exam, by highlighting specific areas where improvements are needed.

Previous Research on the Dimensionality of the SAT

It is useful to evaluate the existing literature on the dimensionality of the SAT because results for the previous version of the SAT may inform the current SAT, given that the two tests are closely related. This literature review is particularly important given the widely held belief that the SAT measures a *unidimensional construct*.[4] Only three studies on the dimensionality of the SAT were found in our review of the literature, and only one of these studies was published. The three studies include:

1. Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics*, *13*, 19–43.
2. Diones, R., Bejar, I. I., & Chaffin, R. (1996, January). The dimensionality of responses to SAT analogy items. *ETS Research Report*. Princeton, NJ: ETS.
3. Lawrence, I. M., & Dorans, N. J. (1987, April). *An assessment of the dimensionality of SAT-Mathematical.*

[3] Richard Snow was the Howard H. and Jessie T. Watkins University Professor Emeritus of Education at Stanford University. He died in 1997. Snow was one of the early advocates for combining cognitive principles with measurement practices. His 1989 chapter in *Educational Measurement, 3rd Edition*, coauthored with David Lohman, titled "Implications of Cognitive Psychology for Educational Measurement" is a seminal work in the disciplines of both cognitive psychology and educational measurement.

[4] The perception that the current SAT measures a unidimensional contruct became apparent to us when Dr. Joan Herman, discussant at a NCME 2004 session in San Diego organized by Dr. Kristen Huff called "Connecting Curriculum and Assessment Through Meaningful Score Reports," claimed that subscale reporting was not feasible because it was well-known that the SAT is unidimensional. However, Dr. Herman did not present any empirical evidence to support her claim about this important property of the SAT.

Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

In the first study, Cook et al. (1988) used confirmatory factor analysis to assess the dimensionality of the SAT verbal (SAT-V) section. Three different forms of the SAT-V were compared. The results from the factor analysis revealed that the SAT-V was "slightly multidimensional"[5] and that the three forms of the SAT-V were not strictly parallel (which, again, suggests that different constructs may be measured across forms). The authors concluded by calling for additional dimensionality studies on the SAT because this type of research "…might yield diagnostics that could be used to arrive at more informed psychometric decisions about test specifications, and about the equating and scoring of the SAT" (p. 40).

In the second study, Diones et al. (1996) evaluated two item clusters—intensional and pragmatic—believed to underlie student performance on the analogy items from the SAT. The intensional-pragmatic dichotomy was originally used by Bejar, Chaffin, and Embretson (1991) and Chaffin and Pierce (1987) to describe the type of knowledge required to solve items on the GRE. The purpose of this study was to replicate Bejar et al. (1991) in order to assess whether the analogy items on the SAT would also produce a two-dimensional structure. Both confirmatory and exploratory factor analysis was used. For the confirmatory factor analysis, the intensional-pragmatic bidimensionality was not found for the specified item clusters. Rather, the test remained "unremittingly unidimensional." However, using exploratory factor analysis, different item clusters were produced, resulting in a two-factor solution. The outcomes from the confirmatory and exploratory analysis caused the authors to conclude "…the [analogy] test was not in all forms unidimensional" (p. 16).

In the third study, Lawrence and Dorans (1987) used exploratory and confirmatory factor analysis to assess the dimensionality of the SAT math section (SAT-M) across six forms. Confirmatory analyses of item parcels "strongly indicate that the SAT-Mathematical Test is unidimensional" (p. 22). Exploratory analyses of the item-level data, on the other hand, revealed a "slight departure from unidimensionality" attributable to the geometry items.

Taken together, the results from these three studies are diffuse, especially as they pertain to the question, what is the dimensionality of the SAT? It appears that the SAT verbal section measures a multidimensional construct whereas the SAT math section measures a unidimensional construct. However, this conclusion is far from certain for a number of reasons. First, only a small number of studies have been conducted to evaluate the dimensionality of the SAT. Given the tentative conclusions presented by the authors in these studies, far more research is needed to evaluate the dimensionality of the SAT. Second, the dimensionality studies cited draw on one procedure—factor analysis. No studies were found that use a multimethod approach or that implement the procedures associated with recent developments in nonparametric dimensionality assessment such as DIMTEST, DETECT, or HCA (Douglas, Kim, Roussos, Stout, and Zhang, 1999; Stout, Habing, Douglas, Kim, Roussos, and Zhang, 1996). Third, the studies cited were not designed to evaluate the diagnostic potential of the SAT. Rather, the dimensionality analyses were content-based (particularly the first and third studies) and, as a result, the outcomes from these studies shed no light on the dimensional structure of the SAT, as it might relate to diagnostic assessment.[6] From our review of the literature, we conclude that it is not accurate to assume that the SAT measures a well-defined unidimensional construct—far more empirical evidence is required to substantiate this claim.

Methods

In this section, we describe the SAT test items, the 2003 Field Trial used to collect the SAT student-response data, and the dimensionality analyses used in the current study.

SAT Test Items

The SAT is a standardized test designed to measure college readiness. Both critical thinking and reasoning skills are evaluated. The test contains three sections: mathematics, critical reading, and writing. However, only the dimensionality in the first two sections will be evaluated in our study.

The math section contains 54 items administered in two 25-minute sections and one 20-minute section. For these items, students are expected to solve unfamiliar problems using key mathematical concepts in the areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. Multiple-choice and constructed-response item formats are used, but the items for both formats are scored dichotomously.

The critical reading section contains 67 items also administered in two 25-minute sections and one 20-minute section. For these items, students are expected to draw inferences from text, synthesize information, distinguish between main and supporting ideas, understand word meaning, follow the logic of an argument, and recognize genres. Students solve sentence-completion items in addition to critical reading items associated with short and long reading passages using content drawn from natural

[5] Using data from the verbal section of the SAT to illustrate the distinction between item-level and test-level dimensionality, Dorans and Lawrence (1999) also concluded that the SAT-Verbal was multidimensional at both the micro- and macro-level of analysis (see pp. 21–32).

[6] This point is particularly important in light of the fact that any model of the "correct" dimensionality is not unique, meaning that other statistical models can also be found that fit the data. As a result, Tate (2002), among others, suggests that substantive considerations guide the selection of the final statistical model, given that different models can be specified that fit the data.

sciences, social studies, literary fiction, and humanities. All items are multiple choice and, therefore, scored dichotomously.

2003 Field Trial Sample Data

Data from Design 1 of the SAT and new PSAT/NMSQT® (Preliminary SAT/National Merit Scholarship Qualifying Test) spring 2003 field trial (Liu, Feigenbaum, and Walker, 2004) were used in the current study. Design 1 was the major component of the field trial in which the content and the statistical properties for the SAT were evaluated. This design included 13 booklets containing different combinations of SAT and PSAT/NMSQT items spiraled within classroom to achieve comparable groups and appropriate sample sizes for all follow-up analyses (see Table 1 in Liu, Feigenbaum, and Walker, 2004). The examinees in the sample were primarily high school juniors who attended schools that volunteered to participate in the field trial. The field trial sample was deemed to be similar to but not entirely representative of the baseline cohort of college-bound seniors who took the SAT in 2002. Nevertheless, the results from comprehensive analyses allowed Liu, Feigenbaum, and Walker (2004) to conclude that the student sample who completed the field test items would allow researchers to adequately evaluate important psychometric issues on the SAT, including the dimensionality of the test.

Data from three different books were used in the current study: Books 2a, 2c, and 5.[7] These books were used for three reasons. First, the use of multiple forms allowed us to initially test our hypotheses and models and then cross-validate the results to ensure the outcomes were stable and consistent across samples. As a result, Book 2a was designated the primary sample and Books 2c and 5 were seen as the cross-validation samples. All preliminary analyses were conducted using the Book 2a data. Second, the field test design promoted comparisons across forms because only student samples differed across the books (i.e., all books contained the same test items). As a result, the stability of the item characteristics and the dimensions was easily assessed. Third, the sample sizes in Books 2a and 2c were relatively large and amenable to dimensionality analyses because these samples were also used in the major equating study (Liu, Feigenbaum, and Walker, 2004, p. 8). Book 5 was selected as a second cross-validation sample because these data were used by ETS and the College Board in many of their preliminary SAT analyses (Dr. Kristen Huff, personal communication, June 29, 2004). The content and the sample size for these three test booklets are shown in Table 1. The comparability of the test booklets is apparent in this table: Although the three forms differed in sections 5 and 9, the positions of the forms containing the data used in the current study—mathematics and critical reading—were identical across forms. The sample sizes across these three booklets were also adequate, exceeding 2,000 examinees in each form.

Description of Dimensionality Assessment Procedures

Embretson and Reise (2000), in their recent review and critique of dimensionality assessment in educational and psychological testing, claimed that:

> "…researchers should now be starting to move away from reporting heuristic indices such as 'variance accounted for by the first factor' or 'ratio of first to second eigenvalue' and start implementing the new procedures that tackle these issues in a much more sophisticated manner. [Specifically, as described in this chapter] we recommend more application of Stout's procedure for determining essential unidimensionality and/or applications of appropriate techniques such as those found in TESTFACT, POLYFACT, NOHARM, and LISCOMP." (p. 245)

We acted on this suggestion by conducting extensive parametric and nonparametric dimensionality analyses using data from the mathematics and critical reading sections of the 2003 field trial. The results from this study are designed to provide a point of comparison

Table 1

Field Trial Design 1 Test Book Summary

		Section in Book								
Book	***Sample***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
2a	2,443	*SAT Essay*	SAT Math	SAT Reading	SAT Math	*PSAT/NMSQT Reading*	SAT Reading	SAT Math	SAT Reading	*SAT Writing*
2c	2,336	*SAT Essay*	SAT Math	SAT Reading	SAT Math	*PSAT/NMSQT Math*	SAT Reading	SAT Math	SAT Reading	*SAT Writing*
5	2,202	*SAT Essay*	SAT Math	SAT Reading	SAT Math	*SAT Writing*	SAT Reading	SAT Math	SAT Reading	*Pretest*

Note: The data from sections 1, 5, and 9, highlighted in italics, were not used in the current study.

[7] Book 2a was also denoted as TD Form Code 3ZNN11 and Systems Form Code 111 in the nSAT Field Test documentation; Book 2c was coded as TD Form Code 3ZNN13 and Systems Form Code 113; and Book 5 was specified as TD Form Code 3ZNN18 and Systems Form Code 118.

with results from existing SAT dimensionality analyses and to draw on the potential benefits associated with recent developments in dimensionality assessment. Three dimensional procedures are used extensively in our study: DIMTEST, DETECT, and nonlinear factor analysis.

DIMTEST Overview

DIMTEST is a nonparametric statistical procedure that conducts a hypothesis test to assess the presence of multidimensionality (Froelich, 2000; Froelich and Habing, 2001). This procedure is based on Stout's (1987) concept of "essential unidimensionality," which holds when only one dominant dimension influences the examinees' performance on a set of test items (Hattie, Krakowski, Rogers, and Swaminathan, 1996; Nandakumar, 1991; Nandakumar and Stout, 1993; Stout, Habing, Douglas, Kim, Roussos, and Zhang, 1996). DIMTEST is used to test the hypothesis, H_0:d=1 versus H_1:d>1, where d is the number of dimensions. DIMTEST has undergone two major revisions since it was introduced. The first revision was undertaken by Nandakumar and Stout (1993) and the second by Froelich (2000). In the current study, the most recent version of DIMTEST is used and, as a result, the most recent version is described in our overview.

DIMTEST is based on the idea that if a test is unidimensional then the conditional covariance between any two items on the assessment subtest (AT) is zero after conditioning on the partitioning subtest (PT), or $E[Cov(X_{i1,}\ X_{i2} \mid \theta_{PT})] = 0$. If, on the other hand, a test is multidimensional then the conditional covariance between any two items on the AT is greater than zero after conditioning on the PT, or $E[Cov(X_{i1}\ X_{i2} \mid \theta_{PT})] > 0$ (Zhang and Stout, 1999). Therefore, testing whether the conditional covariance is zero is analogous to testing the assumption of weak local independence (Stout, 1987). Conceptually, this test means that after conditioning on scores associated with the PT, which is the total score for those items that primarily measure θ_1, then the expected value of the covariance for those items that compose the AT, which is the total score for those items that primarily measure θ_2, will be zero if the AT and the PT are measuring the same dimension. This outcome occurs because any dependency between AT item responses is removed by conditioning on the PT, if the test structure is truly unidimensional.

The conditional covariances are estimated by finding the difference between the total test variability and item variability for examinees with the same score, k, on the PT because total test variability (σ^2_x) can be written as a sum:

$$\sigma^2_X = \sum_{i-1}^{N_i} p_i q_i + 2\sum_{i<l} Cov(U_i, U_l)\,.$$

From this expression, the terms can be rearranged to produce the covariance, which, in turn, can be written as the difference:

$$\frac{\sigma^2_X - \sum_{i=1}^{N_1} p_i q_i}{2} = \sum_{i<l} Cov(U_i, U_l)\,.$$

Conditional covariance is obtained by calculating the covariance for examinees with the same score, k, on the PT items, which is expressed as:

$$T_{L,k} = \frac{\sigma^2_{X,k} - \sum_{i=1}^{N_1} p_{i,k}\,, q_{i,k}}{2} = \sum_{i<l,k} Cov(U_{i,k}, U_{l,k})\,.$$

When these three expressions are considered together, the DIMTEST test statistic can then be defined as:

$$T = \frac{T_L - \overline{T}_B}{\sqrt{2}}\,,$$

where the value of T_L in the numerator of the test statistic is based on the sum of the estimated conditional covariances between the AT items for examinees that have obtained the same score, k, on the PT items. Specifically:

$$T_L = \frac{\sum_{k=1}^{K} T_{L,k}}{\sqrt{\sum_{k=1}^{K} S_k^2}}\,,$$

where k represents the subgroup of examinees with the same score on the PT items, $T_{L,k}$ is the conditional covariance, and S_k^2 is the asymptotic variance of $T_{L,k}$. The second term in the numerator of the test statistic is a correction for the bias that is known to exist for a test of finite length when test data are unidimensional. The test statistic, T, is known to have an asymptotic normal distribution as the number of examinees and items approach infinity (Froelich, 2000).

In earlier versions of DIMTEST it was necessary to identify another set of items, distinct from the AT and the PT, to correct for this bias. This set of items needed to be similar in item difficulty to the AT, and was referred to as AT2. Typically, items for AT2 were hard to identify, especially when the total number of items on the test was relatively small. This limitation was eliminated recently by Froelich (2000). Although the bias correction is still necessary, the new version of DIMTEST makes use of a nonparametric IRT bootstrap method to correct for the bias. The bootstrap method works by estimating the item and the ability parameters under the assumption that the test measures a unidimensional composite. Then, examinee responses to all test items are simulated using the estimated unidimensional item-response functions. A test statistic, comparable to T_L, is computed using the simulated data, denoted T_B. To reduce the random variation in T_B, this procedure is repeated a number of times and the average, $\overline{T_B}$, is used as the bias correction.

Recall, however, that dimensionality assessment can be accomplished using either an exploratory or a confirmatory approach with DIMTEST. An exploratory approach is ultimately data driven, meaning that the dimensional structure is first produced from the test data and then interpreted. Exploratory DIMTEST identifies the AT using the items with the highest factor loadings from an unrotated principal-axes factor analysis of the tetrachoric correlation matrix. This approach was used for the exploratory DIMTEST analyses reported in our study. A confirmatory approach can also be conducted where substantive hypotheses guide the formation of both the AT and the PT. With a confirmatory approach, different dimensional structures are first specified based on substantive considerations about the test data and then hypotheses are tested statistically. In the current study, the College Board test specifications and skill categories were used for the confirmatory DIMTEST analyses.

DETECT Overview

DETECT is also a nonparametric dimensionality assessment procedure, but this approach is designed to determine the multidimensional structure underlying test data (Zhang and Stout, 1999). DETECT identifies mutually exclusive, dimensionally homogeneous clusters of items using a genetic algorithm. Because the clusters of items are mutually exclusive, this procedure is most useful when approximate simple structure[8] prevails in the test data. Unlike DIMTEST, DETECT can only be conducted in exploratory mode.

