DOCUMENT RESUME

ED 304 457		TM 012 829

AUTHOR	Angoff, William H.; Cook, Linda L.
TITLE	Equating the Scores of the "Prueba de Aptitud Academica" and the "Scholastic Aptitude Test." College Board Report No. 88-2.
INSTITUTION	College Entrance Examination Board, New York, N.Y.; Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-88-3
PUB DATE	88
NOTE	24p.
AVAILABLE FROM	College Board Publications, P.O. Box 886, New York, NY 10101-0886 ($6.00).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Algorithms; College Bound Students; *College Entrance Examinations; Comparative Analysis; Difficulty Level; English; *Equated Scores; Higher Education; *Latent Trait Theory; Mathematics Tests; Spanish Speaking; Verbal Tests
IDENTIFIERS	Anchor Tests; *Prueba de Aptitud Academica; *Scholastic Aptitude Test

ABSTRACT

With some procedural differences, this study replicated an early study designed to develop algorithms for converting scores on the Scholastic Aptitude Test (SAT) with those on the Prueba de Aptitud Academica (PAA) scale and vice versa. The study involved selection of test items equally appropriate and useful for English- and Spanish-speaking students for use as an anchor test and the equating analysis itself. Once the items were selected, they were administered as pretests, one for each language, to determine whether the two response functions for each item were sufficiently similar for the items to be considered equivalent. On the basis of these analyses, 39 verbal and 25 mathematical items were selected for use as anchor items for equating. The anchor tests were administered at regularly scheduled administrations of the SAT and PAA. An item response theory model was used to equate the two tests. The equating itself showed curvilinear relations in both verbal and mathematical tests, indicating that, in this instance, both sections of the PAA are easier than the corresponding SAT sections. Differences between these findings and those of the previous study by W. H. Angoff and C. C. Modu (1973) are assessed. Six graphs and six data tables are provided. (TJH)

College Board Report

No. 88-2

Equating the Scores of the *Prueba de Aptitud Académica*™ and the *Scholastic Aptitude Test*®

William H. Angoff
Linda L. Cook

Equating the Scores of the *Prueba de Aptitud Académica*™ and the *Scholastic Aptitude Test*®

William H. Angoff
Linda L. Cook

College Board Report No. 88-2
ETS RR No. 88-3

College Entrance Examination Board, New York, 1988

William H. Angoff is distinguished research scientist at Educational Testing Service, Princeton, New Jersey.

Linda L. Cook is principal measurement specialist at Educational Testing Service, Princeton, New Jersey.

Acknowledgments

The authors express their deep appreciation for the fine work of many professionals who have contributed in important ways to the success of the study: Eugene Mohr, Rosemarie Cruz, Aileene Alvarez, Eugene Francis, and Victor Garcia, professors at the University of Puerto Rico, and Winifred Melendez, professor at the Inter American University of Puerto Rico. These able professionals carried out the important and sensitive work of translating test items that were available in their English form from English to Spanish, other items from Spanish to English, and all the items back again to their original languages. The authors also wish to express their gratitude to Edward Curley, Nancy Anderson, and Protase Woodford of the Test Development group at Educational Testing Service (ETS), who coordinated the assembly of the special equating items and test sections and worked with the translations, to Nancy Wright and Edwin Blew of the College Board Statistical Analysis staff at ETS for their coordination and production of the statistical analyses for this study, and to Marilyn Wingersky for her technical assistance in this study. Finally, the authors wish to acknowledge the important contributions of Carlos J. Lopez-Nazario, deputy director, and the staff of the College Board Puerto Rico Office for their dedicated involvement and able support throughout the course of the study.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $6.

CONTENTS

Abstract 1

Introduction 1

Method 2
Phase 1: Selection of Items for Equating 3
Phase 2: Equating 9

Summary and Discussion 15

References 17

Figures

1. Plots of item response functions for verbal and mathematical items given to PAA and SAT groups, illustrating poor and good agreement between groups 4

2. Plot of b's for pretested verbal items 6

3. Plot of b's for pretested mathematical items 7

4. Item response theory conversions for verbal tests 13

5. Item response theory conversions for mathematical tests 13

Tables

1. Summary Statistics for Pretested Items, by Language of Origin, before and after Selection of Equating Items 8

2. Summary Statistics for Pretested Items, by Item Type 8

3. Distribution of Pretested Items, by Item Type and Language of Origin 9

4. Frequency Distributions and Summary Statistics for Verbal Operational and Equating Sections of the SAT and PAA 11

5. Frequency Distributions and Summary Statistics for Mathematical Operational and Equating Sections of the SAT and PAA 12

6. Final Conversions between PAA and SAT 14

ABSTRACT

The present study is a replication, in certain important respects, of an earlier study conducted by Angoff and Modu (1973) to develop algorithms for converting scores expressed on the College Board Scholastic Aptitude Test (SAT) scale to scores expressed on the College Board Prueba de Aptitud Académica (PAA) scale, and vice versa. Because the purpose and the design of the studies, though not all of the psychometric procedures, were identical in the two studies, the language of this report often duplicates that of the earlier study. The differences in procedure, however, are worth noting, and it is hoped that this study will contribute in substance and method to the solution of this important problem.

The study described in this report was undertaken in an effort to establish score equivalences between two College Board tests—the Scholastic Aptitude Test (SAT) and its Spanish-language equivalent, the Prueba de Aptitud Académica (PAA). The method involved two phases: (1) the selection of test items equally appropriate and useful for English- and Spanish-speaking students for use as an anchor test in equating the two tests; and (2) the equating analysis itself. The first phase called for choosing a set of items in each of the two languages, translating each item into the other language, "back-translating" independently into the original language, and comparing the twice-translated versions with their originals. This process led to the adjustment of the translations in several instances and, in other instances, to the elimination of some items considered too difficult to be translated adequately. At this point both sets of "equivalent" items, each in its original language mode, were administered as pretests, chiefly to determine whether the two response functions for each item were sufficiently similar for the items to be considered equivalent.

On the basis of these analyses two sets of items—one verbal and the other mathematical—were selected for use as anchor items for equating. These were administered again (in the appropriate language) at regularly scheduled administrations of the SAT and the PAA. An item response theory (IRT) model was used to equate the PAA to the SAT, with the anchor items serving as the link in the equating process.

The equating itself showed definite curvilinear relationships in both verbal and mathematical tests, indicating in this instance that both sections of the PAA are easier than the corresponding SAT sections. The results also showed good agreement between the current conversions and the 1973 Angoff-Modu conversions for the mathematical tests, but not so close agreement for the verbal tests. The reasons for the difference are (speculatively) attributed to improved methodology in the present study, especially for the more difficult verbal equating, and to the possibility of scale drift in one or the other test (or both tests) over the intervening 12 to 15 years since the last study.

INTRODUCTION

Although the study of cultural differences has been of central interest to educators and social psychologists for many years, attempts to develop a deeper understanding of such differences have been frustrated by the absence of a common metric by which many comparisons could be made. The reasons for this are clear. If two cultural groups differ from each other in certain ways that cast doubt on the validity of direct comparisons between them in other respects—if, for example, they differ in language, customs, and values—then those very differences also defy the construction of an unbiased metric by which we could hope to make such comparisons.

We find, however, that there are times when comparisons are nevertheless made, even though the basic differences in language, customs, and values, for example, which sometimes invalidate these comparisons, are known to exist. The present study has been designed in an attempt to develop a method to help make such comparisons in the face of these difficulties by providing a common metric. Specifically, it purports to provide a conversion of the verbal and mathematical scores on the College Board Spanish-language Prueba de Aptitud Académica (PAA) to the verbal and mathematical scores, respectively, on the College Board English-language Scholastic Aptitude Test (SAT). Both tests, it is noted, are administered to secondary school students for admission to college. The PAA is typically administered to Puerto Rican students who are planning to attend colleges and universities in Puerto Rico; the SAT is typically administered to mainland students who are planning to attend colleges and universities in the United States. It was expected that if conversion tables between the score scales for these two tests were made available, direct comparisons could be made between subgroups of the two language-cultures who had taken only that test appropriate for them. For the immediate purpose, however, it was expected that these conversion tables would help in the evaluation of the probable success of Puerto Rican students who were interested in eventually attending college on the mainland and were submitting PAA scores for admission. As already indicated in the Abstract, the study was conducted in an effort to repeat the earlier study by Angoff and Modu, but with some modifications and improvements in method, and to confirm that the earlier results are still valid.

