CollegeBoard
connect to college success™

Research Notes

Office of Research and Analysis RN-31, January 2007

Comparability of Scores on the New and Prior Versions of the SAT Reasoning Test™

Jennifer L. Kobrin and Gerald J. Melican

Introduction

In March 2005, the SAT Reasoning Test™ was revised to strengthen its alignment with curriculum and instructional practices in high school and college. An important assumption underlying the changes to the SAT® was that scores on the SAT Reasoning Test would be fully comparable to and interchangeable with scores on the prior test, the SAT I: Reasoning Test. This assumption was essential for allowing test users to track score trends across years and test administrations, and to enable colleges and universities to treat scores from all SAT administrations equally when making admissions decisions.

When the SAT was revised in 2005, the test specifications were changed. Analogy items were removed, and paragraph reading items were added to the critical reading (formerly verbal) section. Third-year college-preparatory math content was added to the mathematics section, and quantitative comparison items were removed. Furthermore, a writing section was added, changing the total testing context. When a test undergoes changes in specifications and/or administrative conditions, it is necessary to assess whether these changes have a significant impact on the constructs measured by the test, and it is important to assess the degree of equatability as these changes may alter the meaning of scores.

The College Board carried out extensive research prior to, and following, the introduction of the SAT Reasoning Test to evaluate the comparability of the SAT Reasoning Test and prior SAT I: Reasoning Test scores, and to ensure that test-takers and users could be confident that scores from the two test versions could be interpreted and used in the same way. To meet this requirement, it is necessary that both tests measure the same constructs, and that the scores from both the SAT Reasoning Test and prior SAT I: Reasoning Test can be equated.

The intention of this research note is to synthesize the research to date addressing the construct comparability of the SAT Reasoning Test and prior SAT I: Reasoning Test (Oh and Sathy, 2006), and the series of research studies addressing the equatability and subpopulation invariance of the SAT Reasoning Test and prior SAT I: Reasoning Test performed by Liu, Feigenbaum, and Dorans (2003, 2005); Dorans, Cahn, Jiang, and Liu (2006); Dorans, Liu, Cahn, and Jiang (2006); Jiang, Liu, Cahn, and Dorans (2006); and Liu, Jiang, Dorans, and Cahn (2006). It is noted that these two topics—construct comparability and equitability—are intertwined, because one of the major assumptions underlying equating is that the two versions of a test are measuring the same constructs (Angoff, 1971/1984). To clarify the connection between the two, a brief description of the concept of equating follows.

Dorans (2000) described three levels of score linkage—equating, scaling, and prediction—that fall on a continuum ranging from strict exchangeability of scores (equating) to only an association between the scores. Construct similarity plays an important role in determining the degree of linkage that can be achieved, but statistical indices in conjunction with rational considerations are also needed. The term "exchangeability" in the strictest sense means that the scores from two tests can be used interchangeably; that a test-taker would, on average, obtain the same score on one test form as on the other test form after equating has been performed. The phrase "on average" is required because tests are not perfectly reliable and an individual's score, even on the same form, will vary slightly should he or she take the test again. Similarly, the SAT I: Reasoning Test and the ACT

have concordance tables that allow a limited comparison of the scores from both tests (Dorans, 1999). The scores are not, however, equated because they do not measure exactly the same constructs, nor would an individual be indifferent as to whether they were administered the SAT or the ACT. Therefore, the scores from the ACT and SAT cannot be used interchangeably.

Equating makes the strongest claims concerning the relationship between scores on two test forms and, therefore, the requirements are the most stringent. Again, when the scores from two versions of the same test are equated, it means that a test-taker would be indifferent to which version was administered (Lord, 1980). Lord specified four requirements that must be met for two tests to be equated. Among these are construct comparability (the two tests must measure the same construct) and subpopulation invariance (the equating transformation should be invariant across subpopulations). An additional requirement described by Dorans and Holland (2000), and consistent with Angoff's (1971/1984) and Lord's equity requirements, is the "equal reliability requirement," that is, tests that measure the same construct but which differ in reliability cannot be equated in the strictest sense because the test-taker would not be indifferent to the version administered.

