Listening. Learning. Leading.®

Research Report
ETS RR–14-30

# Demographically Adjusted Groups for Equating Test Scores

Samuel A. Livingston

December 2014

# ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Demographically Adjusted Groups for Equating Test Scores

Samuel A. Livingston

Educational Testing Service, Princeton, NJ

In this study, I investigated 2 procedures intended to create test-taker groups of equal ability by poststratifying on a composite variable created from demographic information. In one procedure, the stratifying variable was the composite variable that best predicted the test score. In the other procedure, the stratifying variable was the composite that best indicated group membership (i.e., the propensity score). Applied to 2 groups taking the same test at different administrations, the composite that best predicted the test score reduced the ability difference by about two thirds; the composite that best indicated group membership reduced the ability difference by about half. Prescreening the predictor variables did not improve the performance of either procedure.

**Keywords** Equating; poststratification; propensity score; demographics; screening predictor variables

doi:10.1002/ets2.12030

Test score equating is the statistical procedure that makes it possible to report comparable scores on two or more forms of a test containing different items but measuring the same types of knowledge and skills. Equating the scores on two forms of a test requires test-taker performance data that link the two forms. There are three basic ways to obtain these data:

- administer both forms of the test to the same test takers in a way that will control for sequence effects;
- administer the two forms of the test to test-taker groups formed in a way that will make them equal in the knowledge and skills measured by the test; and
- administer, with each of the two forms of the test, a common measure of the knowledge and skills measured by the test. The common measure can be a selection of items from the test itself.

These data collection plans are commonly referred to as *equating designs* (see, for example, Kolen & Brennan, 2004, pp. 13–22).

Sometimes the circumstances of testing make it impossible to implement any of these equating designs. The purpose of the present study was to investigate a statistical procedure that might be used in such situations. This procedure uses demographic information about the test takers to transform the groups taking the two different test forms into groups of equal ability by weighting the test takers unequally. The groups created in this way can be called *demographically adjusted groups* (DAG).

## The Demographically Adjusted Groups Procedure

The DAG procedure is essentially a poststratification procedure. The strata are combinations of the demographic variables. Each group of test takers is divided separately into these strata. Within each stratum, the individual test takers in one group are assigned the weight that makes their combined weight equal to the number of test takers in that stratum in the other group. For example, if a particular stratum included twice as many test takers from Group 1 as from Group 2, each test taker in that stratum from Group 1 would receive a weight of 0.5.

With several demographic variables, it is not practical to specify a separate stratum for each possible combination (e.g., Asian American female test takers who received their bachelor's degrees 2 years before taking the test, whose fathers had a graduate or professional degree, and whose undergraduate grade average was A minus). A practical alternative is to combine the information from several demographic variables into a single composite variable, partition the composite variable into intervals, and use the intervals as strata. The composite variable is created by multiplying the test taker's score

*Corresponding author*: S. Livingston, E-mail: SLivingston@ets.org

on each demographic variable by an empirically determined coefficient (which could be zero) and summing the resulting products:

$$x_{\text{comp}} = b_1x_1 + b_2x_2 + b_3x_3 + \dots ,$$

where $x_1$ represents the test taker's score on the first demographic variable, and so on. On many demographic variables, a test taker's score is simply 1 or 0, but some demographic variables can take on three or more possible values. A composite demographic variable created in this way, for the purpose of adjusting the scores, can be called a *DAG composite*.

The study described in this article investigated two different approaches to the task of creating the DAG composite. Each approach focuses on a different part of the answer to the question, "What makes a demographic variable useful for removing differences between groups in the knowledge or skills that a test measures?" One part of the answer is that the demographic variable must be statistically related to the test scores. If the demographic variable is not statistically related to the test scores, an adjustment based on it will not make the groups more similar in the knowledge or skills that the test measures. The other part of the answer is that the two groups must differ on the demographic variable. If the unadjusted groups do not differ on the demographic variable, an adjustment based on it will have no effect; the adjusted groups will be identical to the unadjusted groups.