Interest in developing conversions such as these has been expressed in various other contexts, usually in the assessment of the outcomes of education for different cultural groups living in close proximity: for example, for English- and French-speaking students in Canada, for English- and Afrikaans-speaking students in South Africa, for speakers of one or another of the many languages in India or in Africa. No satisfactory methods to satisfy this interest have been available until recently, however; and the problems attendant on making comparisons among culturally different groups are far more obvious and numerous than are the solutions. For example, to provide a measuring instrument to make these comparisons, it is clearly insufficient simply to translate the test constructed for one language group into the language of the other, even with adjustments in the items to conform to the more obvious cultural requirements of the second group. It can hardly be expected, without careful and detailed checks, that the translated items will have the same meaning and relative difficulty for the second group as they had for the original group before translation.

A method considerably superior to that of simple translation has been described by Boldt (1969). It requires the selection of a group of individuals judged to be equally bilingual and bicultural and the administration of two tests to each individual, one test in each of the two languages. Scores on the two tests are then equated as though they were parallel forms of the same test, and a conversion table is developed relating scores on each test to scores on the other.

One of the principal difficulties with the foregoing procedure, however, is that the judgment "equally bilingual and bicultural" is extremely difficult, perhaps even impossible, to make. More than likely, the individual members of the group, and even the group as a whole, will on average be more proficient in one of the two languages than in the other. This will be especially true, of course, if the group is small.

This study represents an attempt to overcome such difficulties. In brief, it calls for administering the PAA to Puerto Rican students and the SAT to mainland United States students, using a set of "common," or anchor, items to calibrate and adjust for any differences between the groups in the process of equating the two tests. It is noted that these items are common only in terms of the operations used to develop and select them. By the very nature of things they had to be administered in Spanish to the Puerto Rican students and in English to the mainland students. Therefore, to the extent that there is any validity in the notion that a set of test items can represent the same psychological task to individuals of two different languages and cultures, to the extent that the sense of the operations is acceptable, and to the extent that the operations themselves were adequate, the study will have achieved its purpose. There is also the concern that the Puerto Rican and the mainland groups appear to differ so greatly in average ability that with the limited equating techniques available, it is not likely that any set of common items, however appropriate, can make adequate adjustments for the differences, even if the two tests were designed for students of the same language and culture.

There is, finally, the concern about the generalizability of a conversion between tests that are appropriate for different cultural groups. In the usual equating problem, a conversion function is sought that will simply translate scores on one form of the test to the score scale of a parallel form of the test—an operation analogous to that of translating Fahrenheit units of temperature to Celsius units. When the two tests in question are measuring different types of abilities, however, or when one or both of the tests may be unequally appropriate for different subgroups of the population, the conversion cannot be unitary, as would be true of the temperature-scale conversion, but would be different for different subgroups (Angoff 1966). In the present equating attempt, it is entirely possible that the use of different types of subgroups for the equating experiment—Mexicans and Australians, for example, instead of Puerto Ricans and United States mainlanders—would yield conversion functions quite different from those developed in the present study. For this reason the conversions developed here should be considered to have limited applicability and should not be used without verification with groups of individuals different from those studied here.

METHOD

In broad outline the method followed in this study for deriving conversions of scores from the verbal and mathematical scales of the PAA to the verbal and mathematical scales of the SAT was the same as that followed in the Angoff-Modu (1973) study referred to above. As carried out previously, this study was conducted in two phases: The first phase entailed the selection of appropriate anchor items for equating. This phase called for the preparation of sets of items both in Spanish and in English, the translation of each set into the other language by Puerto Rican educators proficient in both languages, and the administration of both sets in the appropriate language mode to Spanish- and English-speaking students. On the basis of an item analysis of the data resulting from this administration, groups of verbal and mathematical items were chosen to fulfill the principal requirement that they be equally appropriate, insofar as this could be determined, for both student groups. Beyond this requirement, the usual criteria for the choice of equating items as to difficulty, discrimination, and content coverage were adhered to, to the extent possible. In addition, care was taken, also where possible, to produce sets of anchor items reasonably balanced as to Spanish or English origin. Once the anchor items were chosen, the second phase of the study was undertaken, which called for a second test administration and an analysis for equating based on the data resulting from that administration.

Phase 1: Selection of Items for Equating

In accordance with the foregoing plan, 105 Spanish verbal items, 110 English verbal items, 62 Spanish mathematical items, and 62 English mathematical items were drawn from the file and submitted to bilingual experts in Puerto Rico for translation. Two experts were assigned to translate the Spanish verbal items into English, and two other experts were assigned to translate the English verbal items into Spanish. After this initial translation was completed, the two sets of experts independently back-translated each other's work into the original language. All translations were then sent to Educational Testing Service, where Spanish-language experts compared the back-translated items with the original items. Adjustments were made in the initial translations by the ETS staff and by the staff of the College Board Puerto Rico Office when the comparisons revealed inadequacies. In some instances it was judged that revisions could not be made adequately, and as a result a number of items were dropped from further use. The same process was carried out for the mathematical items. Because of the smaller number of mathematical items, however, only two translators were used—one for translating items from Spanish to English and the other, items from English to Spanish. Eventually two complete sets of items were compiled, 160 verbal and 100 mathematical; each set appeared in both languages and, to the extent that this could be observed at an editorial level, was equally meaningful in both languages.

The 160 verbal items were of four types, paralleling the item types normally appearing in the operational forms of the PAA and the SAT: antonyms, analogies, sentence completion, and reading comprehension. The 100 mathematical items fell into four content categories: arithmetic, algebra, geometry, and miscellaneous. Detailed quantitative information on the pretested items is given later in this report.

The 160 verbal items and the 100 mathematical items were subdivided into four 40-item verbal sets and four 25-item mathematical sets and administered to spiraled samples of regular College Board examinees. The test items in English were taken by candidates for the English-language SAT at the January 1985 administration; the same test items, in Spanish, were taken by candidates for the Spanish-language PAA at the October 1984 administration. All of the foregoing sets of items were administered in 30-minute periods. All SAT and PAA examinee samples consisted of about 2,000 cases.

Item response theory (IRT) methods were used to compare performance on the verbal and mathematical items by the SAT and PAA groups. Items that functioned most similarly for the two groups were selected to constitute the 40-item verbal and 25-item mathematical equating tests.

Of all the methods currently available for selecting items that function similarly for two groups of examinees, the three-parameter IRT method (Lord 1977; Petersen 1977; Shepard, Camilli, and Williams 1984) used in this study is most preferable. This is so because it minimizes effects related to differences in group performance that seriously confound the results of simpler procedures such as the delta-plot method (Angoff and Ford 1973) used in the previous study.

Item response theory methods may be used to compare two groups of examinees with respect to their responses to a particular item for the full ability (θ) continuum. Item characteristic curves (ICCs), such as those shown in Figure 1, describe the relationship between the probability of a correct response to an item and the degree of ability measured by the item. The curves in Figure 1 are described by the values of three item parameters: a, b, and c. These parameters have specific interpretations: b is the point on the θ metric at the inflection point of the ICC (where the slope of the curve reaches its maximum and begins to decrease) and is taken as a measure of item difficulty; a is proportional to the slope of the ICC at the point of inflection and represents the degree to which the item provides useful discriminations among individuals; c is the value of the lower asymptote of the ICC (where the slope is essentially zero) and represents the probability that an examinee with very low ability will obtain a correct answer to the item.

Studies of differential item difficulty were undertaken by estimating the ICCs of the pretested items separately for the PAA and SAT groups. Theoretically, if the item has the same meaning for the two groups, the probability of a correct response should be the same for examinees of equal ability (i.e., for any value of θ along the continuum). Panel A of Figure 1 contains a comparison of item response functions obtained for a verbal item given to the PAA and SAT groups. It can be seen, from examination of the ICCs in Panel A, that for all levels of ability (θ) the PAA group has a higher probability of obtaining a correct answer to the item; i.e., the item is seen to function in favor of the PAA group. Panel B of Figure 1 contains a comparison of ICCs obtained for a mathematical item given to the PAA and SAT groups. In contrast to the curves shown in Panel A, the ICCs for the mathematical item given to the two groups of examinees are almost identical; i.e., individuals at all levels of ability in both groups have the same probability of obtaining a correct answer to the item. The item favors neither of the two groups.