These assumptions or requirements for equating impose heavy responsibility on the test developer. In order to meet the assumptions necessary for equating, each version of the test must be assembled to extremely detailed content and statistical specifications. A theoretic goal of the test assembly process is to create all test forms with an equal measure of difficulty, a goal that is impossible in practice (Lawrence and Schmidt, 2001). Equating is used to adjust for minor differences in test difficulty from form to form. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) state that "a clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably" (Standard 4.10). The *Standards* also indicate that it is essential to demonstrate that the theoretical assumptions are met. As long as the assumptions are met, the methods used to equate the scores for the SAT are in full accordance with the *Standards* and, therefore, test users can be confident that the meaning of an SAT score remains the same from year to year regardless of the variation in difficulty across test forms, or the variation in ability across test administrations (Educational Testing Service, 2005; Lawrence and Schmidt, 2001). That is, the data collection, analyses, and interpretation required for equating the SAT meet or exceed the requirements of the *Standards* as long as the underlying assumptions are met. The rest of this paper is dedicated to demonstrating that the assumptions of content comparability and subpopulation invariance are met in comparing the SAT Reasoning Test with its predecessor.

Research Studies to Address Construct Comparability

The SAT was designed to measure critical thinking and reasoning skills that test-takers develop over time, both in and out of school. The revisions to the SAT in 2005 were not intended to change the overall constructs measured by the test, although these revisions did align the test more closely to curriculum and instruction in high school and college. Nevertheless, any time a test undergoes changes, it is important to assess whether the changes alter the constructs that are measured by that test.

In the spring of 2003, two years before the reconstructed SAT Reasoning Test debuted, an extensive field trial was conducted to evaluate the content, timing, and statistical specifications for this new test (Liu and Feigenbaum, 2003). One important purpose of the field trial was to assess score comparability and construct continuity of the SAT Reasoning Test and its previous version, the SAT I: Reasoning Test. To this end, one of the field trial test administration designs had participants take one SAT I: Reasoning Test section in its entirety and its SAT Reasoning Test counterpart (e.g., old verbal and new critical reading sections or old mathematics and new mathematics sections).

The field trial results indicated that the new critical reading and mathematics sections had similar reliability and standard errors of measurement to the prior verbal and mathematics sections. The correlation between the scores on the new critical reading and old verbal sections was .91, and the correlation between the scores on the new and old mathematics sections was .92. The specified test difficulty for the new sections was maintained, and the mean and standard deviation for the new sections were very close to those of the prior sections. The attenuated correlations of the old test sections with the new test sections were virtually 1.00. Furthermore, data showed that the new critical reading and mathematics sections were comparable to the prior test

sections in speededness (i.e., test-takers had enough time to complete the new test, as they did with the old test).

Oh and Sathy (2006) analyzed the field trial data to assess the construct comparability of the SAT Reasoning Test and the prior SAT I: Reasoning Test by exploring the dimensional structure of the two test versions using linear factor analysis of item parcel data, or minitests made up of small collections of nonoverlapping items thought to measure the same underlying dimension(s). Exploratory factor analysis was used to examine the factor structure of the item parcels for the two different tests. All of the items on both tests were analyzed together to allow examination of construct continuity between the two tests. The results from the exploratory factor analyses were used to guide model specification for confirmatory factor analyses.

The results suggested that the changes to the SAT had very little impact on the dimensional structure of the test. For the verbal/critical reading section, multiple models were evaluated, and models including or excluding analogy items were very similar, suggesting that the elimination of analogy items has not considerably affected the critical reading construct. Despite several changes on the mathematics section from the SAT I: Reasoning Test (e.g., elimination of quantitative comparison items, addition of third-year mathematics material), the analyses revealed that both mathematics sections are quite similar, essentially measuring a single mathematics reasoning dimension.

Research Studies to Address Subpopulation Invariance

Dorans and Holland (2000) consider subpopulation invariance to be the most important requirement of equating two tests, because fulfilling this requirement will also imply that the tests measure the same thing and are equally reliable. They suggest that if two tests measure different constructs and/or are not equally reliable, then the equating results will not be invariant for some subgroups. Similarly, Angoff (1971/1984) emphasized that the two criteria—construct comparability and population invariance—"go hand in hand" (page 86). The reasoning is straightforward: if two tests measure different constructs, the equating functions would probably be different for different groups. Having the same or nearly the same equating functions for two groups does not, in itself, establish construct comparability; having different equating functions would imply the underlying constructs are not comparable. Thus, population invariance speaks directly to construct comparability.

