NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between NAEP and SAT®

Rebecca Moran, Andreas Oranje, and Dave Freund, ETS

As part of the National Assessment Governing Board's efforts to enable NAEP to report on the preparedness of U.S. twelfth graders for postsecondary education or entry into job training programs, studies were planned to statistically relate performance on NAEP with results from other assessments that serve as indicators of preparedness for college entry, course placement, and entry into the workforce (National Assessment Governing Board, 2009). Such statistical relationships between tests are accomplished using procedures referred to as test equating or scale linking. The purpose of scale linking is to express the results of one test on the scale of another test. In the context of NAEP 12th grade preparedness research, the goal of the statistical relationship studies is to enable interpretation of NAEP results in terms of other assessments and the preparedness benchmarks associated with those assessments.

This document describes the data and procedures used to establish and evaluate statistical relationships between the NAEP 12th grade reading and mathematics assessments and the critical reading and mathematics scales of the SAT. Design and analysis decisions were made with input from a group of technical advisors assembled by the National Assessment Governing Board ("Governing Board"). Among the technical advisors were experts in psychometrics and test linking, large-scale assessment, educational policy, and industrial-organizational psychology.

Data

This study used data from students who were sampled and assessed in NAEP 12th grade reading or mathematics in 2009 and had also taken the SAT by June 2009.

NAEP Samples

From late January through early March of 2009, NAEP assessments in reading, mathematics, and science were administered to samples of grade 12 students. Because the interest was in linking NAEP to the SAT, which has no science component, data for students assessed in NAEP science were not used in the study. Whereas grade 12 NAEP assessments in reading and mathematics were previously administered to only nationally representative samples, 11 states (Arkansas, Connecticut, Florida, Idaho, Illinois, Iowa, Massachusetts, New Hampshire, New Jersey, South Dakota, and West Virginia) volunteered to participate in a twelfth-grade state pilot program in 2009. As a result, larger samples of students from public schools in each of the 11 pilot states (roughly 3,000 students per subject) were drawn and augmented the nationally-representative samples of public and private school students. Overall, approximately 52,000 twelfth-graders were assessed in NAEP reading and 49,000 were assessed in NAEP mathematics in 2009. Sampling weights were used to appropriately represent these larger state samples in national-level analyses.

Matching NAEP and SAT Test Takers

The Governing Board entered into an agreement with the College Board to obtain SAT scores for public school students who were in twelfth grade in 2009 and had taken the SAT by June 2009. The SAT data were matched, using identifiers provided by College Board, to performance records of students who participated in the 2009 NAEP grade 12 assessments in reading and mathematics. The focus on public school students reduced the NAEP samples to 49,000 students assessed in reading and 46,000 assessed in mathematics.

The process of matching SAT scores to NAEP participants was carried out through an agreement between the Governing Board and the National Center for Education Statistics (NCES) to have NAEP contractors Westat and ETS conduct the preparedness research work on behalf of the Governing Board. This agreement involved the NAEP contractors working with College Board to match the needed SAT scores for students in the NAEP samples. A process for matching the student records was developed to protect students' identity and confidentiality. Confidentiality of SAT scores was assured through the assignment of a pseudo ID for students taking the SAT and using that pseudo ID as a way to transfer SAT scores from College Board to ETS. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS. Via the pseudo ID, ETS subsequently matched SAT scores to NAEP files without requiring access to any Personally Identifiable Information (PII) data from the College Board. Student data was limited to questionnaire responses, SAT scores, and the pseudo ID. SAT scores were provided for all SAT administrations through June of 2009. SAT scores were matched for 33% of the students in the 2009 grade 12 reading and mathematics samples (including students in the 11 states that participated in the grade 12 pilot state assessment), resulting in 16,200 students for reading and 15,300 students for mathematics. This match rate compares favorably to the national SAT participation rate of approximately 36% of public school students.

SAT Scores

For each student in the matched, or linking, sample, scores were available from one or more administrations of the SAT, which included separate scores for critical reading and mathematics. The scale scores for each section range from 200 to 800 in 10-point increments. Some, but not all, students also had scores for the writing section of the SAT, which was introduced in 2005, but those scores were not used in this study. Each student's critical reading and mathematics scores were summed to form a composite SAT score. The critical reading and mathematics scores from each student's highest composite SAT score were used in this study because these are the SAT scores most likely considered in college admissions.

NAEP Scores

Students sampled for participation in NAEP are assessed in only one assessment subject. Consequently, each student in the linking sample had SAT scores in both critical reading and mathematics but results for only one NAEP assessment, either reading or mathematics. The grade 12 NAEP reading scale ranges

from 0 to 500 and the grade 12 mathematics scale ranges from 0 to 300. NAEP scales are subject-specific and their associated metrics are arbitrary.