For this study, item parameters and examinee abilities were estimated by use of the computer program

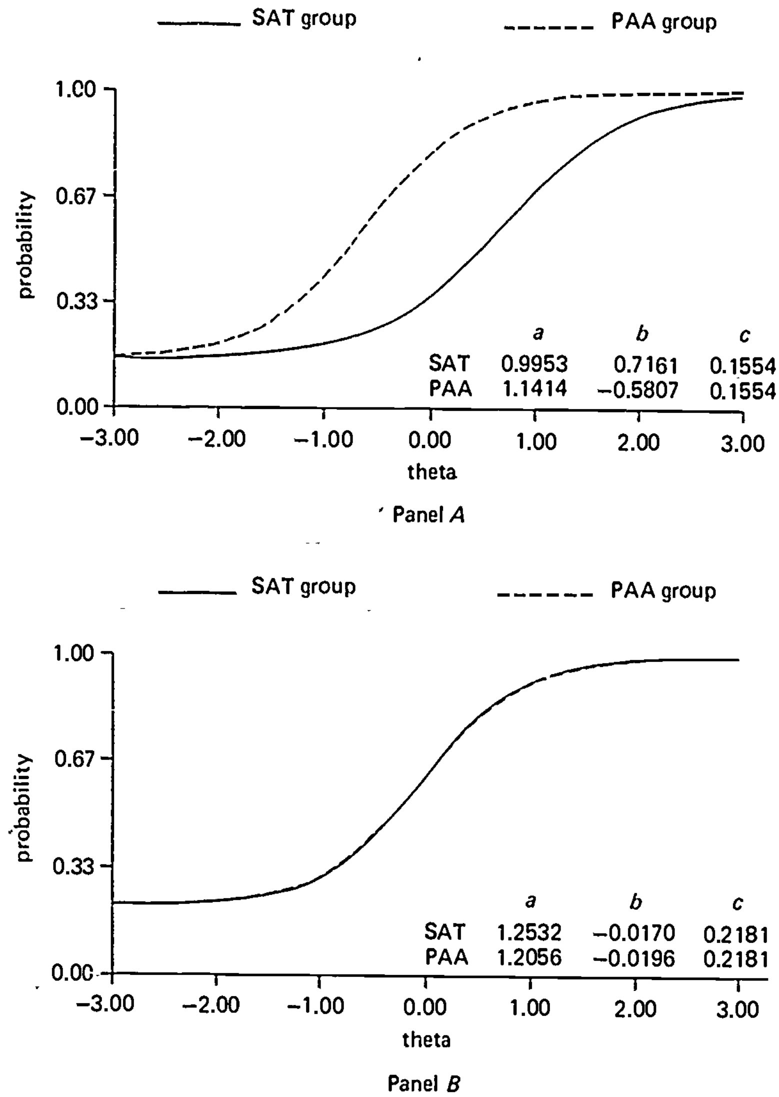

Figure 1. Plots of item response functions for verbal (Panel A) and mathematical (Panel B) items given to PAA and SAT groups, illustrating poor and good agreement between groups.

LOGIST (Wingersky 1983; Wingersky, Barton, and Lord 1982). LOGIST produces estimates of *a*, *b*, and *c* for each item and θ for each examinee. Inasmuch as item parameter estimates for the SAT and PAA groups were obtained in separate calibrations, it was necessary to introduce an item parameter scaling step at this point. The item characteristic curve transformation method developed by Stocking and Lord (1983) was used for this purpose.

The procedure to screen the pretested items for the analysis of differential item difficulty for the PAA and SAT groups has been described by Lord (1980, chap. 14). The method entails the following steps as they were carried out in this study:

1. Data for the combined PAA and SAT groups were used to obtain estimates of the *c* parameters for all the items.
2. Holding *c*'s fixed at these values, *a* and *b* item parameter estimates were obtained separately for the PAA and SAT groups.
3. Following the scaling of the parameter estimates, item characteristic curves for the two groups were compared, and those items that functioned differently in the two groups were identified.
4. Items with significantly different ICCs were removed from the pool of pretested items.
5. Ability estimates were obtained for the com-

bined PAA and SAT groups, using the reduced set of items.

6. Holding ability estimates fixed at values obtained in step 5, a and b parameter estimates were obtained for all pretested items (including those removed in step 4).
7. ICCs and estimates of item parameters were compared for the two groups, and the proposed set of 40 verbal and 25 mathematical equating items was chosen.

Step 1, it is noted, calls for combining the data from the two groups in the calculation of c-parameter estimates and assuming that these estimates are the same in both groups. The reason for this practice is that c-parameter estimates are otherwise often poorly made, are sometimes even indeterminate, and cause difficulties in comparing parameter estimates across groups. The practice does not interfere with testing for significant differences among the a and b parameter estimates inasmuch as the null hypothesis of the IRT χ^2 test used here (Lord 1980, chap. 14) states that the values of a, b, and c are the same for the two groups of interest.

Steps 4 to 6 represent the IRT analogue to criterion purification procedures used with conventional item-bias techniques. Lord (1980, chap. 14) has cautioned that the set of items of interest may not be measuring a unidimensional trait; thus, it is possible that ability estimates (θ) as well as the ICCs obtained for the PAA group may not be strictly comparable to those obtained for the SAT group. One possible solution is to "purify" the test by removing the differentially functioning items and then to use the remaining set of unidimensional items to reestimate the θ's. Finally, the "purified" set of ability estimates is used to obtain the set of item parameter estimates and ICCs (for the total pool of items) being compared.

Many indices are available for quantifying the differences between item characteristic curves or item parameter estimates for two groups of examinees. The two indices chosen for use in this study were the previously mentioned IRT χ^2 (Lord 1980, chap. 14) and the mean of the absolute difference between the ICCs. (See Cook, Eignor, and Petersen 1985 for a description of this statistic.) For each test, verbal and mathematical items were ranked according to their χ^2 values. From the set of items with the smallest χ^2 values, those with the smallest values of the mean absolute difference were chosen. The verbal and mathematical equating tests were constructed by use of this reduced pool of items.

Summary statistics for all pretested items and for the items chosen to constitute the verbal and mathematical equating tests are presented in Figures 2 and 3 and in Tables 1 to 3. Several points should be noted. First, 2 verbal items and 1 mathematical item were eliminated from scoring before the LOGIST calibration. Second, it was not possible to obtain item parameter estimates for 13 verbal items and 6 mathematical items. Finally, 3 verbal and 2 mathematical items were found to be so easy for both groups that stable r-biserial correlation coefficients could not be assured. These items were removed from the study. As a result, the pretested item pools were reduced to 142 verbal and 91 mathematical items.

Figure 2 is a bivariate picture in which the 142 b's for the Spanish verbal items are plotted against the corresponding b's for the same verbal items as they appeared in English. Figure 3 gives a similar plot for the mathematical items. As may be seen from these figures, the verbal plot is much more dispersed than the mathematical plot is, representing a much lower correlation for verbal items ($r = .66$) than for mathematical items ($r = .90$). In general, the correlation between the b's may be regarded as a measure of item-by-group interaction—i.e., the degree to which the items represent, or fail to represent, the same rank order of difficulty in the two languages. In those instances where the two groups are drawn at random from the same general population, it is not unusual to see correlations between item difficulty indices in the neighborhood of .98 and even higher. That the correlation for the verbal items in this study is as low as it is indicates that the verbal items do not have quite the same psychological meaning for the members of the two language groups. Mathematics, on the other hand, with its much higher correlation, appears to be a more nearly universal language. In a sense, this is one of the more significant findings in this study because it reflects the very nature of the difficulties that are likely to be encountered in cross-cultural studies, especially when verbal tests are used. With respect to this study in particular, some doubt is cast on the quality of any equating that could be carried out with tests in these two languages and with groups as different as these. Since the equating items are used to calibrate for differences in the abilities of the PAA and SAT groups, a basic requirement for equating is that the items have the same rank order of difficulty in the two groups; for the verbal items, it is clear that this requirement is not met. Considerable improvement, in the sense of reducing the item-by-group interaction, was achieved in the verbal items (as will be shown below) by discarding the most aberrant items among them and retaining those that showed the smallest differences between the two language groups in their item response curves. Nevertheless, with item-by-group interaction effects even as large as those observed here for the items that were retained, the concern remains that the verbal equating might be much less trustworthy than would be