To specify these clusters, DETECT attempts to maximize the value of the DETECT index, *D*(*P*). This index quantifies the degree of multidimensionality present in *P*. The DETECT index is created by computing all item covariances after conditioning on the examinees' scores using the remaining items. That is,

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le N} \delta_{ij} E[Cov(X_i, X_j \,|\, \Theta_{TT} = \theta)],$$

where *n* is the number of dichotomous items on a test, *P* denotes the partitioning of *n* items into *k* clusters, Θ_{TT} is the test composite, X_i and X_j are scores on items *i* and *j*, and

$$\delta_{ij} = \begin{cases} 1 & \text{if items i and j are in} \\ & \text{the same cluster of P,} \\ -1 & \text{otherwise.} \end{cases}$$

Although many different partitions can exist in a set of test data, *P** serves as the partition that maximizes *D(P)* [herein denoted as *D(P*)* and also called D_{Max} in the literature]. For instance, when the data are truly unidimensional, clusters of items will be found that are not particularly homogeneous. In this case, the within-cluster conditional covariance will be positive for some pairs of items and negative for other pairs of items resulting in a *D(P*)* index that is close to zero. If, however, the underlying structure of the data is truly multidimensional, then clusters of items will be found that have positive within-cluster conditional covariances and negative between-cluster conditional covariances, resulting in a *D(P*)* index that is greater than zero. Based on results from simulation studies, Kim (1994) suggested that when the *D(P*)* index is less than 0.10, the data can be considered unidimensional; an index between 0.10 and 0.50 can be considered a weak amount of dimensionality; an index between 0.51 and 1.00 can be considered a moderate amount of dimensionality; and an index greater than 1.00 can be considered a strong amount of dimensionality.

Another index that is often reported with *D(P*)* is r_{Max}. To determine if the partition *P**, which produced *D(P*)*, is, in fact, the correct partition to produce a simple structure solution, the following ratio can be computed:

where , $$r_{Max} = \frac{D(P*)}{\tilde{D}(P*)}$$

$$\tilde{D}(P*) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le N} \left| \; E[Cov(X_i, X_j \,|\, \Theta_{TT} = \theta)] \; \right|.$$

In other words, r_{Max} is an index representing how well the data approximates simple structure by comparing the maximum value of the partition to the average of the absolute value for the conditional covariance across all item combinations. Values of *r* greater than 0.80 suggest that the data displays approximate simple structure. Conversely, values of *r* less than 0.80 suggest that the data display complex structure (Kim, 1994). Simulation studies conducted by Kim (1994) and Zhang and Stout (1999) demonstrate that DETECT accurately identifies the correct partition when *r* is greater than 0.80, meaning the data possess a simple structure solution. Unfortunately, comprehensive studies have not been conducted to evaluate the accuracy of the DETECT partition when *r* is less than 0.80—in fact, the accuracy of the DETECT partition may be relatively poor when *r* is less that 0.80. This point was alluded to when Zhang and Stout (1999) claimed:

> It is very important to note that DETECT is still informative when approximate simple structure fails to hold. In particular, it can still locate relatively

[8] For a two-dimensional test, when all items lie along the two dimensional coordinate axes, the test displays *simple structure*. A complete description of the five conditions required to achieve Thurstone's definition of simple structure is presented in McDonald (1999, pp. 179–180). Items can also lie in a narrow sector around the two-dimensional coordinate axes. In this case, the test displays *approximate simple structure*. When items lie in space throughout the two-dimensional coordinate axes (i.e., items measure a range of skills in the $\Theta_1 \Theta_2$ composite) the test displays *complex structure*.

> dimensionally homogeneous clusters; *however, there is no longer a unique 'best' or 'correct' partition to be found by DETECT because there will be little to no separation between some of the clusters found* [italics added]. (p. 215)

This lack of separation that results in a nonunique partition implies that the clusters produced across samples may be highly unreliable and, therefore, difficult to interpret because the clusters are not replicable across samples. This outcome is problematic because unreliable statistical results that are difficult to interpret will not lead to an increased understanding of the multidimensional structure underlying the test data.

Nonlinear Factor-Analysis Overview

A nonlinear factor-analytic approach can also be used to model multidimensional data (McDonald, 1967, 1997, 1999). The common factor parameterization for the factor-analytic model is defined as:

$y_i = \lambda_{i1}\theta_1 + \ldots\lambda_{ik}\theta_k + \delta_i$,

where y_i is conceptualized as a continuous latent response propensity meaning that for each item score there exists an underlying item-specific threshold that corresponds to the difficulty level of the item where the examinee must exceed this threshold to get the item correct; $\lambda_i = [\lambda_1, \lambda_2, \ldots, \lambda_k]$ is the factor-loading vector; $\theta = (\theta_1, \theta_2, \ldots\theta_k)$ is the examinee trait vector having mean 0 and covariance Φ; and δ_i is a residual term distributed $(0, \Psi_i)$. The model assumes for each item i there is a latent ability that is required to correctly answer the item. This latent ability is assumed to be continuous and normally distributed. Taken together, each item score is determined by the location of y_i relative to a fixed item threshold of τ_i such that an examinee's response, U_i, can be expressed as:

$U_i = 1$ if $y_i \geq \tau_i$ and $U_i = 0$ if $y_i < \tau_i$.

The proportion of examinees correctly responding to item i (i.e., the p-value or difficulty level) can be expressed as the proportion of area under a normal curve beyond the threshold τ_i as $p_i = N(\tau_i)$, where N denotes the normal ogive function.

The latent trait parameterization of the k-dimensional normal ogive model[9] can also be formulated and linked, conceptually, to the common factor parameterization, as follows:

$$P\{U_i = 1 | \theta_1 .. \theta_k\} = N\{\beta_{i0} + \beta_i^{/}\theta\} = N\{\beta_{i0} + \beta_{i1}\theta_1 + \beta_{i2}\theta_2 + \ldots + \beta_{ik}\theta_k\}’$$

where, $$\beta_{i0} = \frac{\tau_i}{\sqrt{\psi_i}},$$

and, $$\beta_i = \frac{\lambda_i}{\sqrt{\psi_i}}.$$

In these three equations, N is the normal curve, θ is the latent ability, β_{i0} is the difficulty parameter, β_i is the discrimination parameter, ψ_i is the unexplained item variance or 1 minus the communality, given as:

$\Psi_i = 1 - \lambda_i^{/}\Phi\,\lambda_i$,

and, Φ is the covariance matrix of latent abilities.

One advantage of using the common factor parameterization of the multidimensional model is that it allows one to estimate the correlation between factors that underlie test performance. For the two-factor solution, where θ_1 and θ_2 are the two latent abilities, $r_{(\theta_1, \theta_2)} = \phi_{12}$. Then, the underlying composite is scaled so that Ψ_i can be expressed as:

$$\psi_i = \sqrt{1 - \lambda_1^2 - \lambda_2^2 - 2\lambda_1\lambda_2\phi_{12}}.$$

Although other computer programs are available for estimating multidimensional item parameters for dichotomously scored data, NOHARM is frequently used. NOHARM is the acronym for the *normal ogive by harmonic analysis robust method*. The program was written by Fraser (1988) to fit the unidimensional and multidimensional normal ogive models of latent trait theory, as presented by McDonald (1967). This program uses a nonlinear factor analytic approach to estimate item parameters in either an exploratory or confirmatory mode. If the underlying dimensional structure is unclear, then the exploratory mode of NOHARM would be used. If a particular dimensional structure is hypothesized, then the confirmatory mode should be used.

NOHARM estimates the latent trait parameters for the k-dimensional normal ogive model using a two-step approach. First, the threshold parameters, β_{i0}, are estimated using a closed form expression by solving the sample analog of k-dimensional normal ogive model. Second, the discrimination parameters, β_i, are estimated using unweighted least squares (ULS) by minimizing the expression

$$q = \sum_{i \neq j}(p_{ij} - \pi_{ij}^r)^2,$$

where π^r_{ij} is the r^{th}-term approximation of a normalized Hermite-Tchebycheff polynomial estimation of the proportion answering items i and j correctly, using a quasi-Newton algorithm. The main advantage of using a ULS procedure is that it can estimate the parameters for a large number of items because matrix inversion is not required. The main disadvantage of using a ULS procedure is that there are no direct standard errors for the parameter

[9] Some researchers also represent the latent trait parameterization of the multidimensional model as: $P\{U_i = 1 | \theta_1, \theta_2\} = \dfrac{1}{1 + e^{-1.7(a_1\theta_1 + a_2\theta_2 + d)}}$, where a_1 corresponds to λ_1, the discrimination parameter for the θ_1 trait, a_2 corresponds to λ_2, the discrimination parameter for the θ_2 trait, d corresponds to a *location* parameter, and 1.7 is the scaling factor to make the logistic and normal ogive models equivalent.

estimates and, as a result, few indices for assessing the goodness of fit between the model and the data.[10]

To summarize, a multidimensional data structure can be identified and modeled using a nonlinear factor analytic approach. The model can be presented as either the common factor or the latent trait parameterization. The main advantage in using the common factor parameterization stems from the interpretative conventions that can be adopted from factor analysis to interpret the multidimensional solutions. For example, the λs estimated by NOHARM can be interpreted as factor loadings, which represent the correlation between the item-response propensity and the corresponding dimension. The ψs can be interpreted as the item uniqueness, which represents the proportion of variance in the ys not accounted for by the dimensions. The τs can be interpreted as the inverse normal transformation of the item difficulty level, meaning $\tau_i = N^{-1}(p_i)$. Thus, positive τ_is represent easy items and negative τ_is represent difficult items. We used the common factor parameterization, along with conventions used commonly with factor analysis (Preacher and MacCallum, 2003), to identify and interpret dimensions underlying the student response data for the SAT mathematics and critical reading items.

Results

Overview

Dimensionality analyses can be conducted using either an exploratory or a confirmatory approach. With an exploratory approach, analyses are conducted to identify the optimal number of dimensions that can be extracted from the data. An exploratory analysis is often conducted when few hypotheses or substantive explanations are available to describe the underlying structure of the data. The outcomes from an exploratory analysis are expected to yield information and, possibly, insights about the relationship of the items to the dimensions. Unfortunately, this relationship is often difficult to interpret because the analysis lacks a substantive framework to guide the interpretative process, which is a major disadvantage of an exploratory approach. We begin our study with a series of exploratory analyses as a first attempt at identifying the dimensional structure of the SAT. The exploratory results are reported in Part #1 of the Results section.

With a confirmatory approach, constraints are placed on the data to evaluate specific models. Then, fit statistics are used to evaluate the residual covariance, after modeling the data, to see if any systematic variation among the variables remains. Often, the pattern of the parameters relative to the dimensions is specified because hypotheses or substantive explanations are available to describe the structure of the data. Thus, the purpose of the analysis is to test these hypotheses or substantive explanations. The advantage of this approach stems from the logic of hypothesis testing where a *substantive* explanation is used to generate hypotheses and then a *statistical* analysis is used to test the hypotheses. Each confirmatory analysis, therefore, provides a test of the proposed hypotheses (Ackerman, Gierl, and Walker, 2003; Walker and Gierl, 2004). But this important advantage comes at a price—an articulate substantive framework must be available to describe the dimensional structure of the data and guide the analysis. Because diagnostic assessments are designed to identify and report the cognitively based symptoms associated with diverse test performance, the dimensions we identify must contain items that can be linked directly and systematically to these cognitively based symptoms. Thus, two different substantive frameworks were used to guide the confirmatory dimensionality analyses: the College Board test specifications and skill categories. The confirmatory results are reported in Part #2 of the Results section.

Part #1: Exploratory Dimensionality Results

Applying exploratory DIMTEST to the mathematics and critical reading composite (i.e., combining the items across sections) resulted in rejections[11] with small p-values across all three samples, as shown in Table 2 (under the heading "Composite"). This outcome indicates the

Table 2

Exploratory DIMTEST Results for Composite, Mathematics, and Critical Reading Sections

	Composite		***Mathematics***		***Critical Reading***	
Book	***T***	***p***	***T***	***p***	***T***	***p***
2a	11.3583	0.0000	4.4371	0.0000	3.8565	0.0001
2c	12.3312	0.0000	3.7953	0.0001	3.2162	0.0006
5	11.2085	0.0000	2.4738	0.0067	4.3285	0.0000

[10] Despite this important disadvantage, we conclude that nonlinear factor analysis, as implemented with the computer program NOHARM, is the best dimensionality procedure for our SAT analyses because we want parameter estimates for a large number of items. Alternative estimation procedures, such as maximum likelihood or generalized least squares, are used in computer programs such as LISREL and MPlus. These alternative procedures yield direct standard errors along with a host of goodness-of-fit indices. However, these alternative procedures also require inverting large data matrices and seldom work well with more than 20 test items.

[11] A conventional alpha level of 0.05 was used to interpret the results in this study. In some cases, however, we interpret a result as statistically significant when the observed alpha level is close to critical alpha level (i.e., a + 0.015) because we are attempting to find systematic patterns across samples at this early stage in our research program. An outcome is considered systematic if we can replicate the result in at least two samples.

data are not unidimensional. The composite provides an important baseline for the first exploratory analysis because the mathematics and critical reading items, together, clearly measure a multidimensional construct. This multidimensional construct can be identified with exploratory DIMTEST.

When DETECT was applied to the data (see Table 3 under the heading “Composite”), the $D(P^{*})$ index (herein called DETECT$_{Max}$) ranged from 0.4062 to 0.4178, indicating a weak amount of dimensionality, according to Kim’s (1994) classification. However, the r$_{Max}$ index ranged from 0.8069 to 0.8207, indicating that the data, while displaying a “weak” amount of dimensionality, still displayed approximate simple structure for the two test sections across all three samples. In fact, when the exploratory DETECT clusters for mathematics and critical reading were produced, the items were partitioned *perfectly* by section, meaning that the mathematics items were separated from the critical reading items for two of the three samples (the third sample contained only minor item misclassifications).

Next, exploratory DIMTEST analyses were conducted separately using items from the mathematics and critical reading sections. For mathematics, DIMTEST rejected the null hypothesis producing small *p*-values across all three samples, indicating the data are not unidimensional (see Table 2 under “Mathematics”). When DETECT was applied to the data (see Table 3 under “Mathematics”), DETECT$_{Max}$ ranged from 0.1346 to 0.1419, indicating a weak amount of dimensionality. The r_{Max} index was also relatively low, ranging from 0.4238 to 0.4444, indicating the data displayed complex structure across all three samples. The number of DETECT clusters in mathematics ranged from 4 to 5 across the three samples.

For critical reading, DIMTEST again rejected the null hypothesis producing small *p*-values across all three samples, indicating the data are not that unidimensional (see Table 2 under “Critical Reading”). When DETECT was applied to the data (see Table 3 under “Critical Reading”), DETECT$_{Max}$ ranged from 0.1918 to 0.1951, indicating a weak amount of dimensionality. The r_{Max} index was also relatively low, ranging from 0.4788 to 0.4997, indicating that the data displayed complex structure across all three samples. The number of DETECT clusters in critical reading ranged from 4 to 5 across the three samples.

Table 3

DETECT Results for Composite, Mathematics, and Critical Reading Sections

	Composite		*Mathematics*		*Critical Reading*	
Book	*DETECT$_{Max}$*	*r Index*	*DETECT$_{Max}$*	*r Index*	*DETECT$_{Max}$*	*r Index*
2a	0.4178 (2)	0.8192	0.1347 (4)	0.4322	0.1940 (4)	0.4844
2c	0.4146 (2)	0.8207	0.1419 (5)	0.4444	0.1951 (4)	0.4997
5	0.4062 (2)	0.8069	0.1346 (4)	0.4238	0.1918 (5)	0.4788

Note: The number of clusters is in the DETECT$_{Max}$ columns in parentheses.

To further interpret the DETECT clusters, we adopted the analysis algorithm proposed recently by Nandakumar and Ackerman (2004) for modeling test data. They describe their algorithm, as follows:

Step 1. Use DIMTEST to determine if dimensionality, *d*, underlying test data is essentially 1.
Step 2. If *d*=1, then fit a unidimensional model to the data. Stop.
Step 3. If *d*>1, then investigate if test items can be decomposed into unidimensional clusters using DETECT.
Step 4. Test each DETECT cluster using DIMTEST to determine if *d*=1.
Step 5. Combine clusters, if necessary, based on expert opinion of item content of the AT of DIMTEST. Again, test hypothesis *d*=1.
Step 6. If *d*=1, go to step 2. If *d*>1 for any of the clusters, either delete the items from the test or use the multidimensional model.