Studies of the subpopulation invariance of the SAT I: Reasoning Test and the SAT Reasoning Test were conducted using field trial data and actual operational SAT data. Using the 2003 SAT field trial data, Liu et al. (2003, 2005) explored the score invariance of both test constructs with respect to gender. They examined specifically whether SAT Reasoning Test scores could be mapped onto the prior SAT I: Reasoning Test scales so that the characteristics of the new test scores matched those of the prior test very closely. Several indices were used to evaluate the linking of the scores on the two tests. The standardized root mean square difference (RMSD) was used to quantify the differences between the female and male equating functions and the total population equating functions at a given score value. The root expected mean square difference (REMSD) was used to summarize overall differences between the equating functions. The RMSD and REMSD were compared to the score differences that matter (DTM) proposed by Dorans and Feigenbaum (1994) in the context of linking the SAT Reasoning Test to the SAT I: Reasoning Test.[1] The DTM indicates whether the score resulting from an equating function would change the reported score. If the unrounded scaled score resulting from two separate linkings differ by fewer than five points, the scores would be rounded to the same reported score (Liu, Jiang, Dorans, and Cahn, 2006).

Based on these three indices, the findings provided evidence that subpopulation invariance based on gender on the new critical reading and mathematics sections was achieved. That is, the equating functions obtained for females and males were very similar to those obtained for the total group. Liu et al. (2003, 2005) also compared the standardized difference (or effect size) in scores between females and males on the new critical reading and mathematics sections compared to the old verbal and mathematics sections and found very little difference, providing further evidence of subpopulation invariance on the new tests. Finally, the relative score distributions of females and males on the SAT Reasoning Test were very similar to those on the SAT I: Reasoning Test.

Because the field trial was not an actual SAT administration, and the students participating in the field trial

[1] A thorough discussion of the RMSD, REMSD, and DTM, including their equations, is found in Dorans and Holland (2000).

were not completely representative of the regular SAT-taking population, it was very important to replicate the population invariance analyses using operational SAT data. Statisticians at Educational Testing Service performed a series of invariance studies based on gender and ethnicity with operational SAT data (Dorans, Cahn, Jiang, and Liu, 2006; Dorans, Liu, Cahn, and Jiang, 2006; Jiang, Liu, Cahn, and Dorans, 2006; Liu, Jiang, and Cahn, 2006; Liu, Jiang, Dorans, and Cahn, 2006). Separate equating functions were derived for the total group and separately for female, male, African American, Asian American, Latino, and white test-takers for the SAT Reasoning Test critical reading section to the previous SAT I: Reasoning Test verbal section, and the SAT Reasoning Test mathematics section to the previous SAT I: Reasoning Test mathematics section.

The results suggest that subpopulation invariance based on gender was achieved on both the old verbal/new critical reading and old mathematics/new mathematics equating functions. Subpopulation invariance was not found for racial/ethnic groups on any of the equating functions performed in this study. The equating differences of the new critical reading section to the old verbal sections were all smaller than DTM for all scores in the middle range of the SAT scale (from about 350 to 650). In both the low and high end of the scale, African American, Asian American, and Latino test-takers would have received lower scores if the equating function based only on their respective subgroup (rather than the total group's equating function) was used to produce the scores. In the equating of the new mathematics section to the previous mathematics section, the differences for all subgroups were again smaller than DTM in the middle range of the scale, but were larger at the low and/or high end of the scale. While the differences tended to be near the DTM, the maximum difference was approximately 15 at a few score points. However, it must be emphasized that the sample sizes for every nonwhite racial/ethnic group were extremely small at the points where large differences were found. It is not uncommon to find instability at high and low ends of the scale because there are relatively few test-takers at these extremes. Again, for most of the equating links that were studied, African American and Latino test-takers would have obtained lower scores if the equating function based only on their respective subgroup was used. It is important to note that the sample sizes for the minority ethnic groups were very small, and replication of this research is needed with larger subgroups to confirm or disconfirm the results of this study.