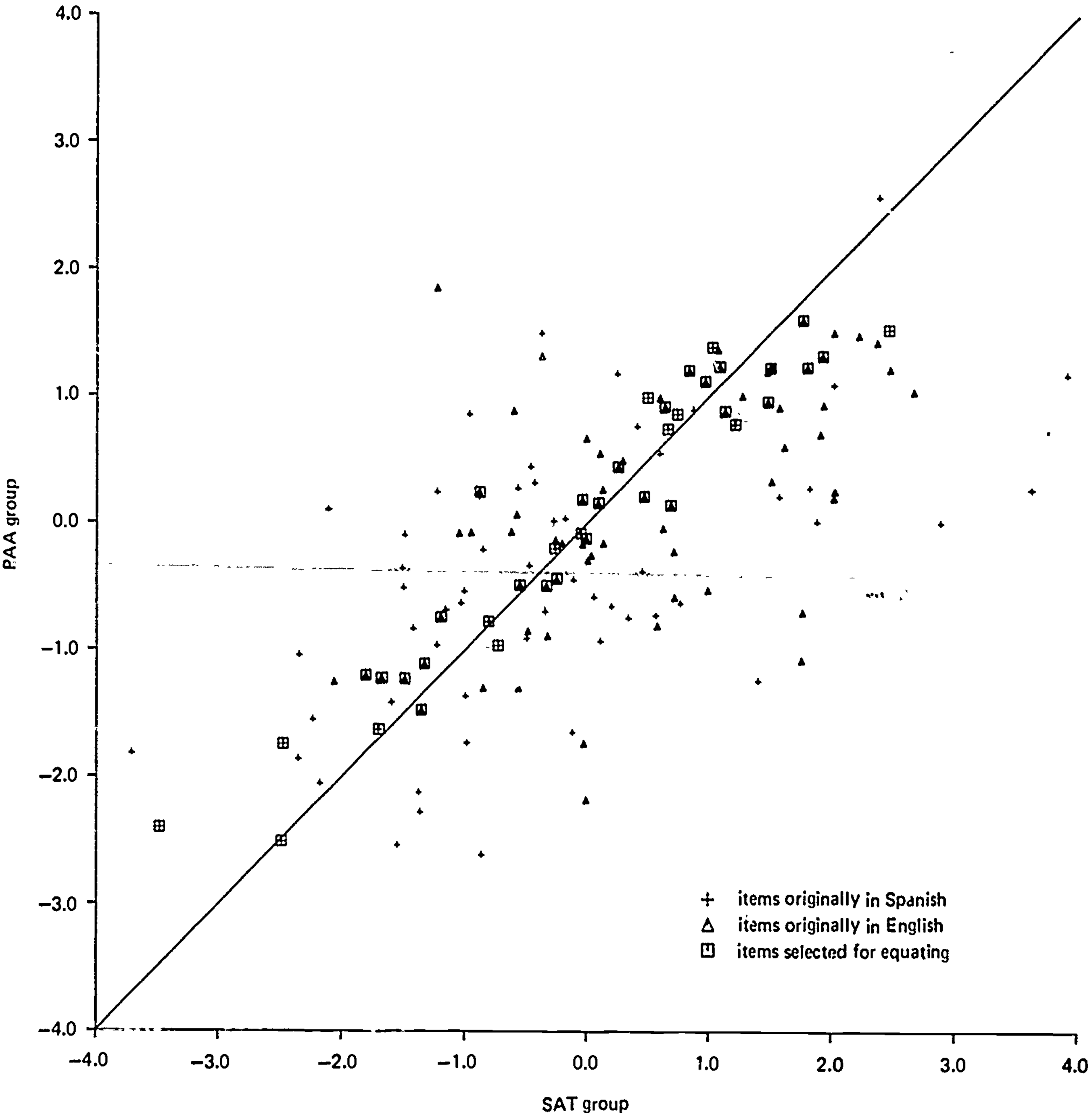

Figure 2. Plot of b's for pretested verbal items (number of items = 142).

expected of an equating of two parallel tests intended for members of the same language-culture.

It bears repetition, however, that these interactions were not entirely unexpected; the observation has often been made that verbal material, however well it may be translated into another language, loses many of its subtleties in the translation process. Even for mathematical items some shift in the order of item difficulty is to be expected, possibly because of differences between Puerto Rico and the United States with respect to the organization and emphasis of the mathematics curriculum in the early grades. But as has already been pointed out, the item-by-group interaction is much less for the mathematical items than for the verbal items.

In Table 1 there is a summary of indices of difficulty (p-values) and discrimination (r-biserials) for the pretested items, as observed in the PAA group and in the SAT group. They are presented separately for the verbal and mathematical items and, within those categories, separately by each item's language of origin. There is also a summary of those indices for the 39 verbal and 25 mathematical items planned for the equating. (It should

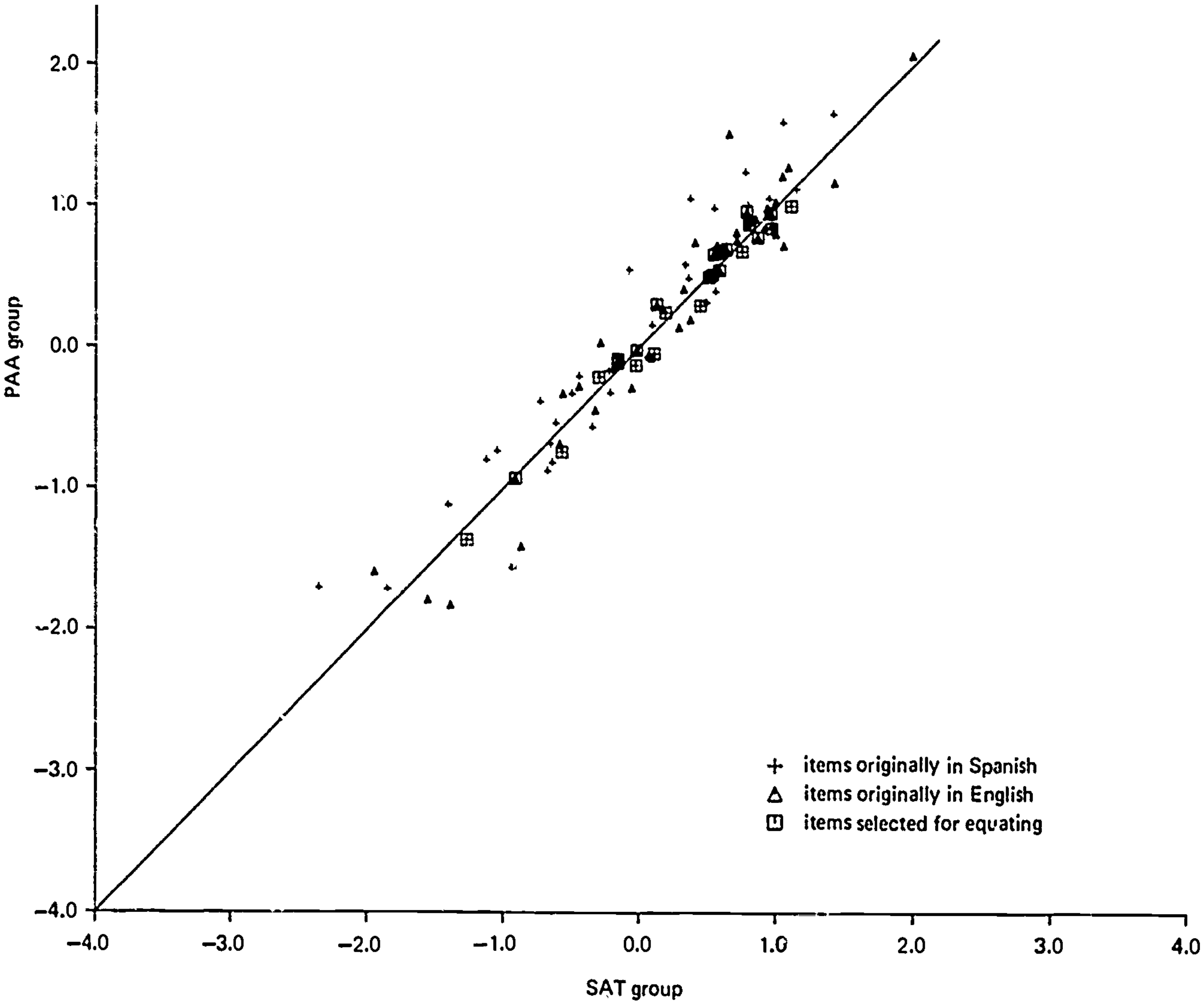

Figure 3. Plot of *b*'s for pretested mathematical items (number of items = 91).

be mentioned here that the original plan was to select 40 verbal items and 25 mathematical items. After the selections were completed, however, it was felt necessary to revise one of the verbal items substantially. Instead it was dropped, reducing the number of verbal equating items from 40 to 39.) The tenth and eleventh columns in Table 1 give the means and standard deviations of the index of discrepancy (the mean absolute difference) between the two item response curves, one of the indices used as the basis for selecting the items. Finally, Table 1 gives the correlations between the *b*-parameter estimates for the two language groups, again by category of item.

As can be seen in Table 1, the items are, on average, considerably more difficult for the PAA candidates than for the SAT candidates. The difference between the mean *p*-values on the verbal items is more than one-half of a standard deviation; the difference on the mathematical items is considerably more than a full standard deviation. For both the PAA and the SAT candidate groups and for both the verbal and the mathematical items, the items originating in Spanish appeared to be relatively easier than those originating in English.