We iterated through these steps until we identified dimensionally homogeneous item clusters in mathematics and critical reading across each sample. That is, DIMTEST was first used to evaluate the dimensionality of each cluster and, if the cluster was shown to possess multidimensionality, DETECT was applied again to the data to produce a smaller cluster. This cycle was continued until either the DIMTEST result was not statistically significant or the item cluster contained four items or less (Nandakumar and Ackerman, 2004).

The results for mathematics are displayed in Figures 1 to 3. Important differences in the final cluster solutions were found across the samples. Using Book 2a, 24 independent clusters were identified in three levels (meaning that the number of levels that contain at least 1 independent cluster is three—the number of levels is presented in the left column of each figure). The smallest cluster contained 1 item and the largest cluster contained 8 items. Using Book 2c, 17 independent clusters were identified in three levels. The smallest cluster contained 1 item and the largest cluster contained 18 items. Using Book 5, 17 independent clusters were identified in three levels. The smallest cluster contained 1 item and the largest cluster contained 7 items. These results reveal the final independent clusters across forms were not similar and, as a result, systematic patterns were not apparent. Thus, these results would be difficult, if not impossible, to interpret because the clusters are inconsistent across the three mathematics forms. This outcome is problematic because unreliable statistical results that are difficult to interpret will not lead to an increased understanding of the multidimensional structure underlying the test data.

Similar results were found for the critical reading items, as displayed in Figures 4 to 6. Using Book 2a,

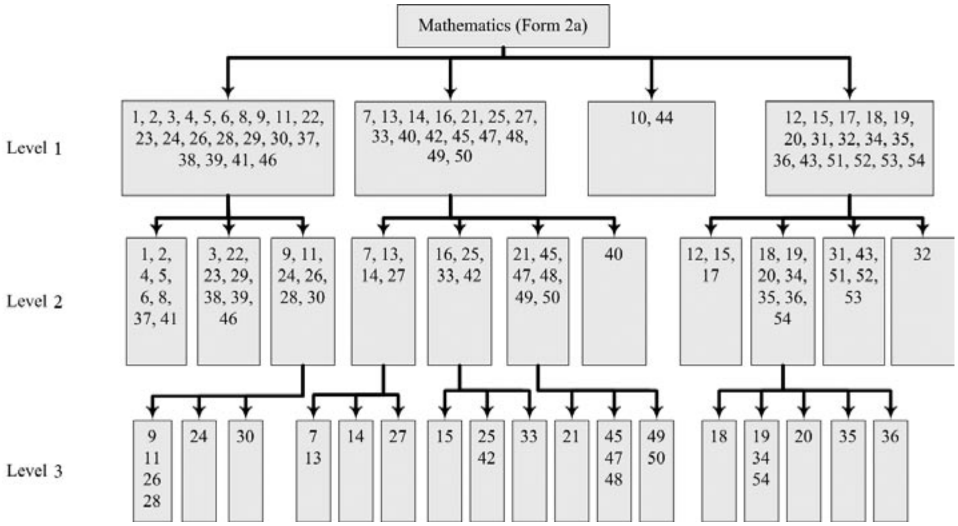

Figure 1. The decomposed DETECT clusters using data from SAT Mathematics Book 2a.

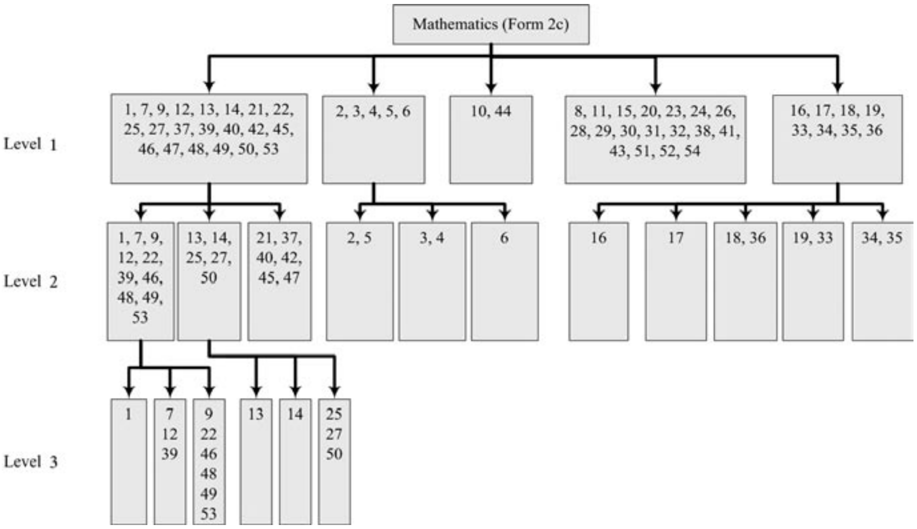

Figure 2. The decomposed DETECT clusters using data from SAT Mathematics Book 2c.

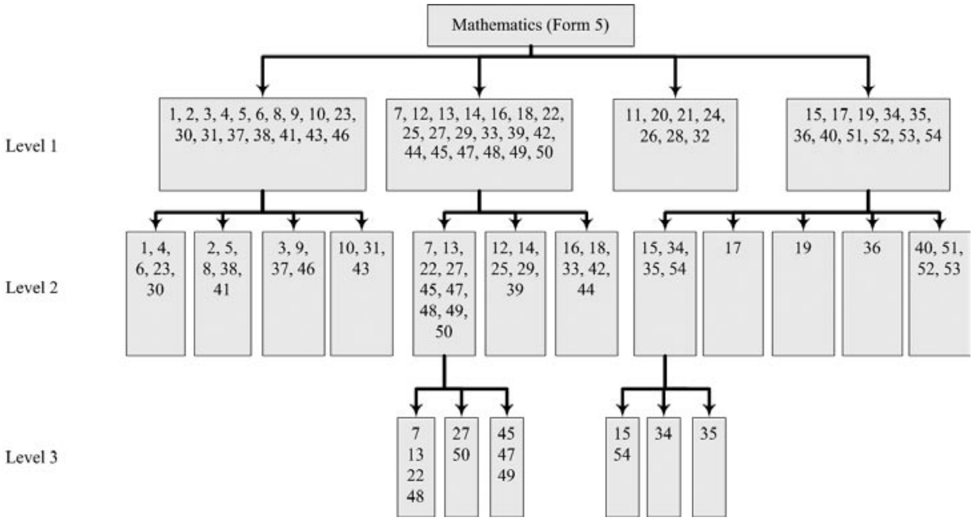

Figure 3. The decomposed DETECT clusters using data from SAT Mathematics Book 5.

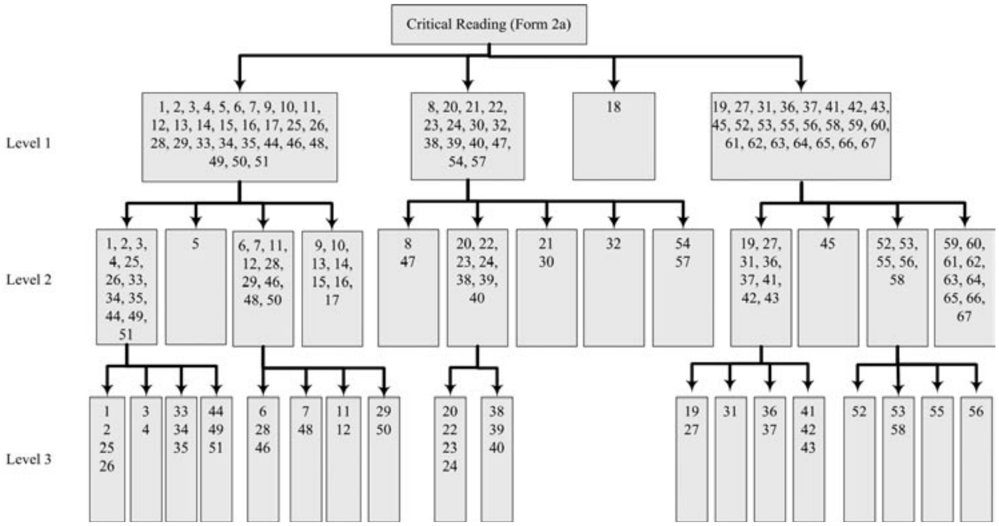

Figure 4. The decomposed DETECT clusters using data from SAT Critical Reading Book 2a.

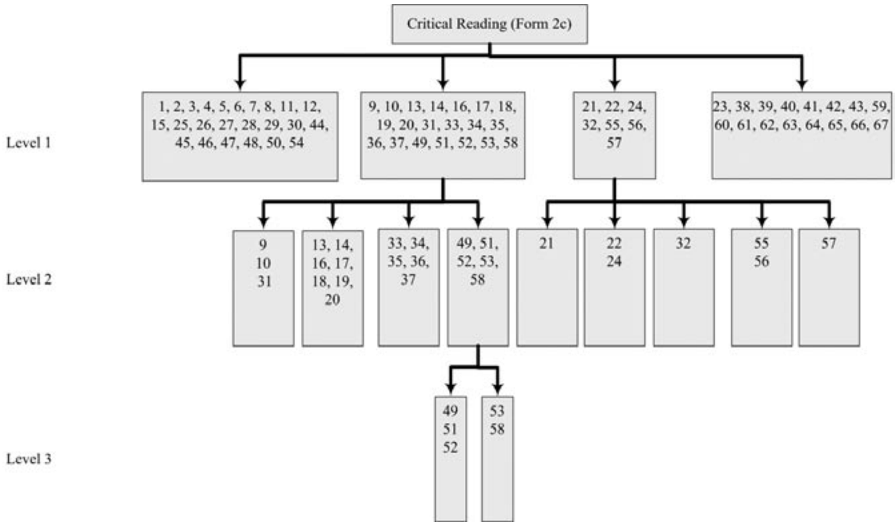

Figure 5. The decomposed DETECT clusters using data from SAT Critical Reading Book 2c.

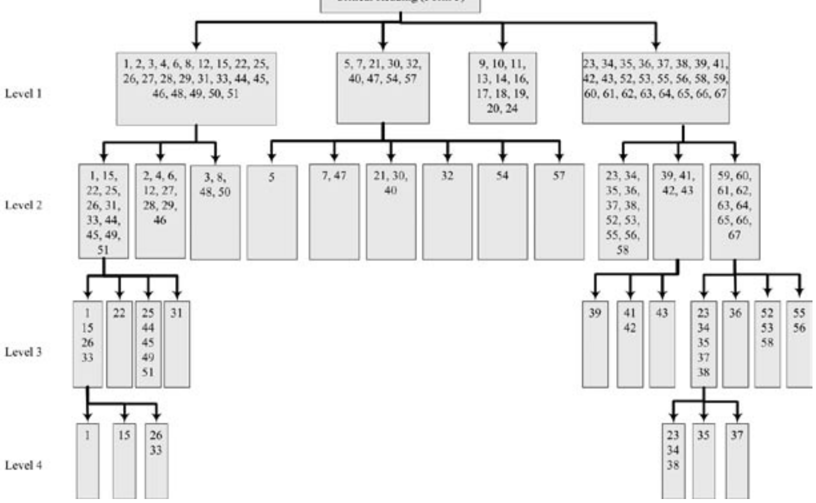

Figure 6. The decomposed DETECT clusters using data from SAT Critical Reading Book 5.

27 independent clusters were identified in three levels. The smallest cluster contained 1 item and the largest cluster contained 9 items. Using Book 2c, 12 independent clusters were identified in three levels. The smallest cluster contained 1 item and the largest cluster contained 24 items. Using Book 5, 25 independent clusters were identified in four levels. The smallest cluster contained 1 item and the largest cluster contained 11 items. As with mathematics, the final independent clusters in critical reading were dissimilar and systematic patterns were not apparent. Again, these results would be difficult to interpret because the clusters are inconsistent across the three critical reading forms.

Initially, these results were surprising to us. We expected a much higher degree of cluster consistency across the three forms. However, it is important to remember that DETECT was created to determine the multidimensional structure underlying test data that display approximate simple structure (Zhang and Stout, 1999; see also Stout et al., 1996, pp. 350–351). No systematic studies have been conducted to evaluate the accuracy of the DETECT partition when the data display differing degrees of complex structure (i.e., a range of r_{Max} values below 0.80).[12] Because DETECT appears to yield unreliable clusters when data display complex structure—as is the case with the SAT mathematics and critical reading items in this study—it is difficult to interpret the clusters substantively. Therefore, to identify a more consistent and, hopefully, interpretable multidimensional structure, the data were fit to exploratory multidimensional models using NOHARM.

We began by computing the results for a series of well-known decision rules used in linear factor analysis for determining the number of dimensions because NOHARM requires the user to specify the expected number of dimensions, even with an exploratory approach. Because many different approaches exist for determining how many dimensions to retain in a linear factor analysis, we applied three widely used decision rules, including Cattell's Scree test, the Kaiser rule, and the minimum average partial (MAP) method. The use of different decision rules is recommended when attempting to determine the number of dimensions to retain in an exploratory analysis (Preacher and MacCallum, 2003).

Cattell's Scree test suggests that if a dimension is significant it will have a large eigenvalue. The magnitude of the eigenvalues can be evaluated graphically by plotting each value. Those eigenvalues that are similar in a plot will form a straight line. The number of eigenvalues that fall above the line are considered the dimensions that account for a majority of the variability in the analysis and, therefore, indicate the number of dominant dimensions. The Kaiser rule states that the dimensionality for a test equals the number of factors where at least three factor loadings equal or exceed 0.30 on any single dimension. Thus, the number of dimensions can be evaluated by assessing the magnitude of the factor loadings in the varimax solution. The MAP method is based on an evaluation of the partial correlation matrices. Initially, the average squared correlations in the off diagonal of the tetrachoric correlation matrix are calculated. The first factor is partialed out of the tetrachoric correlation matrix of the observed variables and the average of the squared partial correlations in the off diagonals of the resulting partial correlation matrix is calculated. Then, the first and second factors are partialed out of the tetrachoric correlation matrix of the observed variables. Likewise, the average of the squared partial correlations is computed. After the minimum average squared partial correlation is obtained, no additional components are extracted. The minimum, which equals the number of dimensions, is reached when the residual matrix directly resembles an identity matrix suggesting that all the off diagonal elements are close to zero.

The results for the three decision rules are presented in Table 4. The number of dimensions for mathematics was consistent across forms at two. Therefore, the number of dimensions that underlie the SAT mathematics items could range from two to five when the factor-analytic and DETECT (see Table 3) results are considered together. The number of dimensions for critical reading ranged from two to three using the three decision rules. Therefore, the number of dimensions that underlie the SAT critical reading items could range from two to five when the factor-analytic and DETECT results are considered together.

To determine which model provided the most parsimonious fit, all four models were fit to the mathematics and critical reading items across samples. To evaluate goodness of fit, NOHARM reports Tanaka's (1993) unweighted least squares goodness-of-fit index and the root mean square residual (RMSR). There are no interpretative guidelines for Tanaka's index, other than a higher value implies better model fit. A RMSR equal to

Table 4

Results for Three Popular Decision Rules Used to Determine the Number of Dimensions Underlying the SAT Mathematics and Critical Reading Items

	Mathematics			*Critical Reading*		
Book	*Scree*	*Kaiser*	*MAP*	*Scree*	*Kaiser*	*MAP*
2a	2	2	2	3	3	3
2c	2	2	2	3	3	2
5	2	2	2	3	3	2

[12] We are currently investigating this problem in a simulation study designed to evaluate cluster consistency under differing degrees of simple and complex structure.

Table 5

NOHARM Fit Indices for 2-, 3-, 4-, and 5-Dimensional Models in Mathematics

	Number of Factors							
	2		*3*		*4*		*5*	
Book	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*
2a	**0.994**	**0.004**	0.996	0.003	0.996	0.003	0.997	0.003
2c	**0.994**	**0.004**	0.995	0.003	0.996	0.003	0.996	0.003
5	**0.994**	**0.004**	0.995	0.003	0.996	0.003	0.997	0.003

Note: The indices for the most parsimonious model are in bold.

or less than four times the reciprocal of the square root of the sample size implies good model fit (Fraser, 1988). Because the sample size ranged from 2,202 to 2,443 across the mathematics and critical reading sections, we interpreted the most conservative RMSR value of 0.0809 to indicate good model fit. The fit indices for the 2-, 3-, 4-, and 5-dimensional models in mathematics and critical reading are presented in Tables 5 and 6, respectively.

