Journal of Educational Measurement
Spring 2009, Vol. 46, No. 1, pp. 104–128

A Comparison of Bridging Methods in the Analysis of NAEP Trends with the New Race Subgroup Definitions

J. Patrick Meyer
University of Virginia
J. Carl Setzer
GED® Testing Service

Recent changes to federal guidelines for the collection of data on race and ethnicity allow respondents to select multiple race categories. Redefining race subgroups in this manner poses problems for research spanning both sets of definitions. NAEP long-term trends have used the single-race subgroup definitions for over thirty years. Little is known about the effects of redefining race subgroups on these trends. Bridging methods for reconciling the single and multiple race definitions have been developed. These methods treat single-race subgroup membership as unknown or missing. A simulation study was conducted to determine the effectiveness of four bridging methods: multiple imputation logistic regression, multiple imputation probabilistic whole assignment, deterministic whole assignment—smallest group, and deterministic whole assignment—largest group. Only the first of these methods incorporates covariate information about examinees into the bridging procedure. The other three methods only use information contained in the race item response. The simulation took into account the percentage of biracial examinees and the missing data mechanism. Results indicated that the multiple imputation logistic regression was often the best performing method. Given that all K-12 and higher education institutions will be required to use the multiple-race definitions by 2009, implications for No Child Left Behind and other federally mandated reporting are discussed.

Since 1977, Federal requirements have stipulated the use of five race and ethnicity categories. These five categories represent mutually exclusive race and ethnicity subgroups. Recent Federal guidelines mandated a change to the method of collecting data on race and ethnicity. National surveys such as the National Assessment of Educational Progress (NAEP) are now required to allow respondents to select multiple (nonmutually exclusive) race categories (Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, 1997). The new race and ethnicity categories ensure accurate representation of Americans. However, they create some technical difficulties in the analysis of race and ethnicity subgroup differences. One particular area of difficulty concerns the analysis of race and ethnicity subgroup trends. NAEP long-term trends date back to the 1970s. Trends in academic achievement are reported for the entire nation and for the White[1], Black, and Hispanic subgroups (Campbell, Hombo, & Mazzeo, 2000). The introduction of multiple-race subgroup categories essentially redefined all subgroups and complicated research spanning both sets of definitions. This study compared methods for bridging the 1997 multiple-race definitions to the 1977 single-race definitions. NAEP reading long-term trend data were used to identify background items that predicted single-race subgroup membership. A simulation study was then conducted to evaluate the effectiveness of each bridging method.

Bridging Methods

Bridging methods assign multiple-race examinees to a single-race subgroup according to some rule or procedure. Some bridging methods use partial or fractional assignment and others use whole assignment. Partial assignment fractionally assigns multiple-race examinees to all race subgroups selected. For example, a White/Black biracial examinee may be placed into the White subgroup with weight .8 and the Black subgroup with weight .2. The choice of weights depends on the particular bridging method used. Whole assignment bridging methods assign a multiple-race examinee exclusively to one subgroup.

Noncovariate Bridging Methods

Noncovariate bridging methods only use information included in the race item response. These methods were developed by the Tabulation Working Group Interagency Committee for the Review of Standards for Data on Race and Ethnicity and endorsed by the Office of Management and Budget (OMB). Four methods were proposed: (a) deterministic whole assignment, (b) deterministic fractional assignment, (c) probabilistic whole assignment, and (d) all-inclusive (Office of Management & Budget [OMB], 2000). Deterministic whole assignment uses fixed, deterministic rules to assign multiracial respondents to one and only one of the single-race subgroups. Two types of deterministic whole assignment are smallest group (DSM) and largest group (DLG). DSM deterministically assigns a multiple-race examinee to the smallest race subgroup selected. Conversely, DLG assigns a multiple-race examinee to the largest race subgroup selected. Deterministic fractional assignment allocates individuals to each of the single-race subgroups but weighted proportionally to the number of subgroups selected, where the fractions, or weights, sum to one across groups for each person. Probabilistic whole assignment (PWA) stochastically classifies respondents into one and only one single-race subgroup. Each multiracial respondent is assigned to one single-race subgroup with a probability equal to the inverse of the number of race subgroups selected. All-inclusive assignment places multiracial individuals in all single-race subgroups that were selected. This method permits respondents to be placed in multiple categories. As such, multiracial examinees are counted multiple times.