The second set of Table 1 columns summarizes the item-test correlations (*r*-biserials) for the items in their Spanish and English forms. In general, both verbal and mathematical items appear to be less discriminating for the PAA candidates than for the SAT candidates, particularly so for the mathematical items. This difference in discrimination is also present in the group of selected items. It is observed that the mathematical items, at least for the SAT group, have higher mean *r*-biserial correlations on average than do the verbal items, an observation that is frequently made in other reviews of these two item types.

As can be seen in the column summarizing the

Table 1. Summary Statistics for Pretested Items, by Language of Origin, before and after Selection of Equating Items

		Difficulty Values (p)				*Item-Test Correlations (r_{bis})*				*Discrepancy Indices*		
		Mean		*SD*		*Mean*		*SD*				
All Pretest Items	*No. of Items**	*PAA*	*SAT*	*PAA*	*SAT*	*PAA*	*SAT*	*PAA*	*SAT*	*Mean*	*SD*	*Correlations between b's*
Verbal												
Originally English	74	.43	.53	.20	.24	.37	.47	.14	.10	.13	.11	.61
Originally Spanish	68	.50	.63	.21	.21	.41	.42	.13	.14	.13	.11	.66
All verbal items	142	.46	.58	.21	.23	.39	.45	.14	.12	.13	.11	.66
Mathematical												
Originally English	44	.28	.54	.17	.18	.36	.57	.14	.09	.04	.03	.96
Originally Spanish	47	.33	.62	.20	.20	.43	.60	.16	.11	.05	.04	.89
All mathematical items	91	.31	.58	.19	.19	.40	.59	.15	.10	.04	.03	.90
Items selected for equating												
Verbal	39	.43	.60	.22	.23	.37	.47	.15	.09	.06	.03	.96
Mathematical	25	.28	.56	.15	.17	.40	.60	.11	.07	.03	.01	.99

*Three of the 145 verbal items and two of the 93 mathematical items were so easy for both groups that stable r-biserial correlation coefficients for these items could not be assured. Consequently these indices were not calculated for the items in question.

Table 2. Summary Statistics for Pretested Items, by Item Type

		Difficulty Values (p)				*Item-Test Correlations (r_{bis})*				*Discrepancy Indices*		
		Mean		*SD*		*Mean*		*SD*				
All Pretest Items	*No. of Items**	*PAA*	*SAT*	*PAA*	*SAT*	*PAA*	*SAT*	*PAA*	*SAT*	*Mean*	*SD*	*Correlations between b's*
Verbal												
Antonyms	43	.44	.47	.22	.23	.37	.43	.13	.13	.18	.13	.59
Analogies	34	.41	.59	.19	.24	.42	.45	.15	.12	.13	.10	.62
Sentence completion	29	.49	.63	.25	.23	.36	.45	.14	.12	.15	.11	.73
Reading comprehension	36	.51	.66	.16	.17	.41	.45	.12	.13	.08	.06	.75
Mathematical												
Arithmetic	21	.29	.58	.16	.20	.42	.60	.19	.10	.05	.03	.93
Algebra	37	.34	.62	.20	.19	.40	.58	.13	.11	.04	.04	.96
Geometry	26	.28	.54	.19	.19	.38	.60	.14	.09	.05	.03	.92
Miscellaneous	7	.29	.54	.16	.23	.35	.53	.18	.07	.03	.01	.99

*Three of the 145 verbal items and two of the 93 mathematical items were so easy for both groups that stable r-biserial correlation coefficients for these items could not be assured. Consequently these indices were not calculated for the items in question.

discrepancies between the item response curves for the two groups, the discrepancies between the two curves for the verbal items are far greater than for the mathematical items. Also, it is observed that the items selected for equating show smaller mean discrepancies than is observed in the entire groups of pretested items. This is to be expected, of course, since the items were selected largely on the basis of the agreement between the two item response curves.

The last column, giving the correlations between the b's, expresses in correlational terms what has already been observed in Figures 2 and 3—namely, the item-by-group interaction with respect to item difficulty. Here we see again that the correlation between the b parameters is much lower for the verbal items than for the mathematical items. And again, we see that the correlations between the b-values for the selected items—especially the verbal items—are higher than for the unselected items.

Table 2 is a summary of the same data as shown in Table 1 but classified by item type rather than by language of origin. The great difficulty of the items for the PAA group is readily observable in this table. It is also clear that the items in all four verbal and mathematical categories are more discriminating for the United States students than for the Puerto Rican students.

It is interesting that the four verbal types arrange themselves into two distinct classes insofar as the correlations between their b-values are concerned: higher correlations (smaller item-by-group interactions) are characteristic of the sentence completion and reading comprehension items, and lower correlations (larger item-by-group interactions) are characteristic of the antonyms and analogies. This result is intuitively reasonable since items with more context probably tend to retain their meaning, even in the face of translation into another language.

Although the item groups are too small to permit easy generalization, it appears that there is considerable and, very likely, significant variation from one verbal item type to another with respect to the similarity of the item response curves for the two candidate groups. (No such interaction is observed in the mathematical items.) The analogy items especially, and to some considerable extent the sentence completion and reading comprehension items, were more difficult relative to antonyms for the Puerto Rican students than for the mainland United States students. This appears to be a subtle effect, very likely characteristic of the item type itself. It is certainly not a function of the origin of these items and their increased relative difficulty upon translation into the other language. As shown in Table 3, very nearly the same proportion of items for each of the categories was drawn from each language.

Table 3. Distribution of Pretested Items, by Item Type and Language of Origin

	Originally English	*Originally Spanish*	*Total*
Verbal			
Antonyms	21	22	43
Analogies	19	15	34
Sentence completion	16	13	29
Reading comprehension	18	18	36
Total	74	68	142
Mathematical			
Arithmetic	11	10	21
Algebra	15	22	37
Geometry	13	13	26
Miscellaneous	5	2	7
Total	44	47	91

Phase 2: Equating

Once the 39 verbal and 25 mathematical items that were to be used as "common"—more properly, "quasi-common"—items were chosen, preparations were made to administer them to groups of candidates taking the PAA or the SAT for admission to college. Accordingly, two samples of candidates were chosen from the December 1985 administration of the SAT—one to take the verbal items in English (N = 6,017) in a 30-minute period, the other to take the mathematical items in English (N = 6,172), also in a 30-minute period, in addition to the regular operational form of the SAT given at that time. Similarly, two samples of candidates were chosen from the October 1986 administration of the PAA—one to take the verbal items in Spanish (N = 2,886) in a 30-minute period, the other to take the mathematical items in Spanish (N = 2,821), also in a 30-minute period, in addition to the regular operational form of the PAA given at that time. Both the SAT and the PAA samples were drawn systematically from their parent candidate groups. The scaled score means for the SAT samples were 405 verbal and 455 mathematical, compared with their parent group means of 404 verbal and 453 mathematical. The scaled score means for the PAA samples were 466 verbal and 476 mathematical, compared with their parent group means of 465 verbal and 485 mathematical. In all instances the sample means approximated their parent means fairly closely.

Before the PAA verbal and mathematical scores were equated to the SAT verbal and mathematical scores, care was taken to evaluate the common items to determine if they were functioning in the same manner for the PAA and SAT samples. The evaluation was carried out by examining plots of item difficulty indices (delta values[1]). Common items in this study were defined as "equally appropriate" to the Spanish- and English-speaking groups on the basis of the similarity of their rank-order position among other items for the two groups. Five verbal and two mathematical items that were considered "outliers" from this plot were deleted from the common-item sections.

Tables 4 and 5 contain information that may be used to evaluate the extent to which the common items are, in fact, appropriate for both groups of examinees and the extent to which the operational tests are appropriate for their groups. Table 4 contains frequency distributions and summary statistics for the verbal operational and equating sections of the SAT and the PAA. It can be seen, from the verbal equating data in Table 4, that the mainland sample is the higher scoring of the two groups by more than a full standard deviation. The difficulty of the 66-item PAA appears to be just about right, on average, for the Puerto Rican sample; the average percentage-pass on that test (corrected for guessing) was .49. The 85-item SAT is clearly difficult for the mainland sample; the average percentage-pass on that test (also corrected for guessing) was .40.

The patterns of standard deviations and correlations observed in Table 4 between the equating test in English and the SAT and between the equating test in Spanish and the PAA suggest that each of these verbal equating tests is reasonably parallel in function to the operational test with which it is paired.

The data presented in Table 5 describe frequency distributions and summary statistics for the mathematical operational and equating sections of the SAT and the PAA. The mathematical equating data in Table 5 reveal even more sharply than do the verbal equating data in Table 4 that the mainland sample is the higher-scoring of the two. The mean difference in the mathematical common items is about 1.4 standard deviations. Also, note that contrary to the verbal test, the operational PAA-mathematical test was as difficult for the PAA sample (percentage-pass, corrected for guessing, was .39) as was the SAT-mathematical test for the SAT sample (percentage-pass, corrected for guessing, was .40).