The first of these two approaches to the creation of the DAG composite focuses on the relationship between the demographic variables and the test scores. In this approach, the DAG composite is the linear combination of the demographic variables that best predicts performance on the test. The weights used to form the composite are the estimated regression coefficients in a linear regression analysis, in which the dependent variable is the test score. This approach will be referred to as the *composite predictor* approach.

The second approach to the creation of the DAG composite focuses on the extent to which the groups differ on the demographic variables. In this approach, the DAG composite is the linear combination of the demographic variables that best discriminates between the groups of test takers taking the two forms of the test. The weights for the individual demographic variables are the estimated regression coefficients in a logistic regression analysis, in which the dependent variable is an indicator variable: 1 for test takers taking one form; 0 for test takers taking the other form. This approach will be referred to as the *propensity score* approach (see Rosenbaum & Rubin, 1984, 1985).

The DAG composite will be useful in making the adjustment only to the extent that it meets both of the requirements specified above. It must be statistically related to the test scores, and it must discriminate between the two groups of test takers. This dual requirement suggests that the effectiveness of the DAG composite might be improved by prescreening the demographic variables before determining the weights. In the composite predictor approach, prescreening would limit the set of predictors to those on which the two groups of test takers differ substantially. In the propensity score approach, prescreening would limit the predictors to those that were substantially related to the test scores.

Both the composite predictor and the propensity score can be computed—with or without prescreening of the variables—from data that are available for a new form of a test as soon as the raw scores have been computed. In the composite predictor approach, the dependent variable in the linear regression is the raw score. In the propensity score approach, the prescreening of the demographic variables is based on their association with the raw scores. None of these analyses requires an assumption of equivalence for raw scores on different forms of the test.

## Purpose of the Study

The study was an attempt to answer the following research questions about the DAG technique:

1. To what extent can the DAG technique produce test-taker groups that are similar in the abilities the test measures?
2. How do the two approaches to creating the demographic composite compare in effectiveness?
3. How much does prescreening of the demographic variables improve the effectiveness of each approach?

The answers to these questions depend on the demographic information available. For either approach to succeed, the information must include variables that are statistically related to the test scores and differentiate between the test takers taking the different forms of the test. This report includes results for only two tests, both in the same testing program, with the same kinds of demographic information available. Therefore, the generality of the results of this study is severely limited. Nevertheless, those results are based on real, operational test-score data, not simulated data. The test scores in this study were not created by entering a set of assumptions into a computer program. They were computed from the actual responses of real test takers to real questions on real tests with real consequences.

## Design of the Study

The basic design of the study was as follows:

1. Choose a testing program in which a substantial amount of demographic information about the test takers is available.
2. In that testing program, choose two (or more) tests that are taken by several hundred test takers at each administration.
3. For each test, choose two test administrations having the following characteristics:
   - The test takers at the two administrations took the same form of the test.
   - The score distributions indicated that the groups of test takers tested at the two administrations (the "admin groups") differed substantially in the knowledge and skills measured by the test.
4. For each pair of test administrations, compute four separate versions of the DAG composite. Then use each DAG composite to create a pair of demographically adjusted admin groups. The four versions of the DAG composite were
   - a composite predictor using all available demographic variables,
   - a composite predictor using a prescreened set of demographic variables,
   - a propensity score using all available demographic variables, and
   - a propensity score using a prescreened set of demographic variables.
5. Compute the score distributions for the original admin groups and for each pair of demographically adjusted admin groups.
6. Compute statistics showing the difference between the two score distributions in each pair.

### The Tests

The study uses data from two of the *GRE*® Subject Tests: the GRE Literature in English Test and the GRE Psychology Test. Table 1 shows, for each of these tests, the number of test takers in each group and the mean and standard deviation of their test scores. The table also shows three statistics describing the extent to which the two groups of test takers differ in their performance on the test. The first statistic is the standardized difference between the mean scores of the two groups (sometimes referred to as the effect size). The second statistic is the ratio of the standard deviations of the scores of the two groups (the larger to the smaller). The third statistic is the Kolmogorov D statistic—the largest difference between the cumulative distribution functions (i.e., the largest difference between the groups in the proportion of the test takers scoring below any given score).

