The College Board

Research Report No. 2003-5

Linking Scores from Tests of Similar Content Given in Different Languages: Spanish Language PAA™ and English Language SAT® I

Alicia S. Cascallar and Neil J. Dorans

College Board Research Report No. 2003-5
ETS RR-03-20

Linking Scores from Tests of Similar Content Given in Different Languages: Spanish Language PAA™ and English Language SAT® I

Alicia S. Cascallar and Neil J. Dorans

College Entrance Examination Board, New York, 2003

Alicia S. Cascallar was president, Assessment Group International, U.S.A., when this report was accepted for publication.

Neil J. Dorans is principal measurement statistician in the Center for Statistical Theory and Practice at Educational Testing Service.

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.

Additional copies of this report (item #998644) 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

An earlier version of this paper was presented at the *Establishing Linkage Between Test Scores: Practical Considerations* symposium, 1999 Annual Meeting of the National Council on Measurement in Education, in Montreal.

The authors thank Miriam Feigenbaum and Wendi Wright for their help with the data analysis. The authors thank Linda Cook, Larry Stricker, and Cathy Wendler for the their helpful comments.

When Alicia Schmitt Cascallar died on August 24, 2003, a productive professional career and a special life came to a premature end. Much of the collaborative work presented in this report occurred toward the end of Alicia's 15-year tenure at the Educational Testing Service, much of which was spent working on College Board testing programs. All those who had the good fortune of working with her will miss her considerateness, her professionalism, her integrity, her gentleness, and her class.

Contents

Tables

Abstract

Score linkages between the Verbal and Math sections of the SAT® I: Reasoning Test and the corresponding sections of the new version of a Spanish-language admissions test, the Prueba de Aptitud Académica® (PAA™), were investigated. A bilingual group design was employed. A language proficiency measure (ESLAT) was used to define the bilingual group and as a predictor variable. Prediction and scaling for concordance results were compared. Results indicated that for both single (PAA Verbal or PAA Math to the corresponding SAT I scores) and composite (PAA–V+M to SAT I–V+M and PAA–V+M+ESLAT to SAT I–V+M) score linkage, prediction is preferable to concordance. Comparison of prediction and concordance results for composite scores versus single construct scores indicates that when PAA Verbal is combined with PAA Math to form a composite, predictions of this composite are better than for Verbal alone but worse than predictions for Math alone.

Key words: SAT I, Prueba de Aptitud Académica (PAA), score linking, prediction, equating, concordance

Linking Scores

The purpose of the current study is to establish score linkages between the Verbal and Math sections of the SAT I: Reasoning Test and the corresponding sections of the new version of a Spanish-language admissions test, the Prueba de Aptitud Académica (PAA), which is administered in Puerto Rico and Latin America. Because the data used in the last concordance study (Angoff and Cook, 1988) was collected more than a decade ago with prior versions of both the SAT I and PAA, new information about the relationship between the SAT I and the new PAA was needed. This study compares results of scaling for concordance and prediction methods to establish linkages between SAT I and PAA scores, and it builds upon work reported in Schmitt, Dorans, Magrina, and Cook (1998).

Hambleton (1993) and Sireci (1996) both offer a number of reasons why researchers are interested in cross-cultural and crosslinguistic assessment. Among the reasons they have given for the rapidly increasing interest in testing in different languages are: (1) to enhance fairness in assessment by allowing examinees to choose the language they wish to be tested in; (2) to facilitate comparative studies across national, ethnic, and cultural groups; and (3) to facilitate comparison of achievement of students in different countries, who receive instruction in different languages.

Ideally, those interested in linking assessments that have been translated into different languages and are given to monolingual examinees in their own language would like to be able to compare the skills and abilities of examinees taking the different assessments as though the scores obtained on the assessments were entirely interchangeable (equated). However, this ideal situation is often difficult to obtain because data collected in crosslinguistic linking studies are typically inconsistent with the requirements of score equating.

Sireci (1996a) provides an overview of technical issues associated with linking tests used in crosslinguistic assessments. Sireci begins his review by noting that some practitioners believe that simply translating a test from one language to another is a sufficient condition for crosslinguistic assessment. Sireci points out the fallacy in this line of reasoning by noting that unintended effects of the translation may produce items that differ in difficulty and other characteristics across the different languages (see Geisinger, 1994; Hambleton, 1993; Olmedo, 1981; Prieto, 1992).

According to Sireci, methods used to link assessments given in different languages fall into three design categories: (1) separate monolingual group designs; (2) bilingual group designs; and (3) matched monolingual designs. Separate monolingual group designs involve some procedure for developing "overlapping items," whereas the latter two designs have as their central requirement overlapping groups of examinees.

The PAA™

This study focuses on how best to develop a relationship between scores obtained on the Spanish language PAA and the English language SAT I: Reasoning Test (SAT I). The PAA is used for several purposes. It is administered to secondary students in Puerto Rico, Mexico, and other Latin American countries and used as an entrance test to universities and colleges in these countries. The PAA is also submitted as part of an admissions portfolio by some students who are interested in attending a mainland United States college or university. The English-speaking colleges and universities may use the PAA scores as an estimate of the Spanish-speaking student's potential for success in college. In addition, PAA scores are sometimes used to compare the verbal and math reasoning skills of English-speaking U.S. secondary-school students and Spanish-speaking students attending secondary school in Puerto Rico and Latin America.