Unlike the results shown for the verbal tests in Table 4, the patterns of standard deviations and correlations in Table 5 between the SAT and the equating test in English and between the PAA and the equating test in Spanish suggest that the equating test may be considered parallel in function to the SAT but not quite so parallel to the PAA.

1. The delta index is a transformation of the proportion of the group who answer an item correctly ($p+$) to a normal deviate (z), and from z to a scale with a mean of 13 and a standard deviation of 4 by means of the equation $\Delta = -4z + 13$.

Two kinds of equating were undertaken, linear and curvilinear. Of the several linear methods, two were chosen for use; one is attributed to Tucker (in Angoff 1984, p. 110) and the other to Levine (1955). Two curvilinear methods were used: equipercentile equating (see Angoff 1984, p. 97) and item response theory equating (Lord 1980, chap. 13). Although the results of all the methods were evaluated, only the item response theory results were used and consequently are the only method and results described in this report.

Item response theory (IRT) assumes there is a mathematical function that relates the probability of a correct response on an item to an examinee's ability. (See Lord 1980 for a detailed discussion.) Many different mathematical models of this functional relationship are possible. As mentioned in the preceding section, the model chosen for this study was the three-parameter logistic model. In this model the probability (P) of a correct response to item i for an individual with ability θ, $P_i(\theta)$, is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}, \qquad (1)$$

where a_i, b_i, and c_i are three parameters describing the item, and θ represents an examinee's ability.

The IRT equating method used in this study is referred to as IRT concurrent equating. (See Petersen, Cook, and Stocking 1983; also Cook and Eignor 1983 for a discussion of several IRT equating methods.) For IRT concurrent equating, all item parameter estimates for old and new test editions are calibrated in a single LOGIST run. This process results in item parameters expressed on a common scale and allows direct equating of the new and the old test editions.

Once item parameter estimates on a common scale have been obtained, there are a number of IRT equating procedures that may be used. This study, however, was concerned only with true formula score equating (Lord 1980). The expected value of an examinee's observed formula score is defined as his or her true formula score. For the true formula score, ξ, we have

$$\xi = \sum_{i=1}^{n} \frac{(k_i + 1)}{k_i} P_i(\theta) - \frac{1}{k_i}, \qquad (2)$$

where n is the number of items in the test and $(k_i + 1)$ is the number of choices for item i. If we have two tests measuring the same ability θ, then true formula scores ξ and η from the two tests are related by Equation (2), given above, and Equation (3):

$$\eta = \sum_{j=1}^{n} \frac{(k_j + 1)}{k_j} P_j(\theta) - \frac{1}{k_j}, \qquad (3)$$

where Equation (3) parallels Equation (2), but for items j (= $1 - n$). Clearly, for a particular θ, corre-

Table 4. Frequency Distributions and Summary Statistics for Verbal Operational and Equating Sections of the SAT and PAA

	Mainland Sample		*Puerto Rican Sample*	
Raw (Formula Score)	*Operational SAT*	*Equating Section*	*Operational PAA*	*Equating Section*
81–83	3			
78–80	16			
75–77	23			
72–74	28			
69–71	39			
66–68	67			
63–65	87			
60–62	122		15	
57–59	137		33	
54–56	185		90	
51–53	217		83	
48–50	277		141	
45–47	333		193	
42–44	333		201	
39–41	363		241	
36–38	416		209	
33–35	459	44	239	
30–32	441	216	260	5
27–29	405	495	219	31
24–26	404	625	233	47
21–23	362	804	164	126
18–20	327	939	152	244
15–17	263	874	166	337
12–14	245	863	94	502
9–11	190	500	80	424
6–8	128	345	45	459
3–5	78	212	20	400
0–2	29	71	7	209
−3–1	27	26	1	100
−6–4	9	2		2
−9–7	4	1		
Number of cases	6,017	6,017	2,886	2,886
Mean	33.80	17.72	32.41	10.58
SD	15.91	7.17	12.64	6.56
Correlation: Operational vs. Equating	.841		.806	
Number of items	85	34	66	34

sponding true scores ξ and η have identical meaning. They are thus said to be equated.

Because true formula scores below the chance-score level are undefined for the three-parameter logistic model, some method must be established to obtain a relationship between scores below the chance level on the two test forms to be equated. The approach used for this study was to estimate the mean and the standard deviation of below-chance-level scores on the two tests to be equated (see Lord 1980). Then these estimates were used to do a simple linear equating between the two sets of below-chance-level scores.

In practice, true score equating is carried out by substituting estimated parameters into Equations (2) and (3). Paired values of ξ and η are then computed for a series of arbitrary values of θ. Since we cannot know an examinee's true formula score, we proceed as if relationships (2) and (3) apply to the examinee's observed formula score.

The final outcome of the IRT equating of the PAA verbal and mathematical tests to the SAT verbal and mathematical tests was two conversion tables; one table relates raw scores on the PAA-verbal to raw scores on the SAT-verbal, and the second table relates raw scores

Table 5. Frequency Distributions and Summary Statistics for Mathematical Operational and Equating Sections of the SAT and PAA

Raw (Formula Score)	Mainland Sample: Operational SAT	Mainland Sample: Equating Section	Puerto Rican Sample: Operational PAA	Puerto Rican Sample: Equating Section
59–60	19			
57–58	14			
55–56	29			
53–54	33			
51–52	56			
49–50	74		5	
47–48	79		9	
45–46	106		24	
43–44	101		28	
41–42	160		4[illegible]	
39–40	160		44	
37–38	211		51	
35–36	248		92	
33–34	258		80	
31–32	300		116	
29–30	323		92	
27–28	366		153	
25–26	336		121	
23–24	416	206	159	2
21–22	369	486	190	10
19–20	341	320	183	10
17–18	370	505	201	25
15–16	321	439	178	27
13–14	320	511	229	40
11–12	248	613	205	75
9–10	257	556	183	119
7–8	200	677	165	172
5–6	187	554	90	295
3–4	136	571	90	444
1–2	64	460	47	733
−1–0	47	188	29	550
−3–2	15	82	4	272
−5–4	7	4	2	45
−7–6	1		2	2
Number of cases	6,172	6,172	2,821	2,821
Mean	24.17	10.82	19.56	2.92
SD	12.48	6.72	10.71	4.43
Correlation: Operational vs. Equating	.879		.740	
Number of items	60	23	50	23

on the PAA-mathematical to raw scores on the SAT-mathematical. Conversion tables showing the relationship between scaled scores on the respective PAA and SAT tests were derived from the raw-to-raw conversion tables. Scaled score conversions for the verbal and mathematical tests are presented in Table 6. It is clear from the Table 6 list of verbal equivalences that the difference between the two scales is as much as 180 to 185 points at a PAA score of 500. The differences are smaller at the extremes of the score scale.

The equivalences for the mathematical tests also show striking differences between the PAA and the SAT scales. In the vicinity of a PAA-mathematical score of 500 there is also a difference of 180 to 185 points. As is the case for the verbal equivalences, the differences are smaller at the extremes of the score scale.

Graphs of the verbal and mathematical IRT equating results appear in Figures 4 and 5. It is evident, even from a cursory glance at these figures that they suggest markedly curvilinear conversions between the PAA and

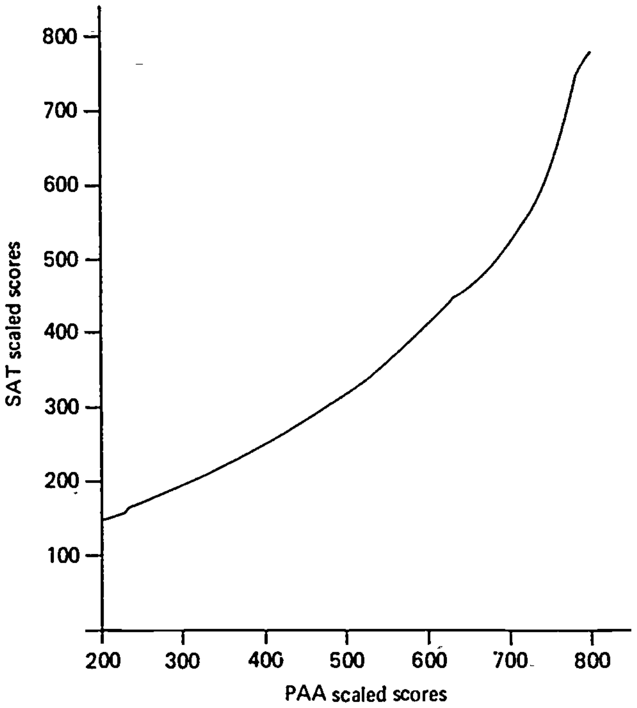

Figure 4. Item response theory conversions for verbal tests.