### The Demographic Variables

Table 2 lists the demographic variables included in the analysis. The data for these variables came from a questionnaire that the test takers completed when they registered for the test. Most of the questionnaire items translate directly into 0/1 indicator variables. The other items asked for such information as the number of years since the bachelor's degree, father's education, mother's education, major-field GPA, and overall GPA. For these items, the responses were numerical values or ordered categories. The first step in using the information from each of these multivalued items was to compute, separately for the two admin groups, the mean test score of the test takers giving each response. The second step was to examine plots of these conditional means and specify the variables to be created from the responses. Those variables are listed in Table 3.

One demographic variable created specifically for this study is the college selectivity score. The questionnaire included an item asking which college the test taker had attended as an undergraduate. That information was potentially useful for predicting test performance, but the large number of colleges made it impractical to create a separate indicator variable for each college. One practical way to use the information was to assign each college a selectivity score. This score was

**Table 1** Comparison of Groups Taking Each Test

| | Number of test takers | Mean raw score | Standard deviation of scores | Standardized mean difference | *SD* ratio | D statistic |
|---|---|---|---|---|---|---|
| Psychology | | | | | | |
| Admin. 1 | 3,954 | 110.28 | 30.57 | 0.30 | 1.16 | 0.13 |
| Admin. 2 | 1,516 | 100.48 | 35.42 | | | |
| Literature | | | | | | |
| Admin. 1 | 1,599 | 129.48 | 36.46 | 0.34 | 1.11 | 0.14 |
| Admin. 2 | 673 | 116.46 | 40.43 | | | |

**Table 2** Demographic Variables Included in the Analysis

| |
|---|
| Sex (male/female) |
| English as primary language (yes/no) |
| Preference for left-handed seating (yes/no) |
| Father's level of education (select 1 of 9 levels) |
| Mother's level of education (select 1 of 9 levels) |
| U.S. citizenship status (citizen/resident/neither) |
| Ethnicity (for U.S. citizens only: "How do you describe yourself?") |
| Undergraduate institution |
| Undergraduate grade-point average in major field (select 1 of 7 levels) |
| Overall undergraduate grade-point average (select 1 of 7 levels) |
| Years since receiving bachelor's degree |
| Plans to attend graduate school (full time/part time) |
| Eventual graduate education objective (degree) |
| Reasons for taking GRE (yes/no for each of 6 reasons listed; could choose more than one) |
| Preferred geographic region for graduate school (6 U.S. regions plus foreign regions) |

based on a list of the 54 most selective U.S. colleges, showing the 25th and 75th percentiles of the *SAT*® scores (verbal plus math) of the entering freshmen at each of those colleges (newengland, 2001). Summing those 2 percentiles for each college yielded a score that ranged from 2,620 to 2,990 for the 54 colleges on the list. Subtracting 2,610 and dividing by 10 resulted in a college selectivity score that ranged from 1 to 38 for those 54 colleges. Any college not on the list was assigned a college selectivity score of 0. Table A1 in the appendix shows the 54 colleges and their selectivity scores.

## Missing Data

One complication in this research—and in the practical application of the DAG procedure—is missing data. About two thirds of the test takers left one or more of the questionnaire items unanswered. Restricting the analysis to the test takers with complete demographic data would have resulted in a greatly reduced and possibly unrepresentative sample. Fortunately, there is an alternative—multiple imputation. The SAS manual (SAS Institute, 2000) described the procedure as follows:

> Instead of filling in a single value for each missing value, multiple imputation (Rubin 1976; 1987) replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The multiply imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining results from different data sets is essentially the same. (p. 131)

In this study, the data analyst used five imputations, creating a multiply imputed dataset with five records for each actual test taker. As a check on the stability of the results, the data analyst performed five independent replications of the multiple imputation procedure for the literature test and computed the linear regressions on the resulting five datasets. The results varied only slightly over the five replications. All further analyses used multiply imputed datasets in which missing values for the demographic variables were imputed five times.