It is important at this point to emphasize that the PAA is not a translation of the SAT I. Although the PAA is designed to measure the same constructs as the SAT I, the PAA contains different items and is developed independently of the SAT I. A decision was made by the College Board, early in the history of the PAA testing program, that because of the complexities and difficulties involved in translating a test from one language to the other, the "parallelism" between the two tests would be better preserved if each test was designed to measure the same construct, but in a different language.

A distinguishing feature of the PAA is that it is designed to be used in multiple Hispanic contexts. The various Hispanic populations, e.g., Mexican and Puerto Rican, differ from one another in much the same way as, say, residents of the United States and residents of Great Britain do; both groups speak English, but the nuances of the language differ in the different countries. Differential item functioning (DIF) analyses are carried out on the PAA to ensure that differential performance at the item level is taken into consideration across the different Hispanic populations.

Changes in PAA and SAT® I

Since the spring of 1994, new editions of the SAT I have been built to revised content and statistical specifications. (See Cook, 1995, for a description of the revisions to the SAT I.) In October 1996, a new version of the PAA was administered. This newer PAA was revised to include new item types and revised content and statistical specifications. Changes to the PAA paralleled the changes introduced to the SAT I. Specifically, as does the SAT I, the new PAA Verbal does not include antonyms and has a higher percentage of Verbal items that relate to critical reading passages (56 percent versus 31 percent). Consistent with the SAT I, in addition to the traditional multiple-choice items, the new PAA Math includes items where the examinee produces his or her own response. (See The College Board, 1995, for an extensive description of the changes to the new PAA.) One difference between the SAT I and the new PAA is that the SAT I allows the use of calculators in the Math section, while calculators are not permitted on the new PAA. Another difference is that the PAA is a right-scored test, while the SAT I is a formula-scored test.

Reasons for linking PAA and SAT I scores fall into two of the categories presented above. The first is fairness. The PAA offers those few students in Puerto Rico and Latin American countries applying for admission to English-speaking colleges and universities the opportunity to test in their native language. The second category is the comparison of international educational achievement. Comparisons are sometimes made of the verbal and math reasoning abilities of mainland United States secondary-school students with those of students attending secondary school in Puerto Rico and Latin American countries. The second reason is haunted, however, by a double dose of self-selection. The propensity to take the SAT I varies across states. Students from states on either coast are more likely to take the SAT than those in the middle, and this differential effect is less prominent among higher-scoring students than lower-scoring students. Likewise, a student from Puerto Rico is more likely to take the PAA than a student from Latin American countries. This double dose of self-selection confounds any comparisons of PAA and SAT I scores across country and for that matter within state.

Previous Research

Angoff and Modu (1973) carried out the first study conducted for the purpose of linking scores on the PAA to scores on the SAT. The results of the Angoff and Modu study were used to link scores on the PAA to scores on the SAT for about a 10-year period. Advances in technology, as well as the realization that it is good practice to repeat and revise the results of linking studies periodically, led to the repetition of the PAA–SAT linking in the study of Angoff and Cook (1988). The Angoff and Cook study followed the basic design of the earlier study, but replaced the classical test theory methodology with item response theory techniques. What follows is a brief critique of the two previous studies and a rationale for the study described in this report.

The two prior studies provided concordance tables between scores on the former versions of these tests (Angoff and Modu, 1973; Angoff and Cook, 1988). These linking studies used "common" items to adjust for any differences between the SAT and PAA groups. The "common" items were translations of PAA Spanish items to English and of SAT English items to Spanish. Although these studies were seen as groundbreaking efforts to develop cross-cultural comparison methodology, the authors of the second study cautioned that the lack of item parallelism between the populations for the verbal common item set and the lower ability level of the PAA sample "...cast [doubt] on the quality of any equating that could be carried out with tests in these two languages and with groups as different as these" (Angoff and Cook, 1988, p. 5).

Because of the problems identified by Angoff and Cook (1988), the current study approached the comparability of scores between the SAT I and the PAA from a different perspective. The purpose of providing concordance tables between the PAA and the SAT I is so that "direct comparisons could be made between subgroups of individuals of the two language-cultures who had taken only that test appropriate for them" (Angoff and Modu, 1973, p. 2). Furthermore, Angoff and Modu (1973, p. 2) add that "it was also expected that these conversion tables would help in the evaluation of the probable success of Puerto Rican students who were interested in eventually attending colleges on the mainland and were submitting PAA scores for admission." Linking procedures can only make the PAA and SAT I score distributions look the same. In addition, equivalence tables obtained though equating methods assume that the two tests are basically alternate forms representing the same construct. Even though the Spanish language PAA is an adaptation designed to parallel the English language SAT I, the items are developed for and pretested with Puerto Rican and other Hispanic populations in Latin America. Therefore, the PAA and SAT I are not alternate forms.

Prior research on other tests is also relevant to the present study. Regression methods were used by Alderman (1981) and by Boldt (1969). In the Alderman study, students were tested on the SAT I, PAA, the Test of English as a Foreign Language (TOEFL) and English as a Second Language Achievement Test (ESLAT). Language proficiency (TOEFL or ESLAT) was viewed as a moderator variable affecting the relationship between SAT I scores and PAA scores. Higher language proficiency as measured by these tests resulted in a stronger relationship between the SAT I and PAA scores. Similar results were found with a recent sample of PAA–SAT I test-takers (Schmitt, Dorans, Magrina, and Cook, 1998). These results underline the importance of using a measure of language proficiency when creating a prediction equation between the SAT I and PAA. Because all PAA examinees also take the ESLAT for admissions purposes, the ESLAT results can be used as such a predictor variable along with a measure of Spanish achievement.