Summary

Several research studies have been completed to substantiate that scores on the SAT Reasoning Test are fully comparable with scores on the prior SAT I: Reasoning Test. The research to date has indicated that the changes to the SAT did not significantly alter the constructs that are measured by the test, and that the scores on the SAT Reasoning Test can be equated to scores on the SAT I: Reasoning Test without adverse effects on any test-takers. Although the research has not completely established subpopulation invariance with respect to racial/ethnic groups, the disparities occurred only at the very high and low ends of the scale where there were very small sample sizes, and the differences in these score intervals tended to favor African American and Latino test-takers (that is, they received higher scores on the SAT Reasoning Test when it was equated to the SAT I: Reasoning Test than they would have if the equating was performed using subgroup-specific functions).

In summary, the research conducted to date provides evidence for comparability of scores on the SAT Reasoning Test and the SAT I: Reasoning Test. Future research will replicate the analyses described in this research note to assess their generalizability. Another indication of score comparability is the extent to which SAT Reasoning Test scores predict college outcomes compared to the prior test. An upcoming, large-scale validity study of the SAT Reasoning Test will ascertain the correlation of the SAT Reasoning Test scores with first-year college grades, and offer comparison with the validity coefficients based on the former test.

Jennifer L. Kobrin is a research scientist at the College Board. Gerald J. Melican is chief psychometrician at the College Board.
The authors would like to thank Rick Morgan and Krista Mattern for their helpful comments on earlier versions of this research note.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Angoff, W.H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service. (Reprinted from *Educational Measurement* [2nd ed.], by R.L. Thorndike, Ed., 1971, Washington, DC: American Council on Education.)

Dorans, N.J. (1999). *Correspondences between ACT and SAT I scores* (College Board Research Report No. 99-1, ETS Research Report No. 99-2). New York: The College Board.

Dorans, N.J. (2000). *Distinctions among classes of linkages* (College Board Research Note RN-11). New York: The College Board.

Dorans, N.J., Cahn, M., Jiang, Y., & Liu, J. (2006). *Score equity assessment of transition from SAT I verbal to SAT math: gender* (SR-2006-63). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Dorans, N.J., & Feigenbaum, M.D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT®. In I.M. Lawrence, N.J. Dorans, M.D. Feigenbaum, N.J. Feryok, & N.K. Wright, *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10). Princeton, NJ: Educational Testing Service.

Dorans, N.J., & Holland, P.W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306.

Dorans, N.J., Liu, J., Cahn, M., & Jiang, Y. (2006). *Score equity assessment of transition from SAT I verbal to SAT critical reading: gender* (SR-2006-61). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2005). *Test analysis report* (SR-2005-123). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Jiang, Y., Liu, J., Cahn, M., & Dorans, N.J. (2006). *Score equity assessment of transition from SAT I verbal to SAT critical reading: ethnicity* (SR-2006-62). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Lawrence, I.M., & Schmidt, A.E. (2001). *Ensuring comparable scores on the SAT I: Reasoning Test* (College Board Research Note RN-14). New York: The College Board.

Liu, J., & Feigenbaum, M.D. (2003). *Prototype analysis of spring 2003 new SAT field trial* (SR-2003-69). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Liu, J., Feigenbaum, M.D., & Dorans, N.J. (2003). *Equatability analysis of the new SAT to the current SAT I* (SR-2003-73). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Liu, J., Feigenbaum, M.D., & Dorans, N.J. (2005). *Invariance of linkings of the revised 2005 SAT Reasoning Test to the SAT I: Reasoning Test across gender groups* (College Board Research Report No. 2005-6). New York: The College Board.

Liu, J., Jiang, Y., & Cahn, M. (2006). *New SAT score equity assessment across gender groups and ethnic groups* (SR-2006-23). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Liu, J., Jiang, Y., Dorans, N.J., & Cahn, M. (2006). *Score equity assessment of transition from SAT I verbal to SAT math: ethnicity* (SR-2006-64). Unpublished statistical report. Princeton, NJ: Educational Testing Service.

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

Oh, H., & Sathy, V. (in press). *Construct comparability and continuity in the SAT.* Unpublished statistical report. Princeton, NJ: Educational Testing Service.

Office of Research and Analysis
The College Board
45 Columbus Avenue
New York, NY 10023-6992
212 713-8000

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 5,000 schools, colleges, universities, and other educational organizations. Each year, the College Board serves seven 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.

CollegeBoard
connect to college success™