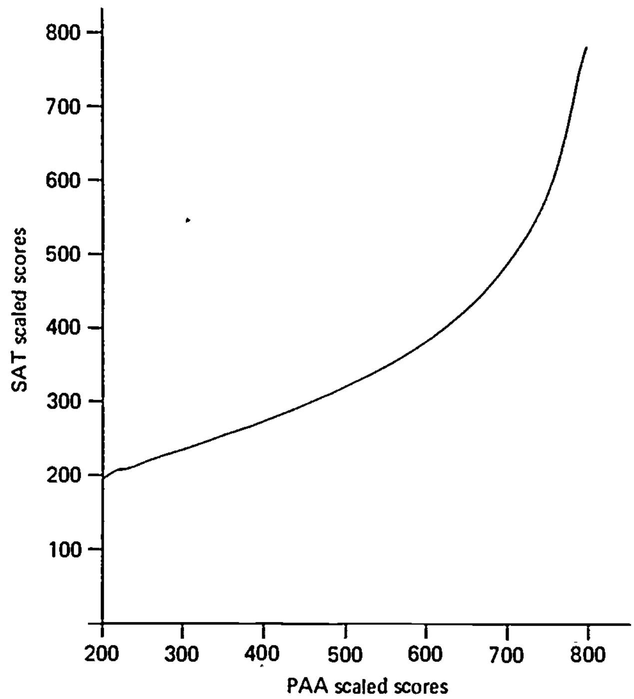

Figure 5. Item response theory conversions for mathematical tests.

SAT, typical of the results of equating two tests that differ pronouncedly in difficulty. Such conversions, which are likely to be very nearly the same irrespective of the particular method of equating employed in producing them, are seen to be concave upward when the easier test is plotted on the horizontal axis and the more difficult test on the vertical axis. In this instance the PAA is clearly the easier test, and—inasmuch as the concavity is deeper for the mathematical test—it appears also that the mathematical tests are more different in difficulty than the verbal tests.

Some attention should be given to the meaning of the differences in the PAA and SAT scales. That a 500 score on the PAA corresponds to a lower-than-500 score on the SAT simply says that if one can assume that the SAT and PAA values have been maintained precisely since the time of their inception, it can be concluded that the original scaling group for the SAT was generally more able in the abilities measured by these aptitude tests than was the original scaling group for the PAA. It does not by itself imply that the SAT candidate group today is necessarily more able than the PAA group, although this appears, in fact, to be the case. Nor does it necessarily suggest any generalization regarding the large populations from which these two examinee groups were self-selected—e.g., that the twelfth-grade students on the mainland score higher than do the twelfth graders in Puerto Rico. We know that the SAT examinee group is about one-third the size of the twelfth-grade population on the mainland and is therefore a more selective group than is its PAA counterpart, which represents a substantial proportion (over 95 percent), of the twelfth-grade population in Puerto Rico. On the other hand, this is not to say that differences between the two twelfth-grade populations do not also exist. There is some evidence, however crude, that marked differences do exist. But this evidence is outside the scope of this study.

In view of these and other possible misinterpretations of the data of this report, it will be useful to restate the limited purpose for which the present investigation was undertaken: to derive a set of conversions between two similar-appearing scales of measurement—one for tests of one language and culture, the other for tests of a different language and culture. Clearly, the accuracy of these conversions is limited by the appropriateness of the method used to derive them and the data assembled during the course of the study. It is hoped that these conversions will be useful in a variety of contexts, but (as suggested by the examples cited here) to be useful, they will need in each instance to be supported by additional data peculiar to the context.

A natural question that would arise at this point is, How well do the equivalences developed in this study compare with those developed in the 1973 Angoff-Modu study? In the earlier study, it is recalled, two linear methods were used in addition to a curvilinear method. The final conversions reported there were

Table 6. Final Conversions between PAA and SAT

*Verbal**		*Mathematical**	
PAA Scaled Scores	*Equivalent SAT Scaled Scores*	*PAA Scaled Scores*	*Equivalent SAT Scaled Scores*
800	785	800	785
787	757	790	743
774	709	779	676
761	660	768	629
749	617	758	593
736	584	747	564
723	557	736	539
710	535	726	518
697	516	715	499
684	500	704	482
672	485	694	467
659	472	683	453
646	460	672	440
633	449	662	429
625	438	651	418
617	428	640	408
609	419	630	399
601	410	619	390
592	401	608	382
584	393	598	374
576	384	587	366
568	376	576	359
560	369	566	353
552	361	555	346
544	354	544	340
535	347	534	334
527	340	523	329
519	333	512	323
511	326	502	318
503	319	491	313

*Scaled scores are not normally reported higher than 800 or lower than 200 on either the PAA or the SAT. Some scores below 200 are reported here to show the nature of the conversions near the ends of the scale.

Note: Care should be exercised in the proper use and interpretation of Table 6. See the text of this report, beginning with the second paragraph on page 16 and continuing through page 17, for a discussion of the limitations of Table 6 and for cautions regarding its possible misuses.

taken to be an average of the three, essentially weighting the curvilinear results equally with the average of the two linear results. In the present study, with the benefit of improved (item response theory) techniques for equating and with somewhat greater understanding of equating theory, it was decided to base the entire operation on the curvilinear equating as determined by the IRT procedure. The results of this study yielded substantially lower conversions to the SAT-verbal scale than was the case in the earlier study, especially in the large middle range between about 450 and about 750. The conversions to the SAT-mathematical scale showed better agreement with the earlier results. One can only speculate regarding the reasons for the agreement in the mathematical and the disagreement in the verbal. Part of it is undoubtedly attributable to a change in equating method. Another reason is the possibility of drift in the equating of either the PAA-verbal scale or the SAT-verbal scale, or both, over the intervening 12 to 15 years, causing a difference between the present results and those found in the earlier study. Yet another reason, as has been discussed in other places in this report, is that verbal equating across two languages and cultures is so much more problematic than is true of mathematical equating. In any case, we suggest that for reasons of improved methodology, the present results are probably more trustworthy than those given in the earlier, Angoff-Modu report.

Table 6. Continued

*Verbal**		*Mathematical**	
PAA Scaled Scores	*Equivalent SAT Scaled Scores*	*PAA Scaled Scores*	*Equivalent SAT Scaled Scores*
495	313	480	308
487	307	470	303
478	301	459	299
470	295	448	295
462	289	438	290
454	283	427	286
446	278	416	282
438	272	406	278
430	267	395	274
421	262	384	269
413	257	374	265
405	252	363	261
397	248	352	257
389	243	342	253
381	238	331	249
373	234	320	245
364	229	310	241
356	225	299	237
348	221	288	232
340	216	278	228
332	212	267	224
324	208	256	220
316	204	246	216
307	200	235	212
299	196	224	209
291	192	214	205
283	188	203	197
275	184	200	188
267	180		
259	176		
250	172		
242	168		
234	163		
226	158		
218	155		
210	152		
202	150		
200	148		

Note: Care should be exercised in the proper use and interpretation of Table 6. See the text of this report, beginning with the second paragraph on page 16 and continuing through page 17, for a discussion of the limitations of Table 6 and for cautions regarding its possible misuses.

SUMMARY AND DISCUSSION

The purpose of this study was to establish score equivalences between the College Board Scholastic Aptitude Test (SAT) and its Spanish-language equivalent, the College Board Prueba de Aptitud Académica (PAA). The method of the study involved two phases: (1) the selection of test items equally appropriate and useful for Spanish- and English-speaking students for use in equating the two tests and (2) the equating analysis itself. The method of the first phase was to choose two sets of items, one originally appearing in Spanish and the other originally appearing in English; to translate each set into the other language; to "back-translate" each set, independently of the first translation, into the original language; and to compare the original version of each item with its twice-translated version and make adjustments in the translation where necessary and possible. Finally, after the items were thought to be satisfactorily translated (some items that appeared to defy adequate translation were dropped), both sets of "equivalent" items—one in English and the other in Spanish—were administered, each in its appropriate language mode, for pretest purposes. These administrations were conducted in October 1984 for the PAA group and in January 1985 for the SAT group; samples of candidates took the PAA or the SAT at regularly scheduled administrations. They provided conventional psychometric indices of the difficulty and discriminating power of each item for each group. In addition, they provided two item response functions for each item, one as it appeared in Spanish and was administered to the Spanish-speaking candidates and one as it appeared in English and was administered to the English-speaking candidates. Both functions were arranged to appear on the same scale so that discrepancies between them could easily be observed. Finally, indices of agreement between the functions and measures of goodness-of-fit of the data to the item response function were also made available.