NAEP does not report or even compute individual student scores. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population subgroups, estimated directly using marginal estimation methods (for a comprehensive description of NAEP estimation procedures, the interested reader should refer to Mislevy, Beaton, Kaplan, & Sheehan, 1992). Reporting statistics of interest for the relevant subgroups (e.g., average scale scores) are often based on computations involving plausible values associated with each assessed student as a matter of convenience. Plausible values are reflections of the relationship between student groups or other variables (e.g., SAT score) and proficiency, as defined and estimated in the model. Relationships between student groups or otherwise variables and proficiency that are not explicitly estimated in the model, are not reflected in the plausible values. This means that plausible values have very specific uses and will under most circumstances provide biased results if used outside what they were defined to represent.

An example of this is the plausible values for either reading or mathematics that were used to report the operational 2009 NAEP results in those subjects. These reflected the relationship of the underlying ability with the large set of group-defining variables that were modeled to support NAEP's extensive reporting goals (i.e., those derived from responses to student and teacher questionnaires in addition to demographic variables). However, the relationship between NAEP and SAT was not estimated and, therefore, not reflected in those plausible values. Subsequently, a model was defined to estimate the relationship between NAEP and SAT and, for convenience, an appropriate reflective new set of plausible values was drawn to facilitate reporting. As described by Mislevy et al. (1992), "Unless a characteristic were incorporated into the marginal analysis from which plausible values were constructed, it would generally not be recovered correctly from analyses of the completed data sets" (p. 147). Therefore, the NAEP plausible values used in this study were derived from a latent regression model that included students' gender, race/ethnicity, state, and SAT scores as population characteristics. SAT scores were included as linear and quadratic main and interaction effects.

Identification and Removal of Outliers

There are important differences in the purposes and reporting goals of the SAT and NAEP assessments. The SAT is a high-stakes test that is taken by a self-selected group of college-bound students. NAEP, as provided by law, estimates the proficiency of relevant population subgroups based on representative samples of students in the nation. Along with the differences in reporting goals, there are also appropriately different levels of measurement precision at the individual student level. SAT scores are reliable measures at the student level. NAEP does not produce individual student-level scores; rather, it is focused on estimating group level results. The procedures used in this study needed to appropriately accommodate this fact. Yet another factor affecting the establishment of a valid and interpretable statistical relationship between NAEP and SAT is the difference in students' motivation for doing well on the respective tests. Since extant data were used for this study, there was no way to adjust for this difference. However, the characteristics of the linking sample were closely examined with an eye

towards screening out any cases that would distort the results (Feuer, Holland, Green, Bertanthal, & Hemphill, 1999).

The linking samples used in the study are a higher-performing subset of the 2009 NAEP national public samples and the 2009 SAT cohort. The average NAEP reading score for the linking sample was 0.42 standard deviations higher than the national public score, and the average NAEP mathematics score for the linking sample was 0.50 standard deviations higher than the national public score. Average SAT scores for the linking sample were 0.13 and 0.11 standard deviations higher than the overall mean for public school students in the 2009 SAT cohort of college-bound seniors (The College Board, 2009).

Initial examination of the joint distribution of NAEP and SAT performance using scatter plots revealed potential outlier cases (e.g., students with relatively high SAT performance and low NAEP performance). Standardized residuals from robust regression (Huber, 1973) were used to identify approximately 1.1% of cases in the reading linking sample and 0.8% of the mathematics linking sample as outliers. These outliers were excluded from the final linking samples and therefore not used in subsequent analyses.

Final Linking Samples

For establishing a statistical relationship between NAEP reading and SAT critical reading, the final linking sample consisted of 16,100 students. The correlation between scores the two reading scales was r = 0.74. The linking sample for NAEP mathematics and SAT mathematics contained 15,200 students, and the correlation between scores on the mathematics scales was r = 0.91.

Method

SAT Preparedness Benchmarks

Central to the purpose of the NAEP 12th grade preparedness statistical relationship studies is relating NAEP performance to preparedness benchmarks that have been established for other assessments. Preparedness benchmarks for the SAT were calculated as the SAT score corresponding to a 65% probability of earning a first-year college grade-point average of 2.67 (B-) or better (Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011). The resulting preparedness benchmark score was 1550 on the SAT composite, with 500 as the benchmark on the individual critical reading and mathematics scales.

The critical reading benchmark score aligns closely with the current overall national average SAT score, and the mathematics benchmark score falls approximately 0.10 standard deviations below the SAT national average score. Wyatt et al. (2011) reported that fifty percent of the 2010 SAT cohort of college-bound seniors met the preparedness benchmark for critical reading and 54% met the mathematics benchmark.