For mathematics, the 2-dimensional model provides the most parsimonious fit to the three data sets with little change in either Tanaka's index or the RMSR when all models are compared. Therefore, we conclude that the 2-dimensional model provides the best description of dimensional structure for the SAT mathematics data using exploratory analyses. The promax-rotated solution using the common factor parameterization for the mathematics Book 2a data is presented in Table 7. The correlation between the dimensions is moderate at 0.686.

For critical reading, the 3-dimensional model provides the most parsimonious fit to the three data sets with a noteworthy change in Tanaka's index and the RMSR between the 2- and 3-dimensional models but little change between the 3-dimensional solution and the 4- or 5-dimensional models. Therefore, we conclude that the 3-dimensional model provides the best description of dimensional structure for the SAT critical reading data using exploratory analyses. The promax-rotated solution using the common factor parameterization for the critical reading Book 2a data is presented in Table 8. The correlations between dimensions were moderate, ranging from 0.655 to 0.726.

Table 6

NOHARM Fit Indices for 2-, 3-, 4-, and 5-Dimensional Models in Critical Reading

	Number of Factors							
	2		*3*		*4*		*5*	
Book	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*
2a	0.989	0.005	**0.992**	**0.004**	0.994	0.003	0.994	0.004
2c	0.989	0.005	**0.992**	**0.004**	0.993	0.004	0.994	0.004
5	0.990	0.005	**0.993**	**0.004**	0.994	0.004	0.994	0.004

Note: The indices for the most parsimonious model are in bold.

Table 7

Promax Rotated 2-Dimensional Exploratory Solution for Mathematics Book 2a

Item	*Dimension*	
	1	*2*
5	**0.902**	-0.291
22	**0.818**	-0.111
3	**0.818**	-0.192
1	**0.787**	-0.194
23	**0.759**	-0.084
2	**0.757**	-0.187
8	**0.737**	-0.068
45	**0.689**	0.035
21	**0.675**	-0.103
39	**0.655**	0.131
9	**0.646**	0.035
4	**0.637**	-0.056
46	**0.632**	0.130
37	**0.611**	-0.011
26	**0.583**	0.189
49	**0.577**	0.065
48	**0.575**	0.253
6	**0.530**	-0.064
13	**0.515**	0.183
7	**0.501**	0.144
11	**0.499**	0.196
29	**0.495**	0.197
38	**0.481**	0.011
28	**0.448**	0.308
51	**0.434**	0.408
12	**0.424**	0.296
25	**0.418**	0.275
30	**0.408**	0.238
41	**0.405**	0.152
50	0.393	**0.431**
14	**0.382**	0.129
24	**0.376**	0.253
27	**0.361**	0.290
52	0.320	**0.552**
43	0.238	**0.288**
32	0.237	**0.269**
15	0.197	**0.510**
40	0.195	**0.340**
31	0.189	**0.352**
53	0.179	**0.577**
42	0.154	**0.415**
33	0.109	**0.525**
10	0.051	**0.471**
17	0.048	**0.529**
54	-0.007	**0.934**
18	-0.011	**0.672**
36	-0.032	**0.519**
16	-0.035	**0.574**
44	-0.111	**0.596**
34	-0.135	**0.774**
19	-0.244	**0.866**
35	-0.246	**0.645**
20	-0.322	**0.706**

Note: The item with the highest factor loading for each dimension is in bold.

Item	*Dimension Correlation*	
	1	*2*
1	1.000	
2	0.686	1.000

Table 8

Promax Rotated 3-Dimensional Exploratory Solution for Critical Reading Book 2a

Item	*Dimension*		
	1	*2*	*3*
1	**0.910**	-0.071	-0.185
25	**0.829**	-0.312	0.188
2	**0.772**	0.061	-0.169
10	**0.770**	-0.336	0.100
49	**0.671**	-0.070	0.104
44	**0.633**	-0.060	0.074
26	**0.609**	0.125	-0.072
3	**0.581**	0.125	-0.149
4	**0.527**	0.015	-0.012
33	**0.525**	0.194	-0.035
16	**0.510**	-0.025	0.091
51	**0.505**	0.046	0.144
9	**0.496**	-0.069	0.009
15	**0.469**	0.166	-0.036
17	**0.465**	-0.025	0.185
28	**0.455**	0.157	-0.065
46	**0.428**	0.188	-0.033
14	**0.421**	0.173	-0.019
19	**0.419**	0.081	0.240
12	**0.405**	0.333	-0.148
35	**0.399**	0.236	0.048
6	**0.395**	0.288	-0.149
13	**0.362**	0.137	0.073
29	**0.359**	0.400	-0.129
60	0.336	-0.044	**0.465**
58	0.335	-0.139	**0.539**
18	**0.321**	0.115	0.076
5	**0.318**	0.178	-0.010
37	**0.313**	0.184	0.123
27	**0.301**	0.066	0.175
31	**0.275**	0.027	0.175
38	0.269	**0.470**	-0.010
23	0.269	**0.345**	0.140
61	0.263	0.142	**0.342**
20	0.249	**0.430**	0.062
34	0.243	**0.444**	-0.012
45	**0.241**	0.066	0.157
52	**0.204**	0.133	0.182
53	0.196	0.152	**0.220**
55	0.190	**0.246**	0.167
50	0.189	**0.568**	-0.174
59	0.179	0.131	**0.443**
22	0.149	**0.457**	0.014
48	0.149	**0.388**	-0.157
56	0.143	**0.224**	0.167
54	0.137	**0.220**	-0.028
11	0.109	**0.592**	-0.187
7	0.092	**0.600**	-0.250
30	0.073	**0.422**	-0.044
36	0.064	**0.359**	0.198
65	0.023	-0.078	**0.914**
41	0.018	**0.402**	0.223
63	0.005	0.033	**0.702**
66	-0.005	0.352	**0.431**
24	-0.023	**0.541**	0.091
62	-0.054	-0.021	**0.721**
42	-0.069	**0.362**	0.331
64	-0.082	**0.428**	0.202
39	-0.091	**0.625**	-0.009
47	-0.115	**0.516**	-0.029
21	-0.153	**0.591**	-0.004
57	-0.167	**0.525**	0.003
40	-0.197	**0.747**	0.016
8	-0.206	**0.768**	-0.150
43	-0.225	**0.467**	0.219
32	-0.239	**0.587**	-0.012
67	-0.239	**0.540**	0.277

Note: The item with the highest factor loading for each dimension is in bold.

Item	*Dimension Correlations*		
	1	*2*	*3*
1	1.000		
2	0.726	1.000	
3	0.703	0.655	1.000

Part #2: Confirmatory Dimensionality Results

Test Specifications

Overview

The first set of confirmatory analyses used the College Board test specifications as the organizing principle to guide the assessment of dimensionality. Test specifications outline the achievement domain and help test developers obtain a representative sample of items from this domain. The specifications also guide item writing and help structure the final form of the test based on the content domain that the test is designed to measure. Thus, a thorough analysis of the content areas measured by the test may help identify a subset of items that measure distinct dimensions. Two categories in the test specifications were used in the first confirmatory analysis for both mathematics and critical reading because these categories highlight important item features according to the test developers, they are readily available and easy to use, and they can guide interpretation. It would also be easy for the College Board to create subscores based on the test specification categories.

Mathematics

The two categories in the College Board mathematics section specifications used for our dimensionality analyses were content area and item type. The first category, content area, included four areas: algebra, arithmetic, geometry, and miscellaneous. Content areas reflect the curricular domain measured by the SAT, which is tied to the examinee's program of studies. By adequately representing the content areas in the test specifications, developers attempt to establish the validity of test score inferences based on content. The second category, item type, measured three areas: comprehension, nonroutine/insightful, and routine. Item type is intended to reflect the nature of the problems examinees are expected to solve on the test. By adequately representing the item

Table 9

Confirmatory DIMTEST Results by Mathematics Content Area

	Algebra		*Arithmetic*		*Geometry*		*Miscellaneous*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	1.5673	0.0585	1.8777	0.0302	2.8976	0.0019	1.8090	0.0352
2c	0.2539	0.3998	1.4623	0.0718	3.7754	0.0001	1.7082	0.0438
5	0.4281	0.3343	1.8288	0.0337	2.6734	0.0038	1.0722	0.1418

type in the test specifications, developers attempt to establish the validity of the test score inferences based the examinees' problem-solving skills.

Using content area as the organizing principle, confirmatory DIMTEST rejected three content areas across samples (see Table 9). That is, DIMTEST rejected the null hypothesis for one sample in algebra (if we consider the *p*-value of 0.0585 in the Book 2a sample as statistically significant), two samples in arithmetic, three samples in geometry, and two samples in miscellaneous. In other words, if we interpret the results from Book 2a as indicative of multidimensionality and attempt to replicate the result across samples, then our findings suggest that arithmetic, geometry, and miscellaneous are dimensionally distinct content areas, but that algebra is not.

Using item type as the organizing principle, confirmatory DIMTEST rejected the null hypothesis for the nonroutine/insightful items across two samples (see Table 10). Thus, nonroutine/insightful items are dimensionally distinct from comprehension and routine items, whereas comprehension and routine items are not distinct from one another.

NOHARM was then used to estimate the parameters for the 4-dimensional model using the items associated with content area. Only the content-area model was fit to the data because two of the three item-type dimensions were not found to be dimensionally distinct from one another. The fit indices for the 4-dimensional model across the three samples are shown in Table 11. Tanaka's index, which has no established interpretive guidelines, was large and comparable across samples, ranging from 0.988 to 0.989. The RMSR was small and comparable across samples at 0.005. Recall, Fraser (1988) suggested an RMSR equal to or less that four times the reciprocal of the square root of the sample size implies good model fit. Using Fraser's suggested outcome, the RMSR implies good model fit across all three samples. The common factor parameter estimates for the Book 2a data are presented in Table 12. The correlations between the dimensions were high, ranging from 0.946 to 1.000.[13]

Table 10

Confirmatory DIMTEST Results by Mathematics Item Type

	Comprehension		*Nonroutine/Insightful*		*Routine*	
Book	***T***	***p***	***T***	***p***	***T***	***p***
2a	-1.2697	0.8979	1.9432	0.0260	0.0840	0.4665
2c	-1.3190	0.9064	0.2689	0.3940	1.1669	0.1216
5	0.3581	0.3601	2.2614	0.0119	0.7496	0.2268

Table 11

NOHARM Fit Indices for Confirmatory Content-Based 4-Dimensional Model in Mathematics Across Samples

	Fit Index	
Book	***Tanaka***	***RMSR***
2a	0.989	0.005
2c	0.988	0.005
5	0.988	0.005

Critical Reading

Two categories in the College Board test specifications for critical reading were used for our dimensionality analyses. The first category was item format. Two item formats were measured: sentence completion and critical reading. The second category was reading passage, which only applied to the critical reading items given these items were divided into short and long reading passages. Eight passages were presented, and each passage contained a different number of test items. Passages 1 and 2 contained 2 items, passage 3 contained 12 items, passage 4 contained 13 items, passages 5 and 6 contained 2 items, passage 7 contained 6 items, and passage 8 contained 9 items.[14] Taken together, the sentence completion items and the reading passages are designed to measure examinees' ability to identify genre, relationships among parts of text, cause and effect, rhetorical devices, and comparative arguments in passages taken from the natural sciences, humanities, social science, and literary fiction. By adequately representing the item formats and reading passages, test developers are attempting to establish the validity of the test score inferences based on the examinees' verbal and critical reading skills.

Using item format as the organizing principle, confirmatory DIMTEST rejected the null hypothesis producing small *p*-values across all three samples indicating the data are not unidimensional (see Table 13). Next, reading passage was used as the organizing principle. Confirmatory DIMTEST rejected the null hypothesis for all passages across samples, except passage 5, indicating the items associated

[13] Due to estimation error, some of the correlations meet or exceeded 1.0. Fortunately, the errors appear to be small as the overestimated correlations are close to the upper bound of 1.0 in all analyses.

[14] The passages are ordered by section according to the nSAT data file we received. Thus, in this report, passages 1, 2, and 3 are passages 1, 2, and 3 of Section 1, respectively; passage 4 is the paired passage in Section 3; passages 5, 6, 7, and 8 are passages 1, 2, 3, and 4 in Section 2, respectively.

Table 12

Four-Dimensional Solution for Confirmatory Content-Based Mathematics Model Using Book 2a

Item	Dimension			
	1	2	3	4
51	**0.772**	0.0	0.0	0.0
39	**0.743**	0.0	0.0	0.0
45	**0.696**	0.0	0.0	0.0
23	**0.666**	0.0	0.0	0.0
12	**0.661**	0.0	0.0	0.0
9	**0.656**	0.0	0.0	0.0
29	**0.643**	0.0	0.0	0.0
47	**0.638**	0.0	0.0	0.0
3	**0.638**	0.0	0.0	0.0
1	**0.627**	0.0	0.0	0.0
7	**0.605**	0.0	0.0	0.0
27	**0.595**	0.0	0.0	0.0
21	**0.574**	0.0	0.0	0.0
33	**0.564**	0.0	0.0	0.0
19	**0.557**	0.0	0.0	0.0
17	**0.510**	0.0	0.0	0.0
43	**0.478**	0.0	0.0	0.0
44	**0.423**	0.0	0.0	0.0
35	**0.341**	0.0	0.0	0.0
26	0.0	**0.726**	0.0	0.0
53	0.0	**0.697**	0.0	0.0
8	0.0	**0.659**	0.0	0.0
11	0.0	**0.650**	0.0	0.0
15	0.0	**0.637**	0.0	0.0
49	0.0	**0.611**	0.0	0.0
30	0.0	**0.599**	0.0	0.0
24	0.0	**0.581**	0.0	0.0
41	0.0	**0.521**	0.0	0.0
14	0.0	**0.478**	0.0	0.0
6	0.0	**0.461**	0.0	0.0
36	0.0	**0.438**	0.0	0.0
54	0.0	0.0	**0.913**	0.0
48	0.0	0.0	**0.777**	0.0
50	0.0	0.0	**0.763**	0.0
28	0.0	0.0	**0.704**	0.0
13	0.0	0.0	**0.659**	0.0
5	0.0	0.0	**0.658**	0.0
25	0.0	0.0	**0.646**	0.0
18	0.0	0.0	**0.598**	0.0
37	0.0	0.0	**0.596**	0.0
2	0.0	0.0	**0.584**	0.0
34	0.0	0.0	**0.568**	0.0
42	0.0	0.0	**0.513**	0.0
40	0.0	0.0	**0.488**	0.0
16	0.0	0.0	**0.472**	0.0
32	0.0	0.0	**0.465**	0.0
52	0.0	0.0	0.0	**0.808**
46	0.0	0.0	0.0	**0.723**
38	0.0	0.0	0.0	**0.478**
31	0.0	0.0	0.0	**0.495**
22	0.0	0.0	0.0	**0.706**
20	0.0	0.0	0.0	**0.339**
10	0.0	0.0	0.0	**0.468**
4	0.0	0.0	0.0	**0.582**

Note: The items in the table are sorted by content area so the factor loadings are easier to interpret. Dimensions 1 through 4 correspond to algebra, arithmetic, geometry, and miscellaneous, respectively.

Item	Dimension Correlations			
	1	2	3	4
1	1.000			
2	0.987	1.000		
3	0.974	0.970	1.000	
4	1.001	0.982	0.946	1.000

Table 13

Confirmatory DIMTEST Results by Critical Reading Item Format

Book	T	p
2a	6.0065	0.0000
2c	9.3176	0.0000
5	7.8681	0.0000

Note: The sentence completion items served as the AT and the critical reading items served as the PT.