Researchers examined the performance of noncovariate bridging methods using data from the 1993, 1994, 1995 National Health Interview Survey (NHIS); the 1995 Supplement to the Current Population Survey (CPS); and the 1998 Washington State Population Survey (WSPS) (OMB, 2000). All these data sets contained responses to race and ethnicity questions defined by the 1977 and the 1997 standards. Their analysis used the 1997 race and ethnicity question to assign multiracial respondents into single-race subgroups. Their classifications were then compared to the participants' responses to the 1977 single-race question. Misclassification occurred if a person was classified into a single-race subgroup that was different from their response to the single-race question. Across all data sets misclassification rates were highest for the smaller single-race subgroups. In the NHIS data misclassification rates were between 0% and 13% across bridging methods for the American Indian/Alaskan Native (AIAN) subgroup, but only between 0% and 1% for the White and Black subgroups. In the WSPS and CPS data, AIAN misclassification rates were as high as 30% and 40% for some noncovaraite bridging methods. For these data, the White subgroup misclassification rates were no larger than 3% and the Black subgroup misclassification rates were no larger than 9%.

In addition to misclassification rates, OMB (2000) considered the effect of misclassification on an outcome measure (a variable other than race) and found that misclassification had little impact. To the contrary, when the all-inclusive method was applied to NAEP data, some statistically significant changes in test scores were noted among the smaller race subgroups (Meyer & Hombo, 2003). Moreover, bridged effect sizes tended to deviate from their true values as the proportion of multiracial examinees increased (Meyer & Huynh, 2008). Unlike the noncovariate methods, this effect was small for covariate bridging methods.

Covariate Bridging Methods

Covariate bridging methods use race subgroup covariates to predict and assign single-race subgroup membership to multiple-race examinees. Discriminant analysis and logistic regression have been used for this purpose (Meyer, McClellan, & Huynh, 2004; Schenker & Parker, 2003). By taking familial, cultural, and other background characteristics into account, covariate bridging methods have shown to be more effective than noncovariate methods.

Schenker and Parker (2003) found that covariate methods were better at estimating race subgroup proportions. Meyer and Huynh (2008) also found that subgroup sample size was best estimated using partial assignment covariate bridging methods. When the outcome of interest was a test score, whole and partial assignment methods performed similarly in their study. The primary factor that affects the effectiveness of bridging was found to be the percentage of multiracial examinees: Sample size and choice of statistical method (e.g., logistic regression versus discriminant analysis) had little effect on the effectiveness of bridging.

Single imputation (Meyer & Huynh, 2008) and multiple imputation (Schenker & Parker, 2003) methods have shown to be effective bridging methods. The latter method has the added benefit of accounting for error due to imputation. Multiple imputation of single-race using logistic regression is accomplished in two steps as detailed in Schenker and Parker (p. 1579). First, background items are used to predict single-race using data from single-race examinees only. The estimated logistic regression coefficients, $\hat{\beta}$, and the estimated covariance matrix, $\hat{\Sigma}$, are used to obtain a vector of logistic regression coefficients from a multivariate normal distribution. Specifically, the coefficients follow an approximate posterior distribution, $\beta_p \sim MVN(\hat{\beta}, \hat{\Sigma})$, where β_p is a vector of logistic regression coefficients drawn from this distribution. Second, for each multiple-race examinee (a) compute the probability of single-race subgroup membership using β_p, and (b) draw a single-race using the probabilities from part (a). Steps (a) and (b) should be repeated M times to obtain multiple imputations of single-race subgroup membership. Partial assignment uses the probabilities from (a) as weights in subsequent analyses. Whole assignment bridging methods round the probabilities from (a) or use the drawn race category from (b), thereby exclusively assigning multiple-race examinees to one subgroup. The analysis of multiple single-race imputations proceeds by computing the statistic of interest (e.g., a mean) M times. A summary of the M statistics and its variance is then computed using formulas provided by Rubin (1987, p. 76).

Note that bridging must be conducted for each multiracial group separately. For example, examinees who selected White and Asian would participate in one bridging analysis along with a random selection of examinees who selected either the White or Asian subgroup but not both. The latter groups of examinees would serve as the single-race training sample. Random selection of these examinees would be necessary to prevent including an examinee in multiple analyses. A separate bridging analysis would be conducted for examinees who selected White and Black with a random sample of examinees who exclusively selected either of these subgroups.

When using multiple imputation, the type of missing data mechanism must be considered. Rubin (1976) defined three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR is assumed to occur when the missingness is not related to the data that are missing and not related to other variables in the data set. This type of missingness would occur when a data point is deleted according to the flip of a fair coin. MAR occurs when missingness is related only to other variables in the data, not the data that are missing after controlling the other variables. Finally, MNAR describes the situation in which missingness is related to that which is missing. The only way to know why something is missing is to know what is missing. MNAR is the most problematic type of missing data mechanism. An additional description of MCAR, MAR, and MNAR may be found in Schafer and Graham (2002).