Types of Statistical Relationships

Different classes of statistical relationships have been established between various tests, and the distinctions correspond to the extent to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer et al., 1999; Holland & Dorans, 2006). In describing the purpose of the NAEP 12th grade preparedness statistical relationship studies, the Governing Board's Technical Panel on 12th Grade Preparedness Research acknowledged that the strongest linking relationship between two tests, equating, is not an option for relating NAEP to SAT and other assessments because they do not share the same content specifications, reliability, or difficulty (National Assessment Governing Board, 2009). Nonetheless, the Technical Panel recommended using the "strongest feasible form of linking" to establish statistical relationships between NAEP and other assessments (National Assessment Governing Board, 2009, p. 7).

Two types of statistical linking were considered in this study: concordance and projection. Concordance establishes a symmetric relationship between the score distributions such that each score on one test has a corresponding score on the other. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric and, in and of themselves, do not assume or result in one-to-one score correspondences. Concordance assumes and requires a much stronger relationship than projection.

The Governing Board's goal is to establish a separate statistical relationship for each subject domain; that is, to separately link NAEP reading to SAT critical reading and NAEP mathematics to SAT mathematics.

The relationship between NAEP reading and SAT critical reading ($r = 0.74$) is not sufficiently strong to support a concordance, given that a generally accepted minimum correlation for concordance is $r = 0.87$ (Dorans, 1999; Dorans & Walker, 2007). Consequently, projection is a more appropriate method based on weaker assumptions for establishing a statistical relationship between the NAEP and SAT reading assessments. The correlation between the NAEP and SAT mathematics assessments ($r = 0.91$), on the other hand, is high enough to support a concordance. However, because it may not be advisable to use two different methods (with different assumptions and implications for interpretation) for the linking of reading and mathematics, results from both projection and concordance were produced for both subject areas.

Smoothing

To produce smooth distributions for use in subsequent analyses, bivariate loglinear presmoothing (Holland & Thayer, 2000) was applied to the joint NAEP-SAT score distribution for each subject area. In doing so, the NAEP sampling weights were used to appropriately represent the population of interest. For both subjects, the loglinear smoothing preserved the first 5 moments of the NAEP distribution, 8 moments of the SAT distribution, and 4 cross-moments. This loglinear smoothing model resulted in the

smallest value of the Akaike Information Criterion (AIC) statistic (Moses & von Davier, 2006) relative to models with more or fewer moments.

An important tool for evaluating statistical links between tests is sensitivity analysis, which examines the extent to which the relationship is subgroup invariant (i.e., does not change across subgroups). To provide data for evaluating the subgroup invariance of the NAEP-SAT relationships, separate smoothed joint distributions were produced for 8 mutually exclusive subgroups: White males, White females, Black males, Black females, Hispanic males, Hispanic females, Asian/Pacific Islander males, and Asian/Pacific Islander females. This permitted separate linkings to be established for these subgroups, in addition to a linking for the overall sample for each subject. It should be noted that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial volatility across student groups for parts of the scale could be quite harmless if sufficiently far away from the location of the benchmark point.

Statistical Projection

For both the reading and mathematics linking samples, the probabilities from the smoothed joint distributions were used to create projection tables containing conditional cumulative distributions of NAEP proficiencies for SAT scores. Separate overall and subgroup-level projections were produced. The range of possible NAEP scores at or above the SAT preparedness benchmark (500 on the critical reading and mathematics scales) were of primary interest. For each subject area, the projected conditional distributions were used to identify the NAEP scale scores associated with 50%, 67%, and 80% of students scoring at or above the SAT benchmark score of 500.

Table 1 shows the NAEP scale scores identified from the projection analysis in each subject area. The proximity of each resulting score to the NAEP *Proficient* cut point in each subject area is also noted. The subgroup-specific projections differed from one another, more so for reading than for math. Across the 8 subgroups studied, the NAEP reading scale scores associated with 50% of students attaining the SAT benchmark ranged from 290 to 318, a difference of 0.74 standard deviation units.

Concordance

An equipercentile concordance of NAEP and SAT scores in each subject area was established using the smoothed joint distribution for each linking sample. SAT critical reading and mathematics scores are reported in 10-point increments along a 200-800 scale. Given that the overall standard deviation is 100, each increment represents 0.10 of a standard deviation unit. For comparability, the NAEP distributions were divided into 3-point increments, which represent roughly 0.10 of the overall standard deviation for the reading and mathematics scales. Separate overall and subgroup-level concordances were produced for each subject area. Based on these concordances, the NAEP scale scores corresponding to the SAT preparedness benchmark scores were identified (table 2). Again, the results show variation in the subgroup-specific concordances, with greater variation for reading than for math.