Table 14

Confirmatory DIMTEST Results by Reading Passage

	Passage 1		Passage 2		Passage 3		Passage 4	
Book	T	p	T	p	T	p	T	p
2a	2.9267	0.0017	2.9879	0.0014	3.3044	0.0005	6.0956	0.0000
2c	3.2240	0.0006	1.8548	0.0318	4.0946	0.0000	3.5783	0.0002
5	2.7549	0.0029	1.8719	0.0306	6.1057	0.0000	5.0210	0.0000
	Passage 5		Passage 6		Passage 7		Passage 8	
Book	T	p	T	p	T	p	T	p
2a	0.2768	0.6090	2.4242	0.0077	4.2212	0.0000	3.8733	0.0001
2c	0.2164	0.4143	3.1612	0.0008	3.6696	0.0001	7.5926	0.0000
5	0.7026	0.2411	1.7746	0.0380	4.0021	0.0000	6.2434	0.0000

Note: Each reading passage served as a separate AT, with the remaining items serving as a PT.

Table 15

Confirmatory DIMTEST Results by Critical Reading Item Format and Passage

	Sentence Completion (19)		Passage 1 (2)		Passage 2 (2)	
Book	T	p	T	p	T	p
2a	6.0065	0.0000	1.8683	0.0309	3.0394	0.0012
2c	9.3176	0.0000	2.6443	0.0041	1.2864	0.0992
5	7.8681	0.0000	2.4838	0.0065	2.2081	0.0136
	Passage 3 (12)		Passage 4 (13)		Passage 5 (2)	
Book	T	p	T	p	T	p
2a	5.9788	0.0000	7.2589	0.0000	-0.6159	0.7310
2c	5.3864	0.0000	4.3784	0.0000	-0.6190	0.7320
5	6.5778	0.0000	5.3429	0.0000	0.9707	0.1659
	Passage 6 (2)		Passage 7 (6)		Passage 8 (9)	
Book	T	p	T	p	T	p
2a	2.6082	0.0046	5.3984	0.0000	7.3766	0.0000
2c	2.5663	0.0051	4.3802	0.0000	8.7440	0.0000
5	2.1970	0.0140	4.7473	0.0000	8.4999	0.0000

Note: Each dimension served as a separate AT, with the remaining items serving as a PT. The number of items measuring each dimension is in parentheses on the top of the column.

Table 16

NOHARM Fit Indices for Confirmatory Item Format and Passage 9-Dimensional Model in Critical Reading Across Samples

	Fit Index	
Book	Tanaka	RMSR
2a	0.988	0.005
2c	0.988	0.005
5	0.990	0.005

with each reading passage measure a distinct dimension (see Table 14). This result is consistent with previous research indicating that reading comprehension passages tend to assess distinct dimensions (see, for example, Bolt, in press; Gierl, 2005; Stout et al., 1996). Finally, item format and reading passage were combined because almost every dimension in each category was dimensionally distinct, producing nine clusters for the confirmatory analysis. DIMTEST rejected the null hypothesis for all item clusters, except sample 2c in passage 2 and all samples in passage 5, indicating that the clusters associated with sentence completion and reading passage are, for the most part, dimensionally distinct from one another (see Table 15).

Because the sentence completion and passage-based reading comprehension items produced interpretable, dimensionally distinct clusters, the factor loadings for a 9-dimensional model were estimated using NOHARM. This model provided adequate fit to the data across all three samples, as shown in Table 16. Tanaka's index was high and comparable across samples ranging from 0.988 to 0.990. The RMSR was low and comparable across samples at 0.005. The common factor parameter estimates for the Book 2a data are presented in Table 17. The correlations between the dimensions were moderate to high, ranging from 0.525 to 1.000 (see Footnote 13).

Table 17

Nine-Dimensional Solution for Confirmatory Item Format and Passage Model in Critical Reading Using Book 2a

Item	*Dimension*								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	**0.675**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	**0.673**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	**0.559**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	**0.522**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	**0.471**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	**0.532**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	**0.456**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	**0.417**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	**0.587**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	**0.742**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	**0.580**	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	**0.663**	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	**0.547**	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	**0.557**	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	**0.584**	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	**0.561**	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	**0.599**	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	**0.487**	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	**0.698**	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	**0.704**	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	**0.411**	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	**0.593**	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	**0.710**	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	**0.570**	0.0	0.0	0.0	0.0	0.0
25	**0.704**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	**0.656**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	**0.510**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	**0.540**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	**0.620**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30	**0.437**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31	0.0	0.0	0.0	0.0	**0.447**	0.0	0.0	0.0	0.0
32	0.0	0.0	0.0	0.0	**0.317**	0.0	0.0	0.0	0.0
33	0.0	0.0	0.0	0.0	**0.664**	0.0	0.0	0.0	0.0
34	0.0	0.0	0.0	0.0	**0.646**	0.0	0.0	0.0	0.0
35	0.0	0.0	0.0	0.0	**0.654**	0.0	0.0	0.0	0.0
36	0.0	0.0	0.0	0.0	**0.573**	0.0	0.0	0.0	0.0
37	0.0	0.0	0.0	0.0	**0.584**	0.0	0.0	0.0	0.0
38	0.0	0.0	0.0	0.0	**0.697**	0.0	0.0	0.0	0.0
39	0.0	0.0	0.0	0.0	**0.503**	0.0	0.0	0.0	0.0
40	0.0	0.0	0.0	0.0	**0.538**	0.0	0.0	0.0	0.0
41	0.0	0.0	0.0	0.0	**0.592**	0.0	0.0	0.0	0.0
42	0.0	0.0	0.0	0.0	**0.562**	0.0	0.0	0.0	0.0
43	0.0	0.0	0.0	0.0	**0.415**	0.0	0.0	0.0	0.0
44	**0.637**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
45	**0.435**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
46	**0.571**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47	**0.354**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
48	**0.382**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
49	0.0	0.0	0.0	0.0	0.0	**0.603**	0.0	0.0	0.0
50	0.0	0.0	0.0	0.0	0.0	**0.507**	0.0	0.0	0.0
51	0.0	0.0	0.0	0.0	0.0	0.0	**0.731**	0.0	0.0
52	0.0	0.0	0.0	0.0	0.0	0.0	**0.531**	0.0	0.0
53	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.558**	0.0
54	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.336**	0.0
55	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.598**	0.0
56	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.527**	0.0
57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.361**	0.0
58	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.713**	0.0
59	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.719**
60	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.731**
61	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.722**
62	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.586**
63	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.682**
64	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.527**
65	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.788**
66	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.740**
67	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.551**

Note: Dimensions 1 through 9 correspond to sentence completion items and passage 1 through 8 items, respectively.

Item	*Dimension Correlations*								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	1.000								
2	0.707	1.000							
3	0.861	0.533	1.000						
4	0.882	0.693	0.799	1.000					
5	0.878	0.661	0.790	0.902	1.000				
6	1.099	0.786	1.008	1.014	1.003	1.000			
7	0.820	0.639	0.736	0.803	0.789	1.038	1.000		
8	0.784	0.693	0.669	0.846	0.870	0.978	0.835	1.000	
9	0.796	0.525	0.680	0.822	0.825	0.912	0.788	0.827	1.000

Cognitive Skill Categories

Overview

The second set of confirmatory analyses used the College Board cognitive skill categories as the organizing principle to guide the assessment of dimensionality. The skill categories were first introduced in 2004 at a symposium conducted at the annual meeting of the National Council on Measurement in Education in San Diego. The symposium was titled, "Connecting Curriculum and Assessment Through Meaningful Score Reports." College Board researchers presented papers on their work related to extracting detailed diagnostic information from the SAT, focusing on the mathematics and critical reading sections. Two of the papers in the symposium focused specifically on developing cognitive skill categories and coding SAT items that represented these skills. To date, only preliminary empirical evidence has been collected to evaluate the reliability and validity of these skills. Therefore, our confirmatory dimensionality analyses were designed to assess empirically the dimensions associated with the cognitive skill categories.

The categories were identified by asking subject-matter experts to identify the skills required to solve the SAT items, keeping in mind the strategies that would be used by the "just qualified candidate" who responded correctly to the item (Huff, 2004, p. 11). These instructions were designed to circumvent potential problems associated with strategy diversity, meaning the coding problems that can occur when examinees are perceived to use different strategies to solve an item (see Gierl, 1997a, 1997b; Gierl, Bisanz, Bisanz, and Boughton, 2003, p. 300; Gierl, Bisanz, and Li, 2004 for a discussion and illustration of this coding issue). In other words, the experts were asked to envision a representative group of examinees who all used the same strategy to solve each item.

After the categories were defined, subject-matter experts also coded the SAT items for these skills. These experts were asked to (a) identify every skill required to solve each item, but also to (b) specify the primary skill, which was defined as the most important skill required to solve the item. These instructions have significant implications for our confirmatory dimensionality analyses. Instruction 1—code all skills—implies the data will display a complex factor structure, especially if the SAT items are believed to elicit multiple skills. In other words, if cognitive skills represent important dimensions and items elicit multiple cognitive skills, then each item will load on more than one dimension. Alternatively, instruction 2—code primary skill only—implies the data will display a simple factor structure because only one cognitive skill will be associated with each item. Taken together, these two instructions, and the item coding they produce, will affect the dimensionality analyses because they prescribe either a complex or a simple factor structure. *For this report, only the data associated with instruction 2 are used.* The data associated with instruction 1 will be evaluated after we conduct additional reliability studies to determine the consistency of the skill category coding across the subject-matter experts.

Mathematics

O'Callaghan, Morley, and Schwartz (2004) identified five skill categories in mathematics. However, the College Board only designated four of these skills as primary. The four primary mathematics skills included applying basic mathematics knowledge, applying advanced mathematics knowledge, managing complexity, and modeling and insight. A short description of each primary mathematics skill is presented in Appendix A.

Using skill as the organizing principle, confirmatory DIMTEST rejected some skill categories across samples (see Table 18). That is, confirmatory DIMTEST rejected the null hypothesis for two samples in basic mathematics (the Book 2a sample was close to the critical value with p=0.0537, so this result was considered statistically significant), one sample in advanced mathematics (the Book 2a sample was close to the critical value with p=0.0547, so this result was considered statistically significant), one sample in managing complexity (the Book 2a sample, again, was close to the critical value with p=0.0642, so this result was considered statistically significant), and all three samples in modeling insight. If we interpret the results from Book 2a as indicative of multidimensionality and an attempt to replicate the result across samples, then our findings suggest that basic mathematics and modeling and insight are dimensionally distinct skills, whereas advanced mathematics and managing complexity are not.

Because the items in the mathematics content areas produced dimensionally distinct clusters when the test specifications were used as the organizing principle, skills and content area were combined and the dimensionality of the resulting clusters was evaluated. Both between- and within-cluster analyses were conducted. For the between-cluster analyses, the dimensional homogeneity of the skills between each content area was evaluated. That is, each skill category for each content area served as a separate AT with all remaining items on the mathematics test serving

Table 18

Confirmatory DIMTEST Results by Mathematics Skill Category

	Basic		*Advanced*		*Complexity*		*Insight*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	1.6096	0.0537	1.6012	0.0547	1.5206	0.0642	4.7194	0.0000
2c	0.8171	0.2069	0.0440	0.4824	0.8509	0.1974	4.2834	0.0000
5	2.7506	0.0030	0.7195	0.2359	1.2272	0.1099	3.3069	0.0005

Note: Each dimension served as a separate AT, with the remaining items serving as a PT.

Table 19

Confirmatory DIMTEST Results by Mathematics Content Area and Skill Category

	Algebra						*Arithmetic*					
	2a		*2c*		*5*		*2a*		*2c*		*5*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Basic	1.1272	0.1298	2.0255	0.0214	0.0207	0.4918	2.6116	0.0045	3.3123	0.0005	2.9777	0.0015
Advanced	0.7400	0.2296	0.8087	0.2093	0.7332	0.2317	1.6206	0.0525	0.9530	0.1703	1.1960	0.1158
Complexity	0.4191	0.3376	0.7641	0.2224	0.6971	0.2429	0.7569	0.2245	-0.5067	0.6938	0.5428	0.2936
Insight	1.8297	0.0336	1.1037	0.1349	1.6876	0.0457	0.4892	0.3123	0.6604	0.2545	1.1488	0.1253
	Geometry						*Miscellaneous*					
	2a		*2c*		*5*		*2a*		*2c*		*5*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Basic	-0.3900	0.6517	0.5642	0.2863	0.6697	0.2515	-0.5774	0.7182	1.2432	0.1069	0.9960	0.1596
Advanced	0.8435	0.1995	1.6715	0.0473	1.3615	0.0867	NA	NA	NA	NA	NA	NA
Complexity	1.6605	0.0484	0.5368	0.2957	0.2151	0.4149	0.8344	0.2020	1.2115	0.1129	1.3159	0.0941
Insight	1.0775	0.1406	0.6525	0.2570	1.8971	0.0289	1.7759	0.0379	0.4081	0.3416	0.3241	0.3729

Note: Each dimension served as a separate AT, with the remaining items serving as a PT. NA means that the cell contained 0 items.

as a PT. The results are presented in Table 19. For algebra, confirmatory DIMTEST rejected the null hypothesis for one sample in basic mathematics and two samples in modeling and insight. For arithmetic, DIMTEST rejected all three samples for basic mathematics and one sample in advanced mathematics (assuming p=0.0525 is considered a statistically significant result). For geometry, DIMTEST rejected one sample in advanced mathematics, one sample in managing complexity, and one sample in modeling and insight. For miscellaneous, DIMTEST rejected one sample in modeling and insight. Taken together, only modeling and insight for algebra and basic mathematics in arithmetic appear to be dimensionally distinct skill clusters across the content areas.

For the within-cluster analyses, the dimensional homogeneity of the skill within each content area was evaluated. In other words, each skill category served as a separate AT with the remaining items within each mathematics content area serving as a PT. The results are presented in Tables 20a through 20d. For algebra, DIMTEST rejected the null hypothesis for one sample in basic mathematics, one sample in advanced mathematics, and three samples in modeling and insight. Thus, it appears that modeling complexity is the only detectable dimension in algebra. For arithmetic, DIMTEST rejected three samples for basic mathematics, two samples for advanced mathematics, and three samples in modeling and insight. Thus, the skills associated with basic mathematics, advanced mathematics, and modeling and insight are detectable in arithmetic. For geometry, DIMTEST rejected two samples in basic mathematics and one sample in managing complexity (in this case, p=0.0567 was considered statistically significant because it was close to the critical value of 0.05). Thus, only basic mathematics skills are detectable in geometry. Finally, for miscellaneous, DIMTEST rejected no samples, implying no dimensionally distinct skill categories exist for these items.

Table 20a

Confirmatory DIMTEST Results by Mathematics Skill Category Within Algebra

	Basic (8)		*Advanced (3)*		*Complexity (3)*		*Insight (5)*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	1.1608	0.1229	0.8452	0.1990	0.0416	0.4834	3.4457	0.0003
2c	1.8435	0.0326	0.4491	0.3267	0.7735	0.2196	2.4064	0.0081
5	1.0645	0.1436	1.9034	0.0285	1.0931	0.1372	3.0284	0.0012

Note: Each skill category served as a separate AT, with the remaining items serving as a PT. The number of items measuring each dimension is in parentheses on the top of the column.

Table 20b

Confirmatory DIMTEST Results by Mathematics Skill Category Within Arithmetic

	Basic (3)		*Advanced (2)*		*Complexity (2)*		*Insight (5)*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	2.7026	0.0034	1.9303	0.0268	1.2970	0.0973	2.4437	0.0073
2c	2.3250	0.0100	1.4589	0.0723	0.8338	0.2022	2.9987	0.0014
5	2.1232	0.0169	1.9772	0.0240	1.1732	0.1204	1.7678	0.0385

Table 20c

Confirmatory DIMTEST Results by Mathematics Skill Category Within Geometry

	Basic (5)		*Advanced (3)*		*Complexity (2)*		*Insight (5)*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	1.1361	0.1280	0.0713	0.5284	1.5832	0.0567	0.3367	0.3682
2c	1.7765	0.0378	1.0469	0.1476	0.3425	0.6340	0.2775	0.3907
5	2.3488	0.0094	0.9502	0.1710	0.0839	0.4666	0.0540	0.4785

Table 20d

Confirmatory DIMTEST Results by Mathematics Skill Category Within Miscellaneous

	Basic (2)		*Advanced (0)*		*Complexity (3)*		*Insight (3)*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	-2.3734	0.9912	NA	NA	1.5034	0.0664	-0.2784	0.6096
2c	0.9525	0.8296	NA	NA	1.1008	0.1355	-1.3265	0.9077
5	-2.6198	0.9956	NA	NA	0.4771	0.3166	-0.8293	0.7965

Note: NA means that the cell contained 0 items.