**Table 3** Predictor Variables Created by Recoding

| Grade-point average in major field | Variable 1 | Variable 2 |
|---|---|---|
| D or lower | 1 | 1 |
| C- | 1 | 1 |
| C | 1 | 1 |
| B- | 2 | 0 |
| B | 3 | 0 |
| A- | 4 | 0 |
| A | 5 | 0 |

| Years since bachelor's degree | Variable 1 | Variable 2 |
|---|---|---|
| 0 | 1 | 0 |
| 1 | 2 | 0 |
| 2 | 3 | 0 |
| 3 | 4 | 0 |
| 4 | 5 | 0 |
| 5 | 6 | 0 |
| 6 | 7 | 0 |
| 7 or more | 8 | 1 |

| Father's education/Mother's education | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|
| No response | 0 | 0 | 1 |
| Grade school or less | 1 | 1 | 0 |
| Some high school | 2 | 0 | 0 |
| High school diploma or equivalent | 3 | 0 | 0 |
| Business or trade school | 3 | 0 | 0 |
| Some college | 4 | 0 | 0 |
| Associate degree | 4 | 0 | 0 |
| Bachelor's degree | 5 | 0 | 0 |
| Some graduate or professional school | 6 | 0 | 0 |
| Graduate or professional degree | 7 | 0 | 0 |

## Screening the Predictor Variables for the Composite Predictor

The prescreening of the demographic variables for the composite predictor was based on the extent to which the two admin groups differed on each demographic variable. The two different types of variables—1/0 indicator variables and multivalued variables—required different screening criteria.

A 1/0 indicator variable was included in the screened predictor set for the composite predictor if it met two requirements. The first requirement was based on a comparison of proportions in the two admin groups. Using the notation $p(0|A_1)$ to represent the proportion of the test takers in Admin Group 1 with a value of 0 on the indicator variable, and so on, the requirement was that one or more of the four ratios,

$$\frac{p\left(1|A_1\right)}{p\left(1|A_2\right)} \text{ or } \frac{p\left(1|A_2\right)}{p\left(1|A_1\right)} \text{ or } \frac{p\left(0|A_1\right)}{p\left(0|A_2\right)} \text{ or } \frac{p\left(0|A_2\right)}{p\left(0|A_1\right)},$$

had to be greater than 1.25. The second requirement was based on the number of test takers in each admin group making each response. Using a notation similar to that above, the requirement was that all four of the frequencies,

$$n\left(0|A_1\right) \text{ and } n\left(1|A_1\right) \text{ and } n\left(0|A_2\right) \text{ and } n\left(1|A_2\right),$$

had to be at least 25.

A demographic variable with many possible values was included in the screened predictor set for the composite predictor only if the two admin groups differed in their mean scores on the predictor variable by at least 0.25 *SD*.

The screened predictor set for the literature test included 25 of the 34 indicator variables and all 13 multivalued variables. The screened predictor set for the psychology test included 18 of the 34 indicator variables and all 13 multivalued variables.

## Screening the Predictor Variables for the Propensity Score

The prescreening of the demographic variables for the propensity score was based on the extent to which each variable was associated with the test scores. Again, the two different types of variables—1/0 indicator variables and multivalued variables—required different screening criteria.

A 1/0 indicator variable was included in the screened predictor set for the propensity score if it met two requirements. The first requirement was based on the difference between the mean test score of the test takers with a 1 on the variable and the mean test score of the full admin group. This difference had to be at least 0.1 *SD* in both admin groups and in the same direction in both groups. The second requirement was that all four of the frequencies,

$$n\left(0|A_1\right) \text{ and } n\left(1|A_1\right) \text{ and } n\left(0|A_2\right) \text{ and } n\left(1|A_2\right),$$

had to be at least 25.