National Assessment of Educational Progress

The NAEP long-term trend assessment in reading began in 1971. It is administered to 9-, 13-, and 17-year-old students. The test is composed of a variety of reading materials such as stories, poems, and essays (Campbell et al., 2000). Students' understanding of these materials is measured with multiple-choice and constructed response items. Results are reported separately for each age group. Overall performance as well as race and gender subgroup performance is reported. NAEP uses a complex sampling design in order to obtain a nationally representative data set and to ensure that minority groups are adequately represented.

NAEP data are collected using a multistage cluster sampling design (Allen, Carlson, & Zelenak, 1999), and weighting is used in the analysis to adjust for the unequal probability of selection. One particular weight available in NAEP secondary use data files is ORIGWT, which adjusts the data to be representative of the population. These weights are a function of the base weight (inverse of the probability of selection), a nonresponse adjustment, and a poststratification weight.

Standard errors that are calculated without accounting for the complex sampling design tend to be underestimated. NAEP uses a jackknife procedure to produce standard errors that account for the sampling design. The procedure uses replicate weights that are created by successively dropping a single primary sampling unit (PSU) from a stratum and reweighting the remaining PSUs in that stratum. An analysis is repeated *K* times, where *K* represents the number of replicates weights. The variance is then estimated by summing the squared deviation of each replicate estimate from the ORIGWT estimate. Further details concerning the NAEP jackknife procedure are provided in the NAEP technical report (Allen et al., 1999).

NAEP long-term trends are evaluated using linear and quadratic tests of trend. These tests are essentially unweighted regressions of the average scale scores on the assessment year, and the square of the assessment year, respectively (Allen et al., 1999). The ratio of each coefficient to its standard error tests the statistical significance of the trend. Linear trends describe an increasing or decreasing order of means across multiple years. Quadratic trends reflect scores that increase over several years followed by a change in the rate of increase. A quadratic trend may also occur when scores decrease over several years but then begin to increase.

Method: Selection of Background Items

The data used for this analysis were taken from the 1990, 1992, 1994, and 1996 NAEP reading long-term trend assessments of 9-year-old students. These 4 years were chosen because they were the most recent 4 years of restricted-use NAEP data made available to researchers. The youngest age group was selected because multiracial groups are more prevalent among younger populations (Lopez, 2003; Meyer & Hombo, 2003).

The first step in identifying background items was to compare the data files from each of the 4 years and determine which background items were included in each of the data sets. Furthermore, selected items were limited to those that were worded exactly the same for each year. This process resulted in the selection of sixteen background items. It is important to note that some of the background items that were *not* consistently included in all data files may actually do a better job of predicting subgroup membership within any given year. From this initial pool of items, further steps were taken to narrow the selection of background items.

The goal was to select approximately ten background items that could be effectively used for multiple imputation. Some background items were selected according to previous research on the analysis of multiracial data. Jones and Smith (2001) found that the population of multiracial people varies considerably by region. The number of books in the home and the level of urbanicity were also found to be related to single-race subgroup membership (Meyer et al., 2004). According to these findings, the REGION, URBAN, and number of books (B000904B) background items were selected for the analysis.

The selection and exclusion of other background items was somewhat arbitrary, but the goal was not to identify all predictors of race and ethnicity. Items thought to be unrelated to race and ethnicity, such as number of teachers and instructional dollars per pupil were eliminated. Other items were excluded because they were thought to be highly correlated with ones already selected. For example, the percentages of White, Black, and Hispanic students were likely correlated with region, school type, and urbanicity. Whether or not a dictionary or encyclopedia is found in a home is

TABLE 1
Description of Background Items Included in Logistic Regression

Variable	Description
Southeast	Region of the country: Southeast
Central	Region of the country: Central
West	Region of the country: West
Private	School type: Private
Catholic	School type: Catholic
Suburban	Urbanicity: Suburban
Rural	Urbanicity: Rural
Newspaper	Does your family get a newspaper regularly?
Books	Are there more than 25 books in your home?
Television	How much television do you usually watch each day?

likely correlated with the presence of 25 or more books in the home. The final list of selected items is shown in Table 1.

The six background items were recoded prior to running the logistic regression analyses. Five of the items were categorical in nature and were dummy-coded into *C-1* new variables, where *C* represents the number of response options. The dummy coded background items are listed in Table 1. The item named "Television" was treated as an eight category ordinal variable.

Logistic regression analyses were performed separately for the 1990, 1992, 1994, and 1996 data sets. In each model, the dichotomous dependent variable was Black versus White, where White was used as the reference group. All background items selected for the analysis were included in the models as predictors of the dependent variable. The logistic regression estimates used the ORIGWT weighting variable and the standard errors were computed using the NAEP jackknife procedure.

Results: Selection of Background Items

The estimated coefficients and standard errors for each year are shown in Table 2. The analysis focused on individual predictors, and no assessment of the overall model fit was performed. Two features exhibited in Table 2