Table 21

NOHARM Fit Indices for Mathematics Skill Category Within Content Area

	Content Area							
	Algebra		*Arithmetic*		*Geometry*		*Miscellaneous*	
Book	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*	*Tanaka*	*RMSR*
2a	0.993	0.005	0.999	0.003	0.995	0.005	0.998	0.003
2c	0.994	0.005	0.998	0.003	0.994	0.005	0.999	0.003
5	0.995	0.004	0.998	0.004	0.994	0.006	0.997	0.004

Because the skills were dimensionally distinct within several content areas, the factor loadings for four content-based 4-dimensional skill models were estimated using NOHARM. In other words, each skill category defined the multidimensional structure within each content area. These models provided adequate fit to the data across all three samples, as shown in Table 21. Tanaka's index was high and comparable across samples ranging from 0.993 to 0.999. The RMSR was low and comparable across samples, ranging from 0.003 to 0.005. The common factor parameter estimates for the Book 2a data are presented in Table 22 a–d. The correlations between the skill dimensions were high, ranging from 0.817 to 1.000.

Table 22a

Four-Dimensional Solution for Skill Items in Algebra Book 2a

Item	*Dimension*			
	1	*2*	*3*	*4*
14	**0.763**	0.0	0.0	0.0
17	**0.699**	0.0	0.0	0.0
5	**0.672**	0.0	0.0	0.0
9	**0.660**	0.0	0.0	0.0
2	**0.655**	0.0	0.0	0.0
1	**0.626**	0.0	0.0	0.0
10	**0.598**	0.0	0.0	0.0
8	**0.584**	0.0	0.0	0.0
18	0.0	**0.658**	0.0	0.0
4	0.0	**0.653**	0.0	0.0
3	0.0	**0.623**	0.0	0.0
11	0.0	0.0	**0.678**	0.0
12	0.0	0.0	**0.567**	0.0
16	0.0	0.0	**0.422**	0.0
6	0.0	0.0	0.0	**0.544**
7	0.0	0.0	0.0	**0.581**
13	0.0	0.0	0.0	**0.360**
15	0.0	0.0	0.0	**0.506**
19	0.0	0.0	0.0	**0.843**

Note: The items in the table are sorted by skill category so the factor loadings are easier to interpret. Dimensions 1 through 4 correspond to basic mathematics, advanced mathematics, managing complexity, and modeling and insight, respectively.

Item	*Dimension Correlations*			
	1	*2*	*3*	*4*
1	1.000			
2	1.004	1.000		
3	0.927	0.908	1.000	
4	0.862	0.819	0.940	1.000

Table 22b

Four-Dimensional Solution for Skill Items in Arithmetic Book 2a

Item	*Dimension*			
	1	*2*	*3*	*4*
2	**0.728**	0.0	0.0	0.0
10	**0.587**	0.0	0.0	0.0
1	**0.490**	0.0	0.0	0.0
3	0.0	**0.670**	0.0	0.0
6	0.0	**0.623**	0.0	0.0
11	0.0	0.0	**0.642**	0.0
4	0.0	0.0	**0.487**	0.0
5	0.0	0.0	0.0	**0.627**
7	0.0	0.0	0.0	**0.739**
8	0.0	0.0	0.0	**0.611**
9	0.0	0.0	0.0	**0.445**
12	0.0	0.0	0.0	**0.655**

Note: The items in the table are sorted by skill category so the factor loadings are easier to interpret. Dimensions 1 through 4 correspond to basic mathematics, advanced mathematics, managing complexity, and modeling and insight, respectively.

Item	*Dimension Correlations*			
	1	*2*	*3*	*4*
1	1.000			
2	0.842	1.000		
3	0.932	0.897	1.000	
4	0.866	0.976	0.897	1.000

Table 22c

Four-Dimensional Solution for Skill Items in Geometry Book 2a

Item	*Dimension*			
	1	*2*	*3*	*4*
6	**0.676**	0.0	0.0	0.0
2	**0.562**	0.0	0.0	0.0
10	**0.561**	0.0	0.0	0.0
1	**0.525**	0.0	0.0	0.0
11	**0.504**	0.0	0.0	0.0
14	0.0	**0.772**	0.0	0.0
7	0.0	**0.694**	0.0	0.0
12	0.0	**0.541**	0.0	0.0
3	0.0	0.0	**0.638**	0.0
9	0.0	0.0	**0.577**	0.0
4	0.0	0.0	0.0	**0.508**
5	0.0	0.0	0.0	**0.611**
8	0.0	0.0	0.0	**0.465**
13	0.0	0.0	0.0	**0.748**
15	0.0	0.0	0.0	**0.877**

Note: The items in the table are sorted by skill category so the factor loadings are easier to interpret. Dimensions 1 through 4 correspond to basic mathematics, advanced mathematics, managing complexity, and modeling and insight, respectively.

Item	*Dimension Correlations*			
	1	*2*	*3*	*4*
1	1.000			
2	0.981	1.000		
3	1.005	1.010	1.000	
4	0.961	1.049	1.013	1.000

Table 22d

Four-Dimensional Solution for Skill Items in Miscellaneous Book 2a

Item	*Dimension*		
	1	*3*	*4*
4	**0.720**	0.0	0.0
1	**0.598**	0.0	0.0
7	0.0	**0.724**	0.0
6	0.0	**0.467**	0.0
2	0.0	**0.460**	0.0
8	0.0	0.0	**0.846**
5	0.0	0.0	**0.534**
3	0.0	0.0	**0.372**

Note: The items in the table are sorted by skill category so the factor loadings are easier to interpret. Dimensions 1, 3, and 4 correspond to basic mathematics, managing complexity, and modeling and insight, respectively.

Item	*Dimension Correlations*		
	1	*3*	*4*
1	1.000		
3	0.990	1.000	
4	0.817	0.955	1.000

Critical Reading

VanderVeen (2004) identified seven skill categories in critical reading. As in mathematics, the College Board only designated four of these skills as primary. The four primary critical reading skills included determining the meaning of words; understanding the content, form, and function of sentences; understanding the content, form, and function of larger sections of text; and analyzing authors' purposes, goals, and strategies. A short description of each primary critical reading skill is presented in Appendix B.

Using skill as the organizing principle, confirmatory DIMTEST rejected some skill categories across samples (see Table 23). More specifically, confirmatory DIMTEST rejected the null hypothesis for all samples in word meaning, all samples in understanding larger sections of text, and one sample in analyzing authors' purposes, goals, and strategies. If we interpret the results from Book 2a as indicative of multidimensionality and an attempt to replicate the results across samples, then our results suggest that word meaning and understanding larger sections of text are the only two dimensionally distinct skills.

Table 23

Confirmatory DIMTEST Results by Critical Reading Skill Category

	Word Meaning		*Understanding Sentences*		*Understanding Larger Sections*		*Analyzing Purpose*	
Book	*T*	*p*	*T*	*p*	*T*	*p*	*T*	*p*
2a	4.7681	0.0000	0.2039	0.4192	2.5645	0.0052	0.3646	0.3577
2c	5.1321	0.0000	0.3301	0.6293	3.7985	0.0001	2.1892	0.0143
5	4.7999	0.0000	1.1444	0.1262	2.6066	0.0046	0.3344	0.6310

Note: Each dimension served as a separate AT, with the remaining items serving as a PT.

Since item format produced dimensionally distinct clusters when the test specifications were used as the organizing principle, skills and item format were combined and the dimensionality of the resulting clusters was evaluated. As in mathematics, both between- and within-cluster analyses were conducted. For the between-cluster analyses, the dimensional homogeneity of the skills between each item type was evaluated, meaning that the items in each skill category served as a separate AT and all remaining critical reading items served as a PT. The results are presented in Table 24. For sentence completion items, confirmatory DIMTEST rejected the null hypothesis for all three samples in word meaning and understanding sentences. For critical reading items, DIMTEST rejected the null hypothesis for all three samples in understanding sentences and understanding larger sections of text and one sample in analyzing purpose. Taken together, word meaning and understanding sentences are dimensionally distinct for sentence completion items, whereas understanding sentences and understanding larger sections of text are dimensionally distinct for critical reading items.

For the within-cluster analyses, the dimensional homogeneity of the skill within each item format area was evaluated where each skill category served as a separate AT and the remaining items within each item format area served as a PT. The results are presented in Tables 25a and 25b. For sentence completion and passage, DIMTEST rejected the null hypothesis for either zero or one sample across almost all categories, indicating that the skills are not dimensionally distinct within item format. The only exception was found for understanding larger sections of text within the critical reading items, where DIMTEST rejected the null hypothesis for Books 2a and 2c, if we interpret a p-value of 0.0511 as statistically significant.

Table 24

Confirmatory DIMTEST Results by Critical Reading Item Format and Skill Category

	Sentence Completion						*Critical Reading*					
	2a		*2c*		*5*		*2a*		*2c*		*5*	
Skill	*T*	*p*	*T*	*p*	*T*	*p*	*T*	*p*	*T*	*p*	*T*	*p*
Word Meaning	5.2510	0.0000	5.8057	0.0000	4.8734	0.0000	-0.0427	0.5170	-0.1366	0.5543	0.5912	0.2772
Understanding Sentences	3.3576	0.0004	2.4194	0.0078	4.0147	0.0000	1.9613	0.0249	2.4984	0.0062	2.7293	0.0032
Understanding Larger Sections	NA	NA	NA	NA	NA	NA	2.4921	0.0063	3.8248	0.0001	2.5599	0.0052
Analyzing Purpose	NA	NA	NA	NA	NA	NA	0.4368	0.3311	2.2251	0.0130	-0.2865	0.6128

Note: Each dimension served as a separate AT, with the remaining items serving as a PT. NA means that the cell contained 0 items.

Table 25a

Confirmatory DIMTEST Results by Critical Reading Skills Within Sentence Completion

	Word Meaning		*Understanding Sentences*		*Understanding Larger Sections*		*Analyzing Purpose*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	0.9588	0.1688	0.0796	0.4683	NA	NA	NA	NA
2c	0.7360	0.2309	0.6862	0.2463	NA	NA	NA	NA
5	1.0180	0.1543	0.8474	0.1984	NA	NA	NA	NA

Note: Each skill category served as a separate AT with the remaining items serving as a PT. NA means that the cell contained 0 items.

Table 25b

Confirmatory DIMTEST Results by Critical Reading Skills Within Critical Reading

	Word Meaning		*Understanding Sentences*		*Understanding Larger Sections*		*Analyzing Purpose*	
Book	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
2a	0.1913	0.4242	0.6408	0.2608	1.6339	0.0511	-0.089	0.5353
2c	-0.067	0.5268	1.0672	0.1429	2.2164	0.0133	1.3115	0.0948
5	1.0816	0.1397	1.0617	0.1442	1.0016	0.1583	-1.068	0.8573

Taken together, the between- and within-cluster dimensionality analyses indicate that word meaning and understanding sentences are dimensionally distinct for sentence completion items, whereas understanding sentences and understanding larger sections of text are dimensionally distinct for critical reading items when *both item formats* serve as the composite measure. The skills are not dimensionally distinct when either sentence completion or critical reading items serve as the conditioning variable.

The reading passages in the test specifications also produced dimensionally distinct clusters. Therefore, skills and reading passages were combined to produce distinct clusters that were tested for dimensional homogeneity. For these analyses, only between-passage comparisons were conducted because many of the cells contained a small number of items and, thus, provided an inadequate representation of the skills for

Table 26a

Confirmatory DIMTEST Results by Critical Reading Skills Between Reading Passages Using Book 2a

	Passage 1		*Passage 2*		*Passage 3*		*Passage 4*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Word Meaning	–	–	–	–	NA	NA	–	–
Understanding Sentences	–	–	NA	NA	4.9493	0.0000	2.5057	0.0061
Understanding Larger Sections	1.8683	0.0309	NA	NA	2.7770	0.0027	4.5915	0.0000
Analyzing Purpose	–	–	–	–	1.3522	0.0882	0.5749	0.2827
	Passage 5		*Passage 6*		*Passage 7*		*Passage 8*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Word Meaning	NA	NA	–	–	NA	NA	–	–
Understanding Sentences	–	–	–	–	0.2208	0.4126	6.1075	0.0000
Understanding Larger Sections	NA	NA	2.6082	0.0046	3.5328	0.0002	5.3539	0.0000
Analyzing Purpose	–	–	–	–	–	–	-0.2443	0.5965

Note: Each skill within a reading passage served as a separate AT, with the remaining items serving as a PT. NA means that the cell contained 0 items and a dash (–) indicates the cell only contain one item, therefore DIMTEST analyses could not be conducted.

Table 26b

Confirmatory DIMTEST Results by Critical Reading Skills Between Reading Passages Using Book 2c

	Passage 1		*Passage 2*		*Passage 3*		*Passage 4*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Word Meaning	–	–	–	–	NA	NA	–	–
Understanding Sentences	–	–	NA	NA	3.2068	0.0007	1.4111	0.0791
Understanding Larger Sections	2.6443	0.0041	NA	NA	1.2313	0.1091	4.9525	0.0000
Analyzing Purpose	–	–	–	–	2.3612	0.0091	0.7277	0.2334
	Passage 5		*Passage 6*		*Passage 7*		*Passage 8*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Word Meaning	NA	NA	–	–	NA	NA	–	–
Understanding Sentences	–	–	–	–	-1.2125	0.8873	6.1591	0.0000
Understanding Larger Sections	NA	NA	2.5663	0.0051	3.2844	0.0005	5.3516	0.0000
Analyzing Purpose	–	–	–	–	–	–	1.7925	0.0365

Table 26c

Confirmatory DIMTEST Results by Critical Reading Skills Between Reading Passages Using Book 5

	Passage 1		*Passage 2*		*Passage 3*		*Passage 4*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Word Meaning	–	–	–	–	NA	NA	–	–
Understanding Sentences	–	–	NA	NA	4.7217	0.0000	-0.4648	0.6790
Understanding Larger Sections	2.4838	0.0065	NA	NA	1.8786	0.0302	4.3749	0.0000
Analyzing Purpose	–	–	–	–	0.2500	0.4013	0.5135	0.3038
	Passage 5		*Passage 6*		*Passage 7*		*Passage 8*	
Skill	***T***	***p***	***T***	***p***	***T***	***p***	***T***	***p***
Word Meaning	NA	NA	–	–	NA	NA	–	–
Understanding Sentences	–	–	–	–	-0.0116	0.5046	5.4684	0.0000
Understanding Larger Sections	NA	NA	2.1970	0.0140	3.4633	0.0003	5.4728	0.0000
Analyzing Purpose	–	–	–	–	–	–	0.4142	0.6606

the short reading passages. The results are presented in Tables 26a to 26c. Across the 13 comparisons, confirmatory DIMTEST rejected the null hypothesis for either two or three samples in all comparisons except 5: analyzing purpose in passage 3, understanding sentences in passage 4, analyzing purpose in passage 4, understanding sentences in passage 7, and analyzing purpose in passage 8. These findings reveal that most of the skill categories in critical reading are dimensionally distinct across the eight reading passages.

Because the skills were dimensionally distinct between the two item formats, the factor loadings for each skill were estimated using NOHARM. In other words, each skill category for each item format defined the multidimensional structure. The 6-dimensional model provided adequate fit to the data across all three samples, as shown in Table 27. Tanaka's index was high and comparable across samples ranging from 0.985 to 0.988. The RMSR was low and comparable across samples, ranging from 0.005 to 0.006. The common factor parameter estimates for the Book 2a data are presented in Table 28. The correlations between the skill-by-item format dimensions were high, ranging from 0.891 to 1.000.