Pennock-Roman (1995) investigated the relationship between scores on graduate-level admissions tests given in English and Spanish to a group of students who were more proficient in Spanish than English and concluded that proficiency in English contributed to the students' scores on the GRE Verbal, Math, Analytic, Psychology, and Biology tests. Pennock-Roman found that English proficiency contributed differently depending on the level of proficiency of the student and the content of the test. She noted the fact that it is quite possible for a talented second-language student to receive a below-average score on the GRE Verbal test simply because it took the student longer to read the passage.

The implications of the Pennock-Roman study for this PAA/SAT I linking study are that if the major purpose of the study is to provide a means for evaluating how well a student in Puerto Rico will do in a mainland college or university, then some measure of ability in English must be taken into consideration. In particular, if the goal is to predict SAT I scores from other scores, the use of ESLAT along with PAA will lead to better predictions than would be obtained with PAA alone. The methodology described in the next section of this paper attempts to take these implications into account.

Tests/Variables

SAT I (Verbal and Math)

The SAT I Verbal test has 78 items given in one 15-minute and two 30-minute sections: 19 sentence completion, 19 analogy, and 40 critical reading (51 percent of Verbal section). The SAT I Math test has 60 items given in one 15-minute and two 30-minute sections: 35 multiple-choice, 15 quantitative comparison, and 10 student-produced response (SPR).

PAA (Verbal and Math)

The PAA Verbal test has 60 items given in two 30-minute sections: 16 sentence completion, 16 analogy, and 28–30 critical reading (47–50 percent of Verbal section). The PAA Math test has 50 items given in one 30-minute and one 35-minute section: 30 multiple-choice, 10 quantitative comparison, and 10 SPR.

ESLAT

The ESLAT has 60 items given in one 45-minute section: 40 grammar and 20 reading.

Description of Data

The sample for this study is all available candidates who took the new PAA and ESLAT in Puerto Rico in test administrations during the period of October 1996 through June 1997 and who also took the SAT I from January 1996 to June 1997. Only the last score of the students who repeated the test within the defined period was considered in the analyses. Each student included in the study had scores from the following tests: (1) SAT I Verbal and Math, from the January 1996 to the June

1997 administrations; (2) PAA, from the October 1996 to the June 1997 administrations; and (3) ESLAT, from the October 1996 to the June 1997 administrations.

Although the sample of the Spanish PAA test-takers was constrained by the fact that the new PAA was only available from the October 1996 administration onward, 1,104 cases with complete data on SAT I, PAA, and ESLAT were obtained.

Design

The design for this study is a bilingual group design where the same group of students took the SAT I, the PAA, and the ESLAT. The ESLAT test was used to determine the degree of bilingualism of the matched group. Because the SAT I and PAA are not identical tests and because they are administered in two different languages, practice effects were not considered a problem. In addition, because the students taking all three tests were a self-selected group, they are assumed to have been equally motivated when taking each test.

Linking Analyses

Several authors, for example, Angoff (1971), Linn (1993), and Mislevy (1992), have discussed distinctions among different types of score linkages. Dorans (2000) presents a conceptual framework for linkages between scores and describes three types of score linkages: equating, scaling, and prediction. Dorans (2000) proposed three approaches to determining which type of linkage makes the most sense. In addition to the construct similarity between two tests, the strength of their empirical relationship and the constancy of the linkage relationship across populations are important factors in determining the type of linkage that can be achieved. Dorans (2000) states that statistical indices in conjunction with rational considerations are needed to make this determination.

Types of Linkages

The goal of equating (Holland and Rubin, 1982; Kolen and Brennan, 1995) is to produce scores that are fully exchangeable. Scores on tests of developed abilities and skills can be equated provided they are constructed to the same set of specifications, and a proper data collection design can be used to establish the equating relationship (Angoff, 1971).

A second type of linkage between two scales is scaling to produce a concordance table. Typically, the data collection designs and the statistical techniques used to establish an equating relationship are also used to establish a concordance relationship. The crucial distinction is that two sets of scores that have been placed on a common metric are considered equated only if they measure the same thing. For example, different editions of the SAT I are placed on the same scale with the intent of producing exchangeable scores. An examinee should be able to take any edition of the SAT I and obtain the same reported scores on the 200 to 800 scale within the precision (reliability) of the test. The same can be said for PAA scores. SAT I scores and PAA scores, however, are not exchangeable. They measure different, albeit highly related, constructs within their respective populations.

The third type of correspondence is prediction of an expected score. Whereas equating strives to achieve fully exchangeable scores and scaling for concordance matches distributions of scores, prediction is merely concerned with doing the best job possible of predicting one set of scores from another. The goal is to minimize the imprecision in the predictions of one score from one or more scores. A classic example of a prediction model is the estimation of grade point average from earlier grades and test scores. Unlike scaling for concordance and equating relationships, prediction relationships are not symmetric; the function that converts scores on test A to scores on test B is not the multiplicative inverse of the function that converts scores on test B to scores on test A.