The criterion for including a demographic variable with many possible values was based on its correlation with the test scores in the two admin groups. The variable was included in the screened predictor set for the propensity score if the size of the correlation was at least .10 in both admin groups and the direction of the correlation was the same in both groups.

The screened predictor set for the literature test included 8 of the 34 indicator variables and 8 of the 13 multivalued variables. The screened predictor set for the psychology test included 11 of the 34 indicator variables and 8 of the 13 multivalued variables.

## Creating the Weighted Score Distributions

The weighted score distributions based on a composite predictor were determined by the following procedure:

1. Perform a linear regression analysis in which the dependent variable was the examinee's test score and the independent variables were admin group membership and the demographic variables.
2. Use the regression coefficients from Step 1 as weights to create a composite predictor variable. Compute each test taker's score on the composite predictor.
3. Compute the distribution of scores on the composite predictor in each of the two admin groups.
4. Divide the range of the composite predictor into intervals.
5. Separately, for each interval, determine the number of test takers in each admin group whose composite predictor scores fall into that interval.
6. Assign a weight of 1 to each test taker in the interval who is a member of the admin group with the smaller frequency in the interval.
7. To each test taker in that interval who is a member of the admin group with the larger frequency in the interval, assign a weight equal to the ratio of the smaller frequency to the larger frequency. This weight will make the sum of the weights for test takers in the interval the same in the two admin groups.

The weighted score distributions based on a propensity score were determined by a similar procedure:

1. Perform a logistic regression analysis with admin group membership as the dependent variable and the demographic variables as independent variables.
2. Use the logistic regression coefficients from Step 1 as weights to create a composite variable. This variable is the propensity score. Compute each test taker's propensity score.
3. Compute the distribution of the propensity score in each of the two admin groups.
4. Divide the range of the propensity score into intervals.
5. Separately for each interval, determine the number of test takers in each admin group whose propensity scores fall into that interval.
6. Assign a weight of 1 to each test taker in the interval who is a member of the admin group with the smaller frequency in the interval.
7. To each test taker in that interval who is a member of the admin group with the larger frequency in the interval, assign a weight equal to the ratio of the smaller frequency to the larger frequency. This weight will make the sum of the weights for test takers in the interval the same in the two admin groups.

Each of these procedures was implemented twice: once with all the demographic variables and once with only those variables that remained after prescreening. Each of these four analyses yielded a set of weights for the individual test takers. Applying the weights to the two admin groups produced a pair of demographically adjusted groups. The test score distributions in these adjusted groups indicated the extent to which the adjustment was successful at eliminating the ability difference between the two admin groups.

## The Results

The objective of the demographic adjustment was to create adjusted admin groups that were equal in the knowledge and skills measured by the test. Since both admin groups in this study actually took the same form of the test, a comparison of the raw score distributions showed the extent to which the adjustment achieved this objective.

Figure 1 shows the difference between the cumulative raw score distributions in the two admin groups taking the same form of the literature test. The figure contains five separate curves. One curve—the highest—shows the difference between the distributions in the two unadjusted admin groups. Each of the other curves shows the same information for one of the four pairs of adjusted groups. The difference between the score distributions in the unadjusted groups was substantial—greater than 8% throughout most of the score range and reaching a maximum of 14.8%. The adjustment based on the propensity score decreased the difference by about half. Prescreening improved this adjustment in the middle portion of the score range but tended to have the opposite effect in the high-middle and low-middle portions of the range. The adjustment based on the composite predictor decreased the difference between the admin groups by a larger amount, particularly in the middle and upper-middle portions of the score range. It decreased the difference between the groups by about two thirds to three fourths. Prescreening, which removed only a few predictor variables from this analysis, had essentially no effect.

Figure 2 presents the results for the psychology test. They are similar to the results for the literature test, except that the propensity score approach was somewhat more effective, particularly in the higher portion of the score range.

Table 4 shows some summary statistics comparing the score distributions in the five pairs of admin groups: the unadjusted groups and the four pairs of demographically adjusted groups.