Table 27

NOHARM Fit Indices for Critical Reading Skill Category Across Item Formats and Reading Passages

	Fit Index	
Book	***Tanaka***	***RMSR***
2a	0.985	0.006
2c	0.986	0.006
5	0.988	0.005

Table 28

Six-Dimensional Solution for Critical Reading Skill Category Across Item Formats and Reading Passages Using Book 2a

Item	***Dimension***					
	1	***2***	***3***	***4***	***5***	***6***
1	**0.679**	0.0	0.0	0.0	0.0	0.0
2	**0.677**	0.0	0.0	0.0	0.0	0.0
29	**0.624**	0.0	0.0	0.0	0.0	0.0
46	**0.575**	0.0	0.0	0.0	0.0	0.0
28	**0.544**	0.0	0.0	0.0	0.0	0.0
6	**0.536**	0.0	0.0	0.0	0.0	0.0
5	**0.474**	0.0	0.0	0.0	0.0	0.0
30	**0.440**	0.0	0.0	0.0	0.0	0.0
45	**0.438**	0.0	0.0	0.0	0.0	0.0
8	**0.420**	0.0	0.0	0.0	0.0	0.0
25	0.0	**0.700**	0.0	0.0	0.0	0.0
26	0.0	**0.651**	0.0	0.0	0.0	0.0
44	0.0	**0.633**	0.0	0.0	0.0	0.0
3	0.0	**0.554**	0.0	0.0	0.0	0.0
4	0.0	**0.518**	0.0	0.0	0.0	0.0
27	0.0	**0.507**	0.0	0.0	0.0	0.0
7	0.0	**0.452**	0.0	0.0	0.0	0.0
48	0.0	**0.379**	0.0	0.0	0.0	0.0
47	0.0	**0.352**	0.0	0.0	0.0	0.0
49	0.0	0.0	**0.583**	0.0	0.0	0.0
21	0.0	0.0	**0.351**	0.0	0.0	0.0
54	0.0	0.0	**0.272**	0.0	0.0	0.0
65	0.0	0.0	0.0	**0.711**	0.0	0.0
23	0.0	0.0	0.0	**0.682**	0.0	0.0
66	0.0	0.0	0.0	**0.675**	0.0	0.0
19	0.0	0.0	0.0	**0.671**	0.0	0.0
60	0.0	0.0	0.0	**0.665**	0.0	0.0
61	0.0	0.0	0.0	**0.659**	0.0	0.0
58	0.0	0.0	0.0	**0.643**	0.0	0.0
33	0.0	0.0	0.0	**0.638**	0.0	0.0
35	0.0	0.0	0.0	**0.628**	0.0	0.0
22	0.0	0.0	0.0	**0.569**	0.0	0.0
16	0.0	0.0	0.0	**0.540**	0.0	0.0
11	0.0	0.0	0.0	**0.490**	0.0	0.0
64	0.0	0.0	0.0	**0.484**	0.0	0.0
18	0.0	0.0	0.0	**0.469**	0.0	0.0
57	0.0	0.0	0.0	**0.330**	0.0	0.0
38	0.0	0.0	0.0	0.0	**0.670**	0.0
51	0.0	0.0	0.0	0.0	**0.638**	0.0
63	0.0	0.0	0.0	0.0	**0.617**	0.0
17	0.0	0.0	0.0	0.0	**0.576**	0.0
12	0.0	0.0	0.0	0.0	**0.556**	0.0
36	0.0	0.0	0.0	0.0	**0.551**	0.0
50	0.0	0.0	0.0	0.0	**0.550**	0.0
42	0.0	0.0	0.0	0.0	**0.541**	0.0
55	0.0	0.0	0.0	0.0	**0.541**	0.0
14	0.0	0.0	0.0	0.0	**0.535**	0.0
10	0.0	0.0	0.0	0.0	**0.529**	0.0
62	0.0	0.0	0.0	0.0	**0.529**	0.0
13	0.0	0.0	0.0	0.0	**0.525**	0.0
40	0.0	0.0	0.0	0.0	**0.517**	0.0
53	0.0	0.0	0.0	0.0	**0.505**	0.0
39	0.0	0.0	0.0	0.0	**0.483**	0.0
56	0.0	0.0	0.0	0.0	**0.477**	0.0
52	0.0	0.0	0.0	0.0	**0.464**	0.0
31	0.0	0.0	0.0	0.0	**0.430**	0.0
9	0.0	0.0	0.0	0.0	**0.415**	0.0
43	0.0	0.0	0.0	0.0	**0.399**	0.0
20	0.0	0.0	0.0	0.0	0.0	**0.670**
59	0.0	0.0	0.0	0.0	0.0	**0.650**
34	0.0	0.0	0.0	0.0	0.0	**0.617**
41	0.0	0.0	0.0	0.0	0.0	**0.565**
37	0.0	0.0	0.0	0.0	0.0	**0.558**
15	0.0	0.0	0.0	0.0	0.0	**0.557**
24	0.0	0.0	0.0	0.0	0.0	**0.544**
67	0.0	0.0	0.0	0.0	0.0	**0.502**
32	0.0	0.0	0.0	0.0	0.0	**0.304**

Table 28 *(continued)*

Note: Dimensions 1 through 6 correspond to sentence completion items and passage 1 through 8 items, respectively.

Item	***Dimension Correlations***					
	1	***2***	***3***	***4***	***5***	***6***
1	1.000					
2	0.999	1.000				
3	1.104	1.109	1.000			
4	0.891	0.924	1.099	1.000		
5	0.902	0.924	1.116	0.990	1.000	
6	0.920	0.903	1.087	1.011	1.006	1.000

Summary of SAT Dimensionality Results

The purpose of this research report is to present the results from our analyses designed to identify dimensions on the mathematics and critical reading sections of the SAT that could promote diagnostic inferences about students' cognitive skills. To identify the dimensions that underlie student performance on the SAT, exploratory and confirmatory analyses were conducted. The outcomes from our dimensionality analyses in mathematics and critical reading are summarized in Tables 29 and 30, respectively.

Exploratory analyses served as a logical first step in our evaluation because the dimensionality of the SAT is not well defined. Moreover, the results from exploratory analyses would allow us to evaluate, using a data-driven approach, whether the SAT was measuring a multidimensional construct. DIMTEST, DETECT, and nonlinear factor analysis were used.

DIMTEST is a nonparametric statistical procedure that conducts a hypothesis test to assess the presence of multidimensionality. This procedure is based on Stout's (1987) concept of "essential unidimensionality," which holds when only one dominant dimension influences the examinees' performance on a set of test items. The DIMTEST results for the SAT revealed that at both the composite and section levels, the data contained more than one dimension.

DETECT is a nonparametric dimensionality assessment procedure designed to determine the dimensional structure underlying test data. This procedure identifies mutually exclusive, dimensionally homogeneous clusters of items that are most easily identified when the data approximate simple structure. The DETECT results for the composite test revealed a weak amount of dimensionality but with interpretable simple structure for the two test sections across two of the three samples. The DETECT results for mathematics also indicated a weak amount of dimensionality but with the data displaying complex structure across all three samples. Similarly, for critical reading, the DETECT results indicated a weak amount of dimensionality with the data displaying

Table 29

Summary of Dimensionality Results in Mathematics

	Book		
Mathematics	***2a***	***2c***	***5***
Exploratory			
DIMTEST	Reject H_o	Reject H_o	Reject H_o
No. of Dimensions[1]			
DETECT	4	5	4
Scree	2	2	2
Kaiser	2	2	2
MAP	2	2	2
Confirmatory[2]			
Content Area (4)	4*	2	2
Item Type (3)	1	0	1
Skills (4)	4*	1	2
Skills Between Content Area (15)	2	2	1
Skills Within Algebra (4)	1	2	2
Skills Within Arithmetic (4)	3	2	3
Skills Within Geometry (4)	1*	1	1
Skills Within Miscellaneous (4)	0	0	0

[1] From these results, we concluded that the number of dimensions was 2. The promax-rotated factor solution for the 2-dimensional model is presented in Table 7.

[2] The total number of identifiable dimensions for each organizing principle is listed in parentheses.

*The observed *p*-values for some of the comparisons in these analyses were very close to the critical value of 0.05. Therefore, these results are considered statistically significant in our summary (see Footnote 11).

Table 30

Summary of Dimensionality Results in Critical Reading

	Book		
Critical Reading	***2a***	***2c***	***5***
Exploratory[1]			
DIMTEST	Reject H_o	Reject H_o	Reject H_o
No. of Dimensions			
DETECT	4	4	5
Scree	3	3	3
Kaiser	3	3	3
MAP	3	2	2
Confirmatory[2]			
Item Format (2)	2	2	2
Reading Passage (8)	7	7	7
Item Format and Reading Passage (9)	8	7	8
Skills (4)	2	3	2
Skills Between Item Formats (6)	4	5	4
Skills Within Item Format (8)	0	0	3
Skills Between Reading Passages (13)	9	9	8

[1] From these results, we concluded that the number of dimensions was 3. The promax-rotated factor solution for the 3-dimensional model is presented in Table 8.

[2] The total number of identifiable dimensions for each organizing principle is listed in parentheses.

complex structure across all three samples. When the Nandakumar and Ackerman (2004) algorithm was applied to the mathematics and critical reading items to identify homogeneous dimensional clusters, highly inconsistent clusters were identified across our three validation samples, suggesting DETECT may be ineffective at determining the dimensional structure when the data display complex structure.

To generate a more consistent dimensional structure, nonlinear factor analysis was used (McDonald, 1967, 1997, 1999), as implemented with the computer program NOHARM (Fraser, 1988). We fit a variety of multidimensional models to the data, as guided by well-known decision rules for identifying the number of dimensions. For mathematics, a 2-dimensional model provided the most parsimonious fit to the data. The correlation between the dimensions was 0.686. For critical reading, a 3-dimensional model provided the most parsimonious fit to the data. The correlations between the dimensions ranged from 0.655 to 0.726. From our exploratory analyses, we conclude that the SAT mathematics and critical reading items measure more than one dimension. However, the nature of these dimensions is unclear because exploratory analyses lack a substantive framework to guide the interpretative process. To overcome this limitation, confirmatory analyses were also conducted.

Confirmatory analyses rely on the logic of hypothesis testing where substantively meaningful hypotheses are first specified and then these hypotheses are tested statistically. For dimensionality assessment, the hypotheses specify the structural characteristics of the data. These characteristics can be specified because substantive explanations are available to describe the structure of the data. Thus, the purpose of the analysis is to test these substantive explanations.

Two different College Board organizing principles were used to guide the confirmatory analyses: test specifications and cognitive skill categories. Each principle allowed us to first specify a model describing the dimensional structure of the data and then test the model because the purpose of the analysis is to evaluate the hypothesized structure. Our organizing principles were selected to shed light on the potential diagnostic value of the SAT. Therefore, dimensions were selected and tested because they may help identify the cognitively based constructs associated with diverse test performance.

Two categories in the College Board test specifications were used in the *first confirmatory analysis* because they highlight important item features according to the test developers, they are readily available and easy to use, and they can guide interpretation. It would also be easy to create subscores based on these test specification categories.

The mathematics test specification categories were content area and item type. Using content area as the organizing principle, confirmatory DIMTEST rejected the null hypothesis for two or more samples in arithmetic, geometry, and miscellaneous indicating three of the four content areas were dimensionally distinct from one another. Using item type as the organizing principle, confirmatory DIMTEST only rejected the null hypothesis for the nonroutine/insightful items, suggesting that item type was not a meaningful organizing principle for identifying dimensionally distinct clusters in mathematics. NOHARM was also used to estimate the parameters for the content-based 4-dimensional model. The fit indices for this model were comparable across samples, indicating good model fit. The correlation between the dimensions were also very high, ranging from 0.946 to 1.000, indicating that these dimensions were closely related to one another.

The critical reading test specification categories were item format and reading passage. Using item format as the organizing principle, confirmatory DIMTEST rejected the null hypothesis for all three samples indicating the sentence completion and critical reading items were dimensionally distinct from one another. Using reading passage as the organizing principle, confirmatory DIMTEST rejected the null hypothesis for all passages across samples, except passage 5, indicating the items associated with the reading comprehension passages also measured distinct dimensions. Because the item format and reading passage categories produced dimensionally homogeneous item clusters, these categories were crossed to evaluate the dimensionality of the resulting item format-by-reading passage item clusters. Again, confirmatory DIMTEST rejected the null hypothesis for almost all item clusters, indicating that the dimensions associated with sentence completion and reading passage were, for the most part, dimensionally distinct from one another. The parameters for the 9-dimensional model were estimated with NOHARM. The correlations between the dimensions were moderate to large, ranging from 0.525 to 1.000.

The *second confirmatory analysis* used the College Board's cognitive skill categories as the organizing principle to guide the assessment of dimensionality. These skills were first described in 2004 at a symposium conducted at the annual meeting of the National Council on Measurement in Education in San Diego where College Board researchers presented papers on their work related to extracting detailed diagnostic information from the SAT. Because only preliminary empirical evidence has been collected to date, our confirmatory dimensionality analyses were designed to assess the dimensions associated with the cognitive skill categories.

In mathematics, the skills included applying basic mathematics knowledge, applying advanced mathematics knowledge, managing complexity, and modeling and insight. Confirmatory DIMTEST rejected the null hypothesis for two samples in basic mathematics and all three samples in modeling insight, suggesting that basic mathematics and modeling and insight were dimensionally distinct skills in mathematics. Mathematics skills were also crossed with mathematics content areas because three of the content areas were found to be dimensionally distinct in the test specification confirmatory analyses. For the between-cluster analyses, confirmatory DIMTEST rejected the null hypothesis in two or more samples for modeling and insight in algebra and for basic mathematics in arithmetic. For the within-cluster analyses, confirmatory DIMTEST rejected the null hypothesis for two or more samples in managing complexity in algebra; basic mathematics, advanced mathematics, and modeling and insight in arithmetic; and basic mathematics in geometry. The parameters for each content-based multidimensional skills model were estimated with NOHARM. The correlations between the dimensions were high, ranging from 0.817 to 1.000.

In critical reading, the skills included determining word meaning, understanding sentences, understanding larger sections of text, and analyzing authors' purposes, goals, and strategies. Confirmatory DIMTEST rejected the null hypothesis for all samples in word meaning and understanding larger sections of text, revealing that these two categories represent dimensionally distinct skills.

Critical reading skills were crossed with item format because item format yielded dimensionally distinct clusters in the test specification confirmatory analyses. As in mathematics, both between- and within-cluster analyses were conducted. For the between-cluster analyses, confirmatory DIMTEST rejected the null hypothesis for all three samples in word meaning and understanding sentences using the sentence completion items. Confirmatory DIMTEST also rejected the null hypothesis for all samples in understanding sentences and understanding larger sections of text for the reading passage items. For the within-cluster analyses, DIMTEST failed to reject the null hypothesis for either two or three samples across all skill categories, except one, indicating the majority of skills were not dimensionally distinct within any item format category.

The reading passages in the test specifications also produced dimensionally distinct clusters, therefore skills and reading passages were combined to produce distinct clusters that were tested for dimensional homogeneity. For these analyses, only between-passage comparisons were conducted because many of the cells contained a small number of items. Confirmatory DIMTEST rejected the null hypothesis for two or more samples in 8 of 13 comparisons, indicating that most of the skill categories in critical reading were dimensionally distinct across the 8 reading passages. The parameters for the passage-based multidimensional skills model were estimated with NOHARM. The correlations between the dimensions were high, ranging from 0.891 to 1.000.

Conclusions and Future Research Directions

The results of this study allow us to conclude that there is a *multidimensional basis for test score inferences* on the mathematics and critical reading sections of the SAT. Results from the exploratory analyses indicate that the data are multidimensional, as mathematics displayed two dimensions and critical reading displayed three dimensions. The correlations between the dimensions were moderate in both test sections.

Results from the confirmatory analyses also indicate that the SAT data are multidimensional. However, the sources designed to account for these dimensions in the test specification and skill categories were not found consistently across the samples in our analyses. This outcome suggests the College Board's organizing principles *approximate*, but may not completely represent, the multidimensional structure of the data. As a result, we cannot claim that the multidimensional structure of the SAT supports diagnostic inferences about students' cognitive skills; we can only claim that the mathematics and critical reading sections of the SAT measure more than one dimension. In mathematics, three of the four content areas were found to be dimensionally distinct. These content-based dimensions were highly correlated. Two of the four skills were also deemed to be dimensionally distinct. When the dimensionality of the skills within each content area was evaluated, the number of detectable dimensions ranged from zero in miscellaneous to three in arithmetic. The skill dimensions within each content area were highly correlated. In critical reading, all item format and reading passage clusters were found to be dimensionally distinct, except one. The correlations among these dimensions were moderate to high. Two of the four skills were found to be dimensionally distinct, as in mathematics. When the skills were evaluated across the item formats and reading passages, 8 of the 13 skills were dimensionally distinct. The skills were highly correlated with one another.