According to Dorans (2000), there are three factors that indicate the degree to which we can achieve exchangeability through equating, concordance through scaling, or prediction. One is the logical evaluation of the similarity of the processes that produced the scores, or in other words, the content parallelism of the two tests. The second is the strength of the empirical relationship between the scores, typically measured by the correlation coefficient. And the third is the population invariance measured by standardized differences between the scaled scores of two groups. In order to evaluate the strength of the empirical relationship between two tests, Dorans (2000) proposed the use of a measure of uncertainty reduction. This index provides a measure of the statistical uncertainty that remains after inclusion of information from the predictor variable and uses the correlation coefficient r,

$$\textbf{reduction of uncertainty} = 1 - \text{coefficient of alienation} =$$

$$1 - \sqrt{(1 - r^2)}.$$

When r equals zero, the coefficient of alienation equals one, which means that there is a zero reduction in uncertainty about scores on the measure to be predicted. In contrast, a 100 percent reduction of uncertainty, represented by a zero coefficient of alienation, is achieved when $r = 1$.

A 50 percent reduction is halfway between 100 percent reduction ($r=1$) and 0 percent reduction ($r=0$). A correlation coefficient of at least .87 is needed to reduce the uncertainty, as measured in score units, of knowing a person's score by at least 50 percent. If a predictor cannot reduce uncertainty by at least 50 percent, it is unlikely that it can serve as a valid surrogate for the score it is supposed to predict.

Dorans (2000) explains that although the selection of any cut point is arbitrary, a 50 percent reduction in uncertainty is parallel to having a test with a reliability of .75. A reliability of .75 is equivalent to a correlation of .87 between true score and observed score. Reductions in uncertainty that fall short of 50 percent may be indicative of scores that are not equivalent or distributions of scores that can be used as if they are interchangeable.

Although the PAA is designed to measure the same constructs as the SAT I, the PAA contains different items; it is not a translation of the SAT I. In addition, based on results of a prior study where back-translations were used, the lack of parallelism between the populations for the verbal common items set and the lower ability level of the PAA sample led the authors to caution against the use of concordance results "…results cast [doubt] on the quality of any equating that could be carried out with tests in these two languages and with groups as different as these" (Angoff and Cook, 1988, p. 5).

This study compares the results of equipercentile scaling and prediction of expected score. Both single (PAA Verbal or PAA Math to the corresponding SAT I) and composite variables (PAA–V+M to SAT I–V+M and PAA–V+M+ESLAT to SAT I–V+M) equipercentile scalings were compared to single and multiple regression models used to predict SAT I scores from PAA scores and ESLAT.

Results

The Sample

In the self-selected sample of 1,104 Puerto Rican test-takers who took the SAT I, the PAA, and the ESLAT, several things are noteworthy, as can be seen in Table 1. First, this group is more able than the general PAA population of 45,810. The PAA Verbal mean is more than 100 points higher, 573 versus 470. The PAA Math mean is even more distant from that of the general population: 634 exceeds 487 by nearly 150 points. And ESLAT means are even further apart, more than 200 points apart, 654 in the sample of 1,104 test-takers versus 446 among the 41,243 PAA test-takers who took ESLAT. This selected group of test-takers is well above average on both PAA measures and ESLAT. In addition, they represent about only 2.5 percent of the full PAA group.

Table 1

Average Exam Performance of Self-Selected Sample and Full PAA Population

Test Score	*Self-Selected Sample*		*Full PAA Group*	
Spanish Language	*N*	*Mean (SD)*	*N*	*Mean (SD)*
ESLAT	1,104	654 (89)	41,243	446 (118)
PAA MATH	1,104	634 (102)	45,810	487 (112)
PAA VERBAL	1,104	573 (97)	45,810	470 (106)
English Language				
SAT I MATH	1,104	456 (97)	NA	NA
SAT I VERBAL	1,104	442 (101)	NA	NA

The Disparity in the SAT and PAA Scales

The second point to note is that the PAA scales have not been recentered, so they still contain a discrepancy between Verbal and Math average scores associated with the original SAT scales. This complicates interpretation of score distributions. Dorans (2002) describes the recentering of the SAT scales. One of the major reasons for recentering was to improve score interpretation. For years the average Verbal score was 50 points lower then the average Math score. Despite this disparity, many users of SAT scores, including students, presumed that the scales were such that the average Verbal score and average Math score were the same, and that they were equal to 500. Recentering corrected these interpretation problems. The problem still exists in the PAA data where there is a 61-point difference between Verbal and Math, as compared to the 14-point difference seen on SAT.

The third point is that the means on SAT I tests of these test-takers are around 450, which is about 50 points below the mean for the cohort of the SAT I. The SAT I cohort has mean scores that are also about 50 points higher than means for the Hispanic American population on the SAT I.

The fourth point to note is that the SAT I and PAA scales are different, despite their common endpoints. The PAA scales have 601 possible score points. Over 25 years ago, the SAT scales went from three active digits (i.e., scores ranged from 200 to 800 in steps of 1) to a three-digit scale in which the last digit is always 0. This reduced the number of score points from 601 to 61 and eliminated unreliable comparisons among nearly identical scores, bringing the scale in line with the scaling proscription against more score intervals than can be supported by the number of items on the test.