The first summary statistic is the weighted absolute difference between the cumulative distributions. This statistic is the size of the difference in the cumulative proportion, computed at each score level, weighted by the proportion of test takers' scoring at that level (in the two admin groups combined), and averaged over score levels. This weighted difference between the unadjusted groups was about .09 on the literature test and .08 on the psychology test. The adjustment based on the propensity score reduced it to about .05 and .03. The adjustment based on the composite predictor reduced it to about .03 and .02.

Another statistic of interest is the largest single difference between the cumulative distributions (the Kolmogorov D statistic). This "worst case" statistic for the unadjusted groups was nearly .15 on the literature test and .13 on the psychology test. The adjustment based on the propensity score reduced it to about .09 and .07. The adjustment based on the composite predictor reduced it to about .07 and .05.

The standardized mean difference between the unadjusted groups taking the same test form was 0.34 for the literature test and 0.30 for the psychology test. The adjustment based on the propensity score reduced it to 0.19 and 0.12. The adjustment based on the composite predictor reduced it to 0.07 and 0.05.

The ratio of the larger standard deviation to the smaller standard deviation was 1.11 for the literature test and 1.16 for the psychology test. The adjustment based on the propensity score reduced it to 1.06 and 1.12. The adjustment based on the composite predictor reduced it to 1.07 and 1.10.

In general, the effects of prescreening the demographic variables were small. For the psychology test, prescreening actually made the adjustment procedures slightly less effective.

## Summary and Discussion

This small-scale study included only two tests, both from the same testing program. The results for these two tests were similar. The adjustment based on the propensity score reduced the between-group differences by about half; the adjustment based on the composite predictor reduced the between-group differences by about two thirds. Prescreening the demographic variables did not increase the effectiveness of either adjustment procedure.