Summary of Results

The intent of this study was to statistically relate NAEP and SAT and use that relationship to identify a reference point or range on the NAEP 12th grade reading and mathematics scales associated with the College Board's preparedness benchmarks on the SAT reading and mathematics measures. The resulting possible reference points are listed in tables 1 and 2, and table 3 displays the percentages of students in the overall 2009 NAEP twelfth-grade sample (from both public and private schools) who attained the potential NAEP reference points derived in this study. However, important limitations must be kept in mind when deciding on how to use the results of this study to report on 12th graders' preparedness. The statistical relationships established between NAEP and SAT for both mathematics and reading are not invariant across major population subgroups. And whereas the relationship between NAEP mathematics and SAT mathematics is sufficiently strong to support using NAEP 12th grade mathematics for reporting on preparedness, the weaker relationship between NAEP and SAT reading requires additional investigation and evaluation to determine how or if preparedness can be reported for NAEP 12th grade reading. The general approach adopted by the Governing Board for the 12th grade NAEP preparedness program of research calls for an overall analysis of results across four sets of studies: content alignment, judgmental standard setting, a survey of higher education institutions, and statistical relationships. The evidence from the composite set of studies is being evaluated to determine if the evidence across studies is mutually confirmatory.

References

The College Board (2009). *College-Bound Seniors 2009: Total Group Profile Report*. New York: Author.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report*.

National Center for Education Statistics (2010). *The Nation's Report Card: Grade 12 Reading and Mathematics 2009 National and Pilot State Results* (NCES 2011-455). Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Wyatt, J., Kobrin, J., Wiley, A., Camara, W. J., & Proestler, N. (2011). *SAT benchmarks: Development of a college readiness benchmark and its relationship to secondary and postsecondary school performance* (Research Report No. 2011-5). New York: The College Board.

Table 1: Results from NAEP-SAT Statistical Projection Analysis: Grade 12 NAEP Reading and Mathematics Scale Scores associated with SAT Preparedness Benchmark Scores

Percentage of students scoring at or above 500 on SAT	NAEP Mathematics Scale Score		NAEP Reading Scale Score	
	Scale Score for the Total Group	Range of Scale Scores for Subgroups	Scale Score for the Total Group	Range of Scale Scores for Subgroups
50%	164	159 - 170	302	290 - 318
67%	169	164 - 175	313	302 - 328
80%	175	169 – 181	325	314 – 338

Note. The subgroups studied were White males, White females, Black males, Black females, Hispanic males, Hispanic females, Asian/Pacific Islander males, and Asian/Pacific Islander females. The cut score for the NAEP *Proficient* achievement level is 176 for grade 12 mathematics; the *Proficient* cut score for reading is 302.

Table 2: Results from NAEP-SAT Concordance Analysis: Grade 12 Reading and Mathematics Scale Scores associated with SAT Preparedness Benchmark Scores

SAT Subscore	NAEP Mathematics Scale Score		NAEP Reading Scale Score	
	Scale Score for the Total Group	Range of Scale Scores for Subgroups	Scale Score for the Total Group	Range of Scale Scores for Subgroups
500	165	162 - 168	303	296 - 313

Note. The subgroups studied were White males, White females, Black males, Black females, Hispanic males, Hispanic females, Asian/Pacific Islander males, and Asian/Pacific Islander females. The cut score for the NAEP *Proficient* achievement level is 176 for grade 12 mathematics; the *Proficient* cut score for reading is 302.

Table 3: Percentages of the 2009 NAEP Reading and Mathematics Samples (Public and Private Schools) Meeting Potential NAEP Preparedness Benchmarks from NAEP-SAT Statistical Relationship Analyses

Student Group	Mathematics					Reading				
	NAEP *Proficient* Cut Score	Linking: Concor-dance	Linking: Statistical Projection (50%)	Linking: Statistical Projection (67%)	Linking: Statistical Projection (80%)	NAEP *Proficient* Cut Score	Linking: Concor-dance	Linking: Statistical Projection (50%)	Linking: Statistical Projection (67%)	Linking: Statistical Projection (80%)
	176	165	164	169	175	302	303	302	313	325
Total	**26%**	**37%**	**39%**	**33%**	**27%**	**38%**	**37%**	**38%**	**27%**	**17%**
Male-White	35%	48%	49%	43%	36%	40%	38%	40%	27%	17%
Male-Black	6%	12%	13%	10%	7%	12%	11%	12%	6%	3%
Male-Hispanic	13%	22%	23%	19%	14%	18%	17%	18%	11%	5%
Male-Asian	52%	63%	64%	59%	53%	45%	44%	45%	34%	22%
Female-White	30%	45%	46%	39%	32%	53%	52%	53%	40%	26%
Female-Black	6%	12%	13%	10%	6%	22%	21%	22%	13%	7%
Female-Hispanic	9%	16%	17%	13%	9%	26%	25%	26%	16%	8%
Female-Asian	51%	63%	64%	59%	52%	53%	52%	53%	41%	29%

Note. The cut scores for the NAEP *Proficient* achievement level are provided for reference.