The next point to note is that the 1,104 bilingual students are very high scorers on the PAA and particularly on the ESLAT. About 50 percent of these students score above 580 on PAA Verbal, above 650 on PAA Math, and above 670 on ESLAT. In contrast, about 50 percent of these same students score below 430 on SAT I Verbal and below 450 on SAT I Math. Achieving a high score on PAA is much easier than achieving a high score on SAT I for this self-selected group of able examinees. This disparity in score performance poses a serious problem to any method that attempted to place these two measures on a common scale. If SAT I had not been recentered, this differential would be even more noticeable. The medians would have been around 350 for Verbal and 410 for Math on the original SAT scale.

The Importance of Language Proficiency

Finally, the correlations between SAT I and PAA/ESLAT in the self-selected sample are very noteworthy. The correlation between PAA Math and SAT I Math is .82, high but not high enough to meet the .87 level that corresponds to at least a 50 percent reduction in uncertainty. In addition, SAT I Math correlates .57 with ESLAT, a correlation that suggests that language proficiency affects prediction of SAT I Math in the bilingual group. Table 2 contains all the correlations.

The important role of language proficiency is even more apparent with ESLAT and Verbal scores. SAT I Verbal and PAA Verbal correlate .62. PAA Math has a lower correlation with SAT I Verbal, .60, but not by much. More significantly, both SAT I Math (.69) and ESLAT (.74) have noticeably higher correlations with SAT I Verbal than does PAA Verbal (.62). Building a concordance between the two verbal measures from these data is a questionable activity because language proficiency plays a critical role even in a group of ostensibly bilingual examinees. The more important question is: Can PAA Verbal add much to the prediction of SAT I Verbal scores beyond what ESLAT can do on its own?

TABLE 2

Correlations Among Test Scores in the Self-Selected Sample

Test Score	*ESLAT*	*PAA–M*	*PAA–V*	*SAT I–M*	*SAT I–V*
ESLAT	1.00	.51	.45	.57	.74
PAA MATH	.51	1.00	.61	.82	.60
PAA VERBAL	.45	.61	1.00	.56	.62
SAT I MATH	.57	.82	.56	1.00	.69
SAT I VERBAL	.74	.60	.62	.69	1.00

ESLAT, on the other hand, is not that highly related with either PAA score: .45 with Verbal and .51 with Math. Compare these to the .74 with SAT I Verbal and .57 with SAT I Math. Notice in particular the relationship that ESLAT has with the two Verbal scores, built to essentially the same content specifications. Clearly, the assessment of developed verbal ability is intimately tied to the language in which verbal ability is developed.

Using ESLAT as a Screening Variable

As noted above, these examinees were ostensibly English proficient. In fact some of them had fairly low ESLAT scores: 3.1 percent scored below the average ESLAT score of 446 obtained by the 41,243 students who took ESLAT. And 7 percent scored below the midpoint of 500 on the 200–800 ESLAT scale.

We decided to use ESLAT to screen out examinees with low English proficiency. We used a normative definition of sufficient English proficiency—scores that are in the top fifth of ESLAT scores for the full ESLAT test-taking population. This operational definition of English proficiency corresponds to using an ESLAT score of 550 to separate those "proficient" in English from those that are not. This normative definition can be supported empirically in the sample of 1,104. Plots of the conditional means of SAT I scores for given ESLAT scores reveal a strong linear trend above about 550 and considerable noise below that point. This transition from scatter to linearity with increasing levels of English proficiency is indicative of a mixture of two or more populations. Similar findings have been noted by Powers (1980) and Wilson (1982) in earlier studies involving the regressions of developed ability measures onto tests of English as a second language. At the very least, there are English proficient and English nonproficient populations. Realistically, there are probably a limited number of degrees of proficiency in these data, and it is likely that the relationship between the predicted scores and the predictors varies across these samples.

We considered being less selective, using a lower cut on ESLAT, and more selective, using a higher cut on ESLAT. More selectivity would have produced steeper regression surfaces, but further restricted the

Table 3

Average Exam Performance of Self-Selected Sample and the Analysis Sample, Which Has ESLAT Scores of 550 or Higher

Test Score	*Analysis Sample*		*Self-Selected Sample*	
Spanish Language	*N*	*Mean (SD)*	*N*	*Mean (SD)*
ESLAT	965	680 (57)	1,104	654 (89)
PAA MATH	965	645 (97)	1,104	634 (102)
PAA VERBAL	965	583 (94)	1,104	573 (97)
English Language				
SAT I MATH	965	469 (93)	1,104	456 (97)
SAT I VERBAL	965	461 (92)	1,104	442 (101)

generalizability of the results, which were already restricted by the nature of the self-selected sample. Less selectivity would have included more generalizable score ranges on PAA and ESLAT, at the expense of flatter regressions than those obtained with a cutoff score of 550.

After screening out scores below 550 on ESLAT, all correlations dropped in magnitude, as would be expected because of the reduction in standard deviations for all scores (see Tables 3 and 4). Direct selection on a variable, in this case ESLAT, leads to indirect selection on all the other variables (Gulliksen, 1950). As a consequence, variances drop on all variables, but mostly on the direct selection variable. We did not screen out low ESLAT scores in an attempt to improve the correlation. Our goal was to arrive at a cleaner analysis sample, one in which English proficiency would have less predictive power than PAA scores.