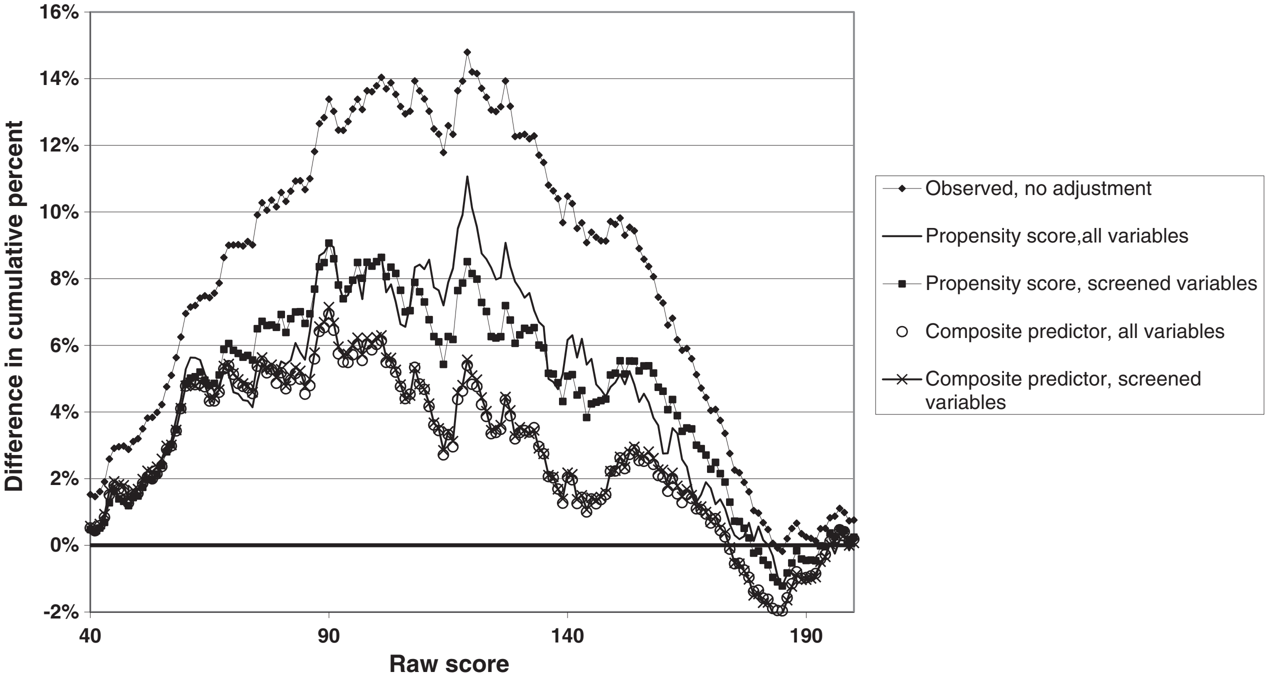

**Figure 1** GRE Literature in English: Difference between raw score distributions in Admin Groups 1 and 2.

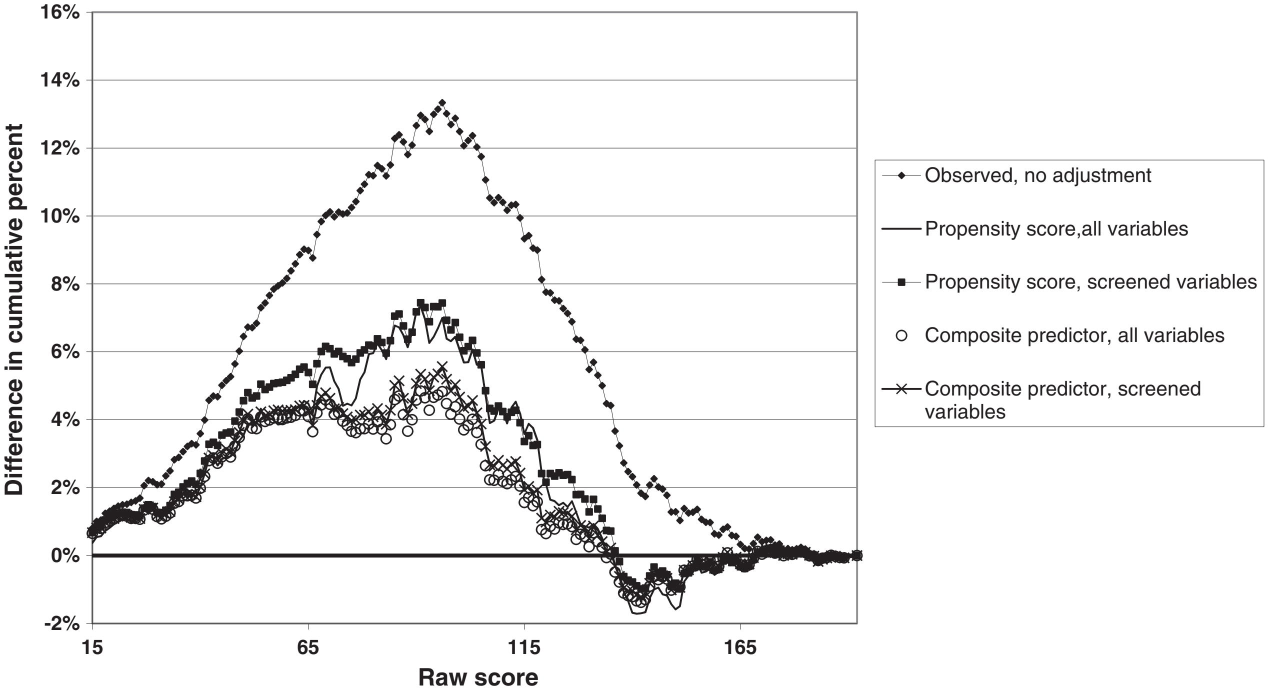

**Figure 2** GRE Psychology: Difference between raw score distributions in Admin Groups 1 and 2.

These results indicate that a demographic adjustment would not be an adequate substitute for an equating based on one of the generally accepted equating designs (e.g., equivalent groups produced by alternating the test forms). However, for the tests and test-taker populations in this study, if none of the generally accepted equating designs could be implemented, the demographic adjustment would be far better than assuming the unadjusted groups to be equal in the knowledge and skills measured by the test.