Our general conclusions are consistent with the findings presented in the literature on the dimensionality of the SAT. For example, Cook et al. (1988), using confirmatory factor analysis to assess the dimensionality of the SAT verbal, concluded it was "slightly multidimensional." Our results reveal that the critical reading section, which is now the primary verbal reasoning measure on the SAT, is clearly multidimensional. Multiple dimensions were relatively easy to identify and they appeared consistently across samples. Lawrence and Dorans (1987), using exploratory and confirmatory factor analysis to assess the dimensionality of the SAT mathematical section, concluded it was unidimensional. However, they also claimed that exploratory analyses of the item-level data revealed a "slight departure from unidimensionality." Our results, conducted exclusively at the item level, indicate that mathematics is multidimensional. However, the dimensions in mathematics were more difficult to identify and were less replicable across samples compared to critical reading. The organizing principles used in mathematics also yielded fewer dimensionally distinct clusters compared to critical reading (see Tables 29 and 30).

Future Direction #1: Single Versus Multiple Cognitive Skills Per Item

Multiple subject-matter experts coded the SAT items using skills in four cognitive categories. These experts were asked to identify every skill required to solve each item, but also to specify the primary skill which was described as the most important skill, required to solve the item. For this report, only the data associated with the primary skill were used. We noted earlier, however, that these instructions have implications for confirmatory dimensionality analyses. Coding all skills implies that the data may display a complex factor structure, especially if the SAT items elicit multiple skills. Coding primary skills, on the other hand, implies the data will display simple structure because only one cognitive skill will be associated with each item. Taken together, these two instructions, and the item coding they produce, will affect the dimensionality analyses because they prescribe either complex or simple structure.

In the current study, primary skills were modeled because the final codes were available from the College Board (these final ratings, in fact, were established by test development experts at the College Board). During the next stage of our dimensionality research, the complete rater codes will be evaluated. These codes may produce a complex dimensional structure that, in turn, may yield a more complete model (and, hence, a better description) for the SAT data. However, these analyses should only proceed if the items are coded consistently across the expert raters. To evaluate rater consistency in the skills coding, we will compute the judge discrepancy from the median (JDM_j) and we will conduct a series of generalizability studies (g-studies).

JDM_j and range are used in reliability studies to assess rater consistency relative to the median ratings. JDM_j is computed as

$$JDM_j = \sum_{k=1}^{K} \left| X_{kj} - Md_k \right|,$$

where j is the rater number, k is the item number, X_{kj} is the jth rater's rating for item k, and Md_k is the median rating for item j. Small JDM_j suggests that rater discrepancy is small. The range also can be used to assess rater discrepancy using the formula $R = X_{jH} - X_{jL} + 1$.

G-studies will also be conducted using the primary and complete rater codes. For the primary skills codes, an $i \times r$: Item by Rater—fully crossed random effect model will be assessed, where the population of items and raters is seen as infinite. The variance components obtained from this model will be interpreted to evaluate rater consistency. The variance components include $\hat{\sigma}^2_i$, $\hat{\sigma}^2_r$ and $\hat{\sigma}^2_{ir,\, e}$. In this model, if the residual is small, then the rater variance component will allow us to make inferences about rater consistency where a small rater variance component implies strong rater agreement. On the other hand, if the residual is large, then additional facets will be modeled. For example, if the residual in mathematics is large, then we may add content as a facet because three of the four content areas were found to be dimensionally distinct. In this case, an $(i{:}H) \times r$: (Item within Content Area by Rater)—nested mixed effect model will be used, where content is viewed as a fixed facet. Because this design would be unbalanced, a multivariate g-study would be conducted. A similar approach will be used with the primary skill categories in critical reading where our initial design will focus on the an $i \times r$: Item by Rater—fully crossed random effect model.

For the complete rater codes, a different set of g-studies will be conducted. Initially, for both mathematics and critical reading, we will conduct $i \times r \times C$: (Item by Rater by Skill Category)—fully crossed mixed effect model where skill category is a fixed facet because only four cognitive skills are available. Again, the magnitude of the variance components can be used to make inferences about rater consistency. When the residual is large, additional facets will be identified and added to the design in an attempt to decrease the error term. By conducting different reliability analyses (i.e., JDM_j and g-studies) and different types of g-studies, we will be able to evaluate rater consistency and to identify important sources of rater inconsistency.

Future Direction #2: Modeling Strategy Use

The subject-matter experts were also asked to identify the skills required to solve the SAT items by envisioning a representative group of examinees who all used the same strategy to solve each item. These instructions were designed to circumvent potential problems associated with strategy diversity that can occur when examinees are perceived to use different strategies to solve an item. Unfortunately, this approach can be problematic because the experts can make an erroneous judgment about strategy use or students can use multiple strategies. The extent of this problem is not clear on the SAT items, and neither is its impact on the analysis and the interpretation of the cognitive skill categories. However, measurement specialists must recognize that strategy diversity is unavoidable when students solve items on tests because multiple strategies are typically used by students to solve problems and that a student may even apply a different strategy to the same problem because, at any one point in time, a student possesses a repertoire of problem-solving strategies. Therefore, complex test-taking performance should not be simplified by describing this performance with a single strategy or process unless it can be demonstrated that an item only elicits a single strategy or process.

As an alternative, subject-matter experts can be asked to identify the skills required to solve the SAT items, specify the items that measure these skills, and then attempt to connect the expert results to the student results using experimental and nonexperimental studies. The alternative approach has some appeal because it does not require the experts to make inferences about students' cognitive item-level performance. Rather, it only requires the experts to make inferences about their own cognitive performance, given the salient demand characteristics of the items.

Regardless of whether students or experts are consulted, a *cognitive model* of test performance is postulated in both cases. Within the information-processing perspective in cognitive psychology, the use of the term "cognitive model" has specifically been used to describe interconnected mental processes that encode, translate, manipulate, and generate information under specific conditions. Cognitive models have currency not only among psychologists but also, increasingly, among educators. Within the last 50 years, there has been increasing interest in using cognitive models to develop better large-scale achievement and aptitude tests (Cronbach and Meehl, 1955; Embretson, 1999; Haladyna and Downing, 2004; Irvine and Kyllonen, 2002; Mislevy, 1996; National Research Council, 2001; Pellegrino, Baxter, and Glaser, 1999; Snow and Lohman, 1993; Sternberg, 1984). Test specialists are particularly interested in using cognitive models to develop large-scale diagnostic tests of student learning and using scores from these tests to make specific inferences about what students know and can do outside of the testing situation. Cognitive models are assumed to be useful in developing diagnostic tests because items can be created to measure distinct processes of the learning cycle (as delineated in the cognitive model). Furthermore, it is believed that when test items are created to measure specific cognitive processes, inferences about student performance will also be specific to the processes measured. Thus, the prospect of using cognitive models in educational measurement is promising, except for one issue: The term *cognitive model* is used broadly in educational measurement, leading to potential confusion about the different types of models that potentially exist for organizing and understanding test performance.

During the next stage of our research, we will describe and give examples of at least three kinds of cognitive models that are used in educational measurement, including the strengths and limitations of each model in elucidating test performance (Leighton, 2004). The first cognitive model we will discuss is of *domain mastery*, which is generated from content experts to establish an extensive set of interconnected knowledge and skills that are believed to conceptualize expertise within a content domain. The second cognitive model we will discuss is of *test specifications*, which can be generated to outline the specific achievement area to be tested and to provide precise guidelines for designing or selecting a representative sample of items from the content domain during test construction. The third cognitive model we will discuss is of *task performance*, which can be generated from students to validate or verify the actual set of interconnected knowledge and skills that students use to respond correctly to test items within a content domain. Our goal in discussing these distinct cognitive models is to provide a framework that can be used by the College Board when organizing evidence and when making inferences about students' cognitive skills.

Future Direction #3: Scalability of Diagnostic Dimensions

Once the dimensions underlying test performance are identified, the scalability of these dimensions must be determined. A scalable score-related multidimensional structure requires statistical evidence demonstrating the subscores derived from the dimensions are invariant over time, across test forms, and across different groups of examinees (e.g., males and females). A test scalability analysis should also address the following questions (Luecht, 2005): (1) Does the data fit the proposed factor structure well enough to explain all nonrandom sources of variance? (2) Can competing models and factor structures be ruled out? (3) Is there sufficient evidence to support educational and policy-related decisions that may arise from the subscores? and (4) Are the diagnostic subscores meaningful and useful to stakeholders (e.g., students, parents, teachers)? In short, a test scalability analysis requires comprehensive studies designed to validate the inferences associated with the cognitive skills the subscores are deemed to measure. These types of studies are required to explicitly connect the SAT multidimensional structure, as identified in our analyses, with cognitive diagnostic test score inferences.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22* (3), 37–53.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and Psychometric Analysis of Analogical Problem Solving*. New York: Springer-Verlag.

Bolt, D. (in press). Limited and full information estimation of IRT models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary advances in psychometrics*. Mahwah, NJ: Erlbaum.

Chaffin, R., & Pierce, L. (1987, April). *Types of verbal analogy relations and academic skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics*, *13*, 19–43.

Cronbach, L. J., & Meehl, P. E. (1955). *Psychological Bulletin*, *52*, 281–302.

Diones, R., Bejar, I. I., & Chaffin, R. (1996, January). The dimensionality of responses to SAT analogy items. *ETS Research Report*. Princeton, NJ: ETS.

Dorans, N., & Lawrence, I. (1999). The role of the unit of analysis in dimensionality assessment. *ETS Research Report* (RR-99-14). Princeton, NJ: ETS.

Douglas, J., Kim, H., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992 administrations*. LSAC Research Report Series. Law School Admission Council, Inc.

Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso (Ed.), *Handbook of applied cognition* (pp. 629–660). Chichester, England: John Wiley & Sons.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.

Froelich, A. G. (2000). *Assessing unidimensionality of test items and some asymptotics of parametric item response theory*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Gierl, M. J. (1997a). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research*, *91*, 26–32.

Gierl, M. J. (1997b). An investigation of the cognitive foundation underlying the rule-space model (Doctoral dissertation, University of Illinois at Urbana-Champaign, 1996). *Dissertation Abstracts International*, *57* (08), 5351-B. (University Microfilms No. AAC 97-02524)

Gierl, M. J. (2004, August). *Identifying cognitive dimensions that affect student performance on the New SAT: A discussion paper*. Paper presented to the College Board, New York, New York.

Gierl, M. J. (2005). Using a dimensionality-based DIF analysis paradigm to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*, 3–14.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement*, *40* (4), 281–306.

Gierl, M. J., Bisanz, J., & Li, Y. Y. (2004, April). *Using the multidimensionality-based DIF analysis framework to study cognitive skills that elicit gender differences*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement*, *38*, 164–187.

Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice*, *19*, 34–44.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretative guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*, 145–220.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *Spring*, 17–27.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: NELS:88 science achievement. *American Educational Research Journal*, *32*, 555–581.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*, 1–14.

Huff, K. (2004, April). *A practical application of evidence-centered design principles: Coding items for skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

Kupermintz, H., Ennis, M., Hamilton, L., Talbert, J. E., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: II. NELS:88 mathematics achievement. *American Educational Research Journal*, *32*, 524–554.

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS:88 mathematics achievement to 12th grade. *American Educational Research Journal*, *34*, 124–150.

Lawrence, I., M., & Dorans, N. J. (1987, April). *An assessment of the dimensionality of SAT-Mathematical*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*, 1–10.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

Liu, J., Feigenbaum, M., Walker, M. E. (2004, April). *New SAT and new PSAT/NMSQT spring 2003 field trial design*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M. (2005). Review of *"Identifying cognitive dimensions that affect student performance on the new SAT, Technical Report #1: Dimensionality Results" by M. Gierl, X. Tan, & C. Wang*. New York, NY: College Board.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometric Monographs, No. 15.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, *28*, 99–117.

Nandakumar, R., & Ackerman, T. A. (2004). Test modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 93–105). Thousand Oaks, CA: Sage.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41–68.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64* (4), 575–603.

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS:88 science achievement to 12th grade. *American Educational Research Journal*, *34*, 151–173.

O'Callaghan, R. K., Morley, M. E., & Schwartz, A. (2004, April). *Developing skill categories for the SAT math section*. Paper presented at the annual meeting at the National Council on Measurement in Education, San Diego, CA.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, *24*, 307–353.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2* (1), 13–43.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.

Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 1–17). Hillsdale, NJ: Erlbaum.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Sternberg, R. J. (1984). What cognitive psychology can and cannot do for test development. In B. S. Plake & J. Mitchell (Eds.), *Social and technical issues in testing: Implications for test construction and usage* (pp. 39–60). Hillsdale, NJ:Erlbaum.

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

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

Tanaka, J. S. (1993). Multifaceted concepts of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

Tate, R. L. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181–211). Mahwah, NJ: Erlbaum.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17* (2), 89–112.

VanderVeen, A. (2004, April). *Toward a construct of critical reading for the new SAT*. Paper presented at the annual meeting at the National Council on Measurement in Education, San Diego, CA.

Walker, C., & Gierl, M. J. (2004). The effect of model misspecification on exploratory and confirmatory multidimensional IRT models. *Applied Psychological Measurement*. Revised manuscript resubmitted for review.

Zhang, J. & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 321–249.

Appendix A

Skill 1

Title: Applying basic mathematics knowledge

Description: Student can apply mathematics content and procedures that are usually learned in Algebra I or before. This content includes basic arithmetic concepts, operations, procedures, translation between verbal expressions and equations, graphing, reading graphs, definitions, and terminology. Basic geometry is also included in this skill category.

Skill 2

Title: Applying more advanced mathematics knowledge

Description: Student can apply mathematics content and procedures taught in high school geometry and Algebra II. Student can also apply some more advanced or complex applications of the knowledge obtained in Algebra I and earlier. This skill category covers procedures, translation between verbal expressions and equations, graphing, definitions, and terminology beyond that covered in the previous skill category.

Skill 3

Title: Managing complexity

Description: Student can keep track of a great deal of information or proceed through many steps to solve a question with a higher level of complexity. Correctly answering a question that is multistep may require the student to have the initiative to continue on for each step in the path to the solution. In some cases, that initiative may be required to take the first step of plunging in and attempting the solution of a nonroutine problem.

Skill 4

Title: Modeling and insight

Description: Student can use insight and modeling to answer a question with a higher level of difficulty. Insights are those realizations that often seem easy when someone else points them out, but are difficult to see by oneself. A student who can model or "create a representation" has the ability to create equations for word problems when that involves more than just rote translation. They can also create or add to graphs or figures to solve problems. Part of the skill of creating representations is deciding what kind of representation would be useful for solving a problem.

Appendix B

Skill 1

Title: Determining the meaning of words

Description: Student determines the meaning of words in context by recognizing known words and connecting them to prior vocabulary knowledge. Student uses a variety of skills to determine the meaning of unfamiliar words, including pronouncing words to trigger recognition; searching for related words with similar meanings; and analyzing prefixes, roots, and suffixes.

Skill 2

Title: Understanding the content, form, and function of sentences

Description: Student builds upon an understanding of words and phrases to determine the meaning of a sentence. Student analyzes sentence structures and draws on an understanding of grammar rules to determine how the parts of speech in a sentence operate together to support the overall meaning. Student confirms that his or her understanding of a sentence makes sense in relationship to previous sentences, personal experience, and general knowledge.

Skill 3

Title: Understanding the content, form, and function of larger sections of text

Description: Student synthesizes the meaning of multiple sentences into an understanding of paragraphs or larger sections of texts. Student recognizes a text's organizational structure and uses that organization to guide his or her reading. Student can identify the main point of, summarize, characterize, or evaluate the meaning of larger sections of text. Student can identify underlying assumptions in a text, recognize implied consequences, and draw conclusions from a text.

Skill 4

Title: Analyzing authors' purposes, goals, and strategies

Description: Student identifies an author's intended audience and purpose for writing. Student analyzes an author's choices regarding content, organization, style, and genre, evaluating how those choices support the author's purpose and are appropriate for the intended audience and situation.

CollegeBoard
connect to college success™

www.collegeboard.com