Even with selection on the basis of ESLAT, ESLAT remains the best predictor of SAT I Verbal scores. This might suggest that we did not screen out enough students. It also could be that language proficiency plays a stronger role in performance on a second language test than might be expected. In any case, ESLAT is still the best predictor of SAT I Verbal scores.

Table 4

Correlations Among Test Scores in Self-Selected Sample (Below Diagonal) and Analysis Sample (Above Diagonal)

	Self-Selected Sample–Analysis Sample Correlations				
Test Score	*ESLAT*	*PAA–M*	*PAA–V*	*SAT I–M*	*SAT I–V*
ESLAT	1.00	.49	.45	.54	.69
PAA MATH	.51	1.00	.58	.82	.56
PAA VERBAL	.45	.61	1.00	.52	.61
SAT I MATH	.57	.82	.56	1.00	.65
SAT I VERBAL	.74	.60	.62	.69	1.00

Linearity

One by-product of screening on ESLAT is that the relationships between predicted and predictor variables became more linear. We examined nonlinear models and found that prediction was hardly improved. When considered against the complexity of depicting these nonlinear results for users, these small improvements in predictability seemed practically insignificant.

Single-Score Linkage: Two-Variable Prediction Tables Versus Single-Score Scalings

The equations for predicting SAT I Verbal and SAT I Math scores from PAA scores and ESLAT scores are approximated by the following formulae. For SAT I Verbal:

Estimated SAT I–V=.366*PAA–V+.848*ESLAT - 329.

This equation yields predictions that correlate .77 with the actual SAT I Verbal scores in this self-selected sample. Note that the weight assigned to ESLAT is more than twice that assigned to PAA Verbal. The two-dimensional table (Table 5) for SAT I Verbal reflects this greater dependence on ESLAT than PAA Verbal. In fact the multiple correlation of .77 is only slightly larger than the correlation of .74 between ESLAT and SAT I Verbal, indicating that PAA Verbal adds little to the prediction of SAT I Verbal beyond what can be explained by ESLAT.

The first column of Table 5 contains PAA Verbal scores ranging from 450 to 800 in steps of 50, plus a row for 580, which is near the PAA Verbal mean of 583. The top row contains ESLAT scores ranging from 550 to 800 in steps of 50, plus a column for 680, which is the ESLAT mean. The table starts at an

Table 5

SAT I Verbal Predicted and Scaled Values in Sample with ESLAT Scores of 550 or Higher

	Prediction							
	ESLAT							
PAA–V	*550*	*600*	*650*	*680*	*700*	*750*	*800*	*Scaling*
450	300	340	390	**410**	430	470	510	330
500	320	360	400	**430**	450	490	530	380
550	340	380	420	**450**	470	510	550	430
580	**350**	**390**	**430**	**460**	**480**	**520**	**560**	**460**
600	360	400	440	**470**	480	530	570	480
650	380	420	460	**490**	500	540	590	530
700	390	440	480	**500**	520	560	610	570
750	410	450	500	**520**	540	580	620	620
800	430	470	510	**540**	560	600	640	670

ESLAT score of 550 for an obvious reason. No examinees in the analysis sample had scores below 550 on ESLAT. Approximately 7 in 8 of the original matched sample (965/1,104) were included in the analysis sample, because they had scores of 550 or higher on ESLAT. Nearly as many of the original 1,104 had PAA Verbal score of 450 or greater. So 450 was chosen as the table cutoff for PAA Verbal; in the actual analysis sample 84 examinees scored below 450 on PAA Verbal.

The body of the table contains the predicted value of SAT I Verbal for each combination of ESLAT and PAA Verbal and the linear scaling of PAA Verbal to SAT I Verbal. The importance of ESLAT relative to PAA Verbal is evident in this table. For example, a 550 on both ESLAT and PAA Verbal yields a predicted SAT I Verbal score of 340. If ESLAT remains fixed at 550, and PAA Verbal is allowed to increase to 750, the predicted SAT I Verbal score becomes 410, a gain of 70 points for a 200-point gain in PAA Verbal. In contrast, if ESLAT is allowed to increase to 750, while PAA Verbal stays fixed at 550, then the predicted SAT I Verbal score increases from 340 to 510, a 170-point gain for a 200-point gain in ESLAT. If the gain is 100 points from 550 to 650 on ESLAT and PAA Verbal, the predicted SAT I Verbal score becomes 460, a 120-point gain, halfway between the 70 and 170 point gains noted above. ESLAT is the primary predictor of SAT I Verbal, and PAA Verbal adds some predictive power beyond that contained in ESLAT alone.

Although the correlation between the PAA Verbal and SAT I Verbal is only .62, linear scaling results are presented at the extreme right of Table 5 for comparison purposes. Comparison of the estimated SAT I Verbal score, keeping ESLAT constant at 680, to the scaling results shows that both coincide at the PAA Verbal mean (580), but at every other point the scaling is much steeper than the average predicted line with differences ranging between 10 to 130 points.

Table 6 depicts the prediction of SAT I Math from PAA Math and ESLAT and the linear scaling of PAA Math to SAT I Math. As with the Verbal table, ESLAT ranges from 550 to 800. PAA Math ranges from 500 to 800; in the matched sample, 132 of 1,104 had scores below 500 on PAA Math, which is quite close to the 139 of 1,104 observed for ESLAT. Here the primary predictor is clearly PAA Math. A 550 on both Spanish language tests produces a predicted SAT I Math score of 360, which increases to 420 if ESLAT increases to 750 holding PAA Math at 550, and which increases to 500 if ESLAT is held at 550 while PAA Math increases to 750. Although higher than the correlation between the Verbal scores, the correlation between the PAA Math and SAT I Math (.82) is still lower than the criteria of .87. The scaling results are presented at the extreme right of Table 6. Comparison of the estimated SAT I Math score, keeping ESLAT constant at 680, to the scaling results shows that both basically coincide at a PAA Math score of 650 (the PAA Math mean was 645); below the 650 score, the scaling gives scores that are between 10 to 40 points lower than the predicted scores, while above the 650 score, the scaling gives scores that are 10 to 40 points higher (much closer to the predicted scores than for Verbal).

Table 6

SAT I Math Predicted and Scaled Values in Sample with ESLAT Scores of 550 or Higher

	Prediction							
	ESLAT							
PAA–M	*550*	*600*	*650*	*680*	*700*	*750*	*800*	*Scaling*
500	330	340	360	**370**	370	390	400	330
550	360	380	390	**400**	410	420	440	380
600	400	410	430	**440**	440	460	470	420
640	**430**	**440**	**460**	**470**	**470**	**490**	**500**	**460**
650	430	450	460	**470**	480	490	510	470
700	470	480	500	**510**	510	530	540	520
750	500	520	530	**540**	550	560	580	570
800	540	550	570	**580**	580	600	610	620

The equation for predicting SAT I Math from PAA Math and ESLAT makes it more obvious that PAA Math is the more important predictor:

Estimated SAT I–M = .697*PAA–M + .290*ESLAT - 178.

PAA Math has a weight that is more than twice as large as that for ESLAT. The multiple correlation of .84 is higher than the .77 obtained for SAT I Verbal. Here ESLAT, while not dominant as it was with SAT I Verbal, still adds something to the prediction of SAT I Math. The multiple correlation of .84 is notably higher than the simple correlation of .69 between PAA Math and SAT I Math.

Composite-Score Linkage: Expectation Tables Versus Concordances

The equation for predicting SAT I Verbal + SAT I Math scores from PAA Verbal + PAA Math scores and ESLAT scores is:

Estimated SAT I–V + M =
.576*PAA–V + M + 1.08*ESLAT - 515.

The weight assigned to ESLAT is almost twice that assigned to the composite of PAA–V+M because the composite score is twice as large as ESLAT. The multiple correlation for this prediction is the same (to two decimals) as that obtained for predicting SAT I Math, .83, indicative of the fact much of the linkability across languages resides in the math portions of the test.

The first column of Table 7 contains PAA–V+M scores ranging from 1000 to 1600 in steps of 100, plus a row for 1230, which is close to the PAA–V+M mean of 1228. The top row contains ESLAT scores ranging from 550 to 800 in steps of 50, plus a column for 680, which is the ESLAT mean.

The body of the table contains the predicted value of SAT I–V+M for each combination of ESLAT and PAA–V+M and the linear scaling of PAA–V+M to SAT I–V+M. When comparing the PAA–V+M and ESLAT, bear in mind that because the PAA–V+M is a composite, it is on a different scale than ESLAT (if the two measures were perfectly related, then for every 100 points of gain in PAA–V+M the expected gain would be 50 points in ESLAT). Although on a different scale, the importance of ESLAT relative to PAA–V+M is also evident in this table. For example, a 550 on ESLAT and a 1000 on PAA–V+M yields a predicted SAT I–V+M score of 660. If ESLAT remains fixed at 550, and PAA–V+M is allowed to increase to 1200, the predicted SAT I–V+M score becomes 770, a gain of about 100 points for a 200-point gain in PAA–V+M. In contrast, if ESLAT is allowed to increase to 600, while PAA–V+M stays fixed at 1000, then the predicted SAT I–V+M score increases from 660 to 710, a 50-point gain for about the same gain in ESLAT. ESLAT and PAA–V+M are basically equally predictive of SAT I–V+M.

The correlation between the composite PAA–V+M and SAT I–V+M (.78) is higher than for the individual Verbal scores (.62) but lower than for the individual Math scores (.82). Although this composite correlation of .78 is lower than the criteria of .86, linear scaling results are presented at the extreme right of Table 7 for comparison purposes. Comparison of the estimated composite SAT I–V+M score keeping ESLAT constant at 680 to the scaling results show that both coincide at the PAA–V+M score of 1230 (which is very close to the mean of 1228). The scaling is much steeper than the predicted line, as expected for measures that are only moderately correlated.

Table 7

SAT I Verbal+Math Predicted and Scaled Values from Linkages in the Sample with ESLAT Scores of 550 or Higher

	Prediction							
PAA–	*ESLAT*							
V+M	*550*	*600*	*650*	*680*	*700*	*750*	*800*	*Scaling*
1000	660	710	770	**800**	820	870	930	700
1100	710	770	820	**860**	880	930	990	800
1200	770	830	880	**910**	940	990	1040	900
1230	**790**	**840**	**900**	**930**	**950**	**1010**	**1060**	**930**
1300	830	880	940	**970**	990	1050	1101	1000
1400	890	940	1000	**1030**	1050	1100	1160	1100
1500	950	1000	1050	**1090**	1110	1160	1220	1200
1600	1000	1060	1110	**1140**	1170	1220	1270	1300

Discussion

This study employed a bilingual group design, using a language proficiency measure (ESLAT) to help define the bilingual group. Prediction and scaling results were compared. Results indicated that for single and composite score correspondence, particularly for the Verbal score, prediction is preferable to scaling. Comparison of prediction and scaling results for composite scores versus single scores indicates that when Verbal is added to Math, the prediction for the resultant composite score is better than that obtained for Verbal alone but worse than that obtained for Math alone.

The single and composite-score prediction equations have limited generalizability. A certain level of English proficiency is required. Scores for examinees with ESLAT scores below 550 had an erratic relationship with the test score variables, which resulted in a noisy regression at these levels. In other words, a certain level of ability as measured by ESLAT is needed before scores become systematically related to the other test scores and, more importantly, before the relationships between scores of similar constructs measured in Spanish and English stabilize.

The prominent role of ESLAT in predicting SAT I Verbal scores even in this group of high ESLAT scores (550 is almost a standard deviation [118] above the mean of 446 in the full PAA population) brings to the fore the problems of trying to link scores across languages. The large differences between prediction and scaling when ESLAT was kept constant at the mean exemplify this. Perhaps prediction may be the best that can be achieved when linking PAA and SAT I scores. Nonetheless, prediction can be used to represent the range of students' performance on the SAT I (Verbal or Math) using PAA and ESLAT scores to predict how students in Puerto Rico might do in U.S. mainland colleges or universities.

One potential drawback of the current design is that the sample was not representative of all examinees taking the PAA. The studied sample consisted primarily of students from private high schools in Puerto Rico with higher levels of English language proficiency who were serious candidates for postsecondary study in U.S. English-speaking colleges or universities. The sample, however, is appropriate for the development of predictive information about such Puerto Rican students' performance in U.S. mainland colleges or universities. The results do not generalize to students with low levels of ESLAT scores.

Generalizations to other groups not represented by the sample (including groups taking the PAA in Latin America beyond Puerto Rico) may not be appropriate because the relationship between SAT I and PAA and ESLAT may differ in these other countries among students with adequate levels of English proficiency. Again, any generalization to studies with lower levels of English proficiency is unwise.

Another potential drawback of the current design was that a comparable Spanish language proficiency measure was not necessarily available for U.S. English-speaking students taking the SAT I and the PAA. Thus, a comparable prediction of PAA scores given an SAT I and a Spanish proficiency score could not be studied. This drawback, however, may not be important because prediction of PAA scores from Spanish proficiency and SAT I scores may be of limited practical interest.

One final drawback of the procedure used for the current study was that the end result of the study will not be a single concordance that will permit direct comparison of subgroups of students taking the PAA with subgroups of students taking the SAT I. This elusive goal has been sought in earlier studies. The current study has focused on a tractable practical goal.

References

Alderman, D.L. (1981). *Language proficiency as a moderator variable in testing academic aptitude* (TOEFL Research Report #10, RR-81-41). Princeton, NJ: Educational Testing Service.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, D.C.: American Council on Education.

Angoff, W.H., & Cook, L.L. (1988). *Equating the scores of the Prueba de Aptitud Académica and the Scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Angoff, W.H., & Modu, C.C. (1973). *Equating the scales of the Prueba de Aptitud Académica and the Scholastic Aptitude Test* (Research Report No. 3). New York: College Entrance Examination Board.

Boldt, R.R. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade County high school volunteers* (Statistical Report 69-31). Princeton, NJ: Educational Testing Service.

The College Board (1995). *Cambio en el examen de admision del College Board: La nueva PAA.* Oficina de Puerto Rico y de Actividades Latinoamericanas.

Cook, L.L. (1995, April). *Lessons learned: Implementing change in the SAT.* Paper presented at the annual meeting of the National Council on Educational Measurement, San Francisco.

Dorans, N.J. (2002). *The recentering of SAT score scales and its effects on score distributions and score interpretations.* College Board Research Report (No. 2002-11), ETS RR-02-04. 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.

Geisinger, K.F. (1994). Cross-cultural normative assessment: Transition and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley & Sons.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9,* 57–68.

Holland, P.W., & Rubin, D. B. (1982). *Test equating.* New York: Academic Press.

Kolen, M.J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83–102.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Olmedo, E.E. (1981). Testing linguistic minorities. *American Psychologist, 36*, 1078–1085.

Pennock-Roman, M. (1995). *Measuring developed academic abilities using Spanish vs. English-language tests: PAEG/GRE relationships for Puerto Ricans who are more proficient in Spanish than in English* (GRE Research Report No. 89-01). Princeton, NJ: Educational Testing Service.

Powers, D.E. (1980). *The relationship between scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* (ETS Research Report RR-80-31). Princeton, NJ: Educational Testing Service.

Prieto, A.J. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43,* 1–14.

Schmitt, A., Dorans, N.J., Magrina, A., & Cook, L. (1998, April). *Predicting scores on the English Language SAT from the Spanish Language PAA and the Spanish Language English as a Second Language Achievement test.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Sireci, S.G. (1996, April). *Technical issues in linking assessments across languages*. Paper presented at the annual meeting of the National Council on Educational Measurement, New York.

Wilson, K.M. (1982). *GMAT and GRE Aptitude Test performance in relation to primary language and scores on TOEFL* (ETS Research Report RR-82-28). Princeton, NJ: Educational Testing Service.

www.collegeboard.com