**Table 4** Statistics Describing Differences Between Score Distributions for Admin Groups

| Test and type of adjustment | Weighted absolute difference in cumulative proportion | Largest difference in cumulative proportion (D statistic) | Standardized mean difference | Ratio of standard deviations |
|---|---|---|---|---|
| Literature test | | | | |
| No adjustment | .094 | .148 | 0.34 | 1.11 |
| Propensity score, without prescreening | .055 | .111 | 0.19 | 1.06 |
| Propensity score, with prescreening | .053 | .091 | 0.19 | 1.06 |
| Composite predictor, without prescreening | .032 | .069 | 0.12 | 1.07 |
| Composite predictor, with prescreening | .033 | .071 | 0.12 | 1.08 |
| Psychology test | | | | |
| No adjustment | .079 | .133 | 0.30 | 1.16 |
| Propensity score, without prescreening | .033 | .073 | 0.12 | 1.12 |
| Propensity score, with prescreening | .037 | .074 | 0.14 | 1.12 |
| Composite predictor, without prescreening | .023 | .048 | 0.09 | 1.10 |
| Composite predictor, with prescreening | .026 | .056 | 0.10 | 1.10 |

The similarity of the results from these two tests suggests that the findings of this study would generalize to other GRE Subject Tests. Whether they would generalize to other kinds of test-taker populations and to other sets of available demographic information remains an open question.

## References

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

newengland. (2001). *A new ranking of American colleges on laissez-faire principles, 1999–2000.* Retrieved from http://collegeadmissions.tripod.com

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33–38.

SAS Institute. (2000). *SAS OnlineDoc*™( Version 8, p. 131). Cary, NC: Author.

## Appendix

**Table ??** College Selectivity Score

| College | Selectivity score |
|---|---|
| Caltech | 38 |
| Harvard | 37 |
| MIT | 35 |
| Harvey Mudd | 31 |
| Princeton | 29 |
| Stanford | 29 |
| Yale | 29 |
| Dartmouth | 26 |
| Swarthmore | 24 |
| Pomona | 23 |
| Rice | 23 |
| Amherst | 19 |
| Williams | 19 |
| Brown | 18 |
| Duke | 18 |
| Columbia | 17 |
| U. of Pennsylvania | 17 |

**Table ??** Continued.

| College | Selectivity score |
|---|---|
| Johns Hopkins | 16 |
| Middlebury | 16 |
| Cooper Union | 14 |
| Haverford | 14 |
| Carnegie Mellon | 13 |
| Northwestern | 13 |
| Emory | 12 |
| Carleton | 11 |
| Cornell | 10 |
| Georgetown | 10 |
| U. of Chicago | 10 |
| Columbia Engineering | 10 |
| Grinnell | 9 |
| Claremont McKenna | 8 |
| California--Berkeley | 7 |
| Reed | 7 |
| Washington and Lee | 7 |
| Macalester | 6 |
| New College (FL) | 6 |
| Tufts | 6 |
| Washington Univ. | 6 |
| Bowdoin | 5 |
| Davidson | 5 |
| Wellesley | 5 |
| Wesleyan U. (CT) | 5 |
| Vanderbilt | 4 |
| Vassar | 4 |
| Barnard | 3 |
| Notre Dame | 3 |
| Bates | 2 |
| Brandeis | 2 |
| Bryn Mawr | 2 |
| NYU | 2 |
| Colby | 1 |
| Georgia Tech | 1 |
| US Naval Academy | 1 |
| U. of Virginia | 1 |
| All others | 0 |

## Suggested citation:

Livingston, S. A. (2014). *Demographically adjusted groups for equating test scores* (ETS Research Report No. RR-14-30). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12030

**Action Editor:** Gautam Puhan

**Reviewers:** Charles Lewis and Neil Dorans