On the basis of the analyses of these data, two sets of items—one verbal and the other mathematical—were chosen and assembled as "common" items to be used for equating. In the second, or equating, phase of the study these common items, appearing both in Spanish and in English, were administered in the appropriate language along with the operational form of the SAT in December 1985 and with the operational form of the PAA in October 1986. The data resulting from the administrations of these common items were used to calibrate for differences in the abilities of the two candidate groups and permitted equating the two tests by means of the item response theory method. The final conversion tables relating the PAA-verbal scores to the SAT-verbal scores and the PAA-mathematical scores to the SAT-mathematical scores are given in Table 6. Because of the scarcity of data at the upper end of the distribution of PAA scores, score equivalences in that region are not considered highly reliable.

The general approach followed in conducting this study requires special discussion, perhaps all the more so because the method is simple, at least in its conception. On the other hand, from a psychological viewpoint the task of making cross-cultural comparisons of the kind made here is highly complex. In the extreme the task is inescapably impossible, and although the present study may represent a reasonably successful attempt, it should be remembered that the cultural differences confronted by the study were minimal and relatively easily bridged. If, for example, the two cultures under consideration were very different, then there would be little or no common basis for comparison.

Given, then, that the cultures out of which the tests in this study were developed are to some extent similar, and that there is indeed a basis for comparison, the approach and method offered do appear to have some likelihood of success. Indeed, the method itself is useful in providing a type of metric for utilizing the common basis for comparison. For example, it allows a comparison of the two cultures only on a common ground, which is to say only on those items whose item response functions were relatively similar, items that apparently had the same "meaning" in both languages and cultures. This being the case, those characteristics of the two cultures that make them uniquely different are in essence removed from consideration in making the comparisons. Thus, while we are afforded an opportunity to compare the two cultures on a common basis—i.e., on the items that are "equally appropriate"—at the same time we are also afforded an opportunity to examine the differences in the two cultures in the terms provided by the divergent, or "unequally appropriate," items. It is noteworthy that what emerges from this study is that the method described here also yields a general measure of cultural similarity, expressed in the index of discrepancy between the two item response functions. The index—rather, the reciprocal of the index—summarizes the degree to which members of the two cultures perceive the item stimulus similarly. Additional studies of the similarity of any two cultures would have to be based on other stimuli examined in a wide variety of different social contexts.

It should also be made clear that the method has its limitations, as do the results of this study, which has followed the method. For example, the present study has relied on the usefulness of translations from each of the two languages to the other, and the assumption has been made that biases in translation, if they exist, tend to balance out. This assumption may not be tenable, however. Quite possibly translation may be easier and freer of bias when it is from Language A to Language B than in the reverse direction, and if items do become somewhat more difficult in an absolute sense as a result of translation, this effect would be more keenly felt by speakers of Language A than by speakers of Language B. Also, implicit in the method of this study is the assumption that language mirrors all the significant cultural effects. This may not be so, and it is possible that the translatability of words and concepts across two languages does not accurately reflect the degree of similarity in the cultures represented by those two languages. If, for example, there are greater differences in the languages than in the cultures, then again the method is subject to some bias.

Aside from matters of methodology and possible sources of bias, a point that has been made earlier in this report deserves repeating: In this study the comparison was made between Puerto Rican and mainland United States students; the resulting conversions between the PAA and the SAT apply only between these two groups of students. Whether the same conversions would also have been found had the study been conducted between the PAA and the SAT as taken by other Spanish speakers and other English speakers is an open question. Indeed, it is an open question whether the conversion obtained here also applies to variously defined subgroups of the Puerto Rican and mainland populations—liberal arts women, engineering men, urban blacks, etc.

It is also to be hoped that the conversions between the two types of tests will not be used without a clear recognition of the realities: A Puerto Rican student with a PAA-verbal score of 503 has been found here to have an SAT-verbal score "equivalent" of 319. This is not to say that an SAT-verbal score of 319 would actually be earned were the student to take the SAT. The student might do better or might do worse, depending, obviously, on the student's facility in English. The conversions do offer a way, however, of evaluating a general aptitude for verbal and mathematical materials in terms familiar to users of SAT scores; depending on how well the student can be expected to learn the English language, the likelihood of success in competition with native English speakers in the continental United States can be estimated. Continuing study of the comparative validity of the PAA and the SAT for predicting the performance of Puerto Rican students in mainland colleges is indispensable to the judicious use of these conversions.

It will be useful, finally, to describe the ways in which the conversions may and may not be used appropriately. A glaring misuse has already been alluded to above: It would be entirely inappropriate to conclude without further consideration that a student who has earned a PAA-verbal score of 503 would therefore earn an SAT-verbal score of 319, were he or she to take it, simply because the table reads that these two scores are listed as "equivalent." As already indicated above, the student might score lower than 319, depending on his or her facility in English. Thus, intelligent use of the table requires the additional knowledge of the student's facility in English. For this purpose scores on a test like the Test of English as a Foreign Language (TOEFL), measuring the student's understanding of written and spoken English, would be useful. (Another possibility is a test, if one exists, that can accurately predict how rapidly a student is likely to learn a foreign language.) If the Spanish-speaking student's TOEFL scores are high, indicating a level of facility in English equivalent to that of a native speaker of English, these conversions may be taken at face value with appropriate cautions for their generalizability, as described earlier. If, however, the student's English-language ability is not high, the conversions given here will be inapplicable to the degree that English is an unfamiliar language to that student. Further, it would be expected that inasmuch as the SAT-verbal test depends more heavily on English language ability than does the SAT-mathematical test, the verbal conversion for the PAA to the SAT will therefore be more sensitive to the inadequacies of the student's knowledge of English than will be true of the mathematical conversion. But these guidelines are at best based only on educated intuition. As already indicated above, the continuing conduct of validity studies will yield the best guidance for the proper use of these scores.

REFERENCES

Angoff, W. H. 1966. Can useful general-purpose equivalency tables be prepared for different college admission tests? In *Testing problems in perspective*, ed. A. Anastasi. Washington, D.C.: American Council on Education, pp. 251–64.

Angoff, W. H. 1984. *Scales, norms and equivalent scores.* Princeton, N.J.: Educational Testing Service. Reprint of chapter in *Educational measurement*, 2d ed., ed. R. L. Thorndike. Washington, D.C.: American Council on Education, 1971.

Angoff, W. H., and S. F. Ford. 1973. Item-rate interaction on a test of scholastic ability. *Journal of Educational Measurement* 10:95–106.

Angoff, W. H., and C. C. Modu. 1973. Equating the scales of the Prueba de Aptitud Académica and the Scholastic Aptitude Test. Research Report 3. New York: College Entrance Examination Board.

Boldt, R. F. 1969. Concurrent validity of the PAA and SAT for bilingual Dade County high school volunteers. College Entrance Examination Board Research and Development Report 68-69, No. 3. Princeton, N.J.: Educational Testing Service.

Cook, L. L., and D. R. Eignor. 1983. Practical considerations regarding the use of item response theory to equate tests. In *Applications of item response theory*, ed. R. K. Hambleton. Vancouver: Educational Research Institute of British Columbia.

Cook, L. L., D. R. Eignor, and N. S. Petersen. 1985. A study of the temporal stability of item parameter estimates. ETS Research Report 85-45. Princeton, N.J.: Educational Testing Service.

Levine, R. S. 1955. Equating the score scales of alternate forms administered to samples of different ability. Research Bulletin 23. Princeton, N.J.: Educational Testing Service.

Lord, F. M. 1977. A study of item bias using item characteristic curve theory. In *Basic problems in cross-cultural psychology*, ed. N. H. Poortinga. Amsterdam: Swets & Vitlinger.

Lord, F. M. 1980. *Applications of item response theory to practical testing problems.* Hillsdale, N.J.: Erlbaum.

Petersen, N. S. 1977. Bias in the selection rule: Bias in the test. Paper presented at Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Petersen, N. S., L. L. Cook, and M. L. Stocking. 1983. IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics* 8:137–56.

Shepard, L. A., G. Camilli, and D. M. Williams. 1984. Validity of approximation techniques for detecting item bias. Paper presented at annual meeting of the American Educational Research Association, New Zealand.

Stocking, M. L., and F. M. Lord. 1983. Developing a common metric in item response theory. *Applied Psychological Measurement* 1:201–10.

Wingersky, M. S. 1983. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In *Applications of item response theory*, ed. R. K. Hambleton. Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S., M. A. Barton, and F. M. Lord. 1982. *LOGIST V user's guide*. Princeton, N.J.: Educational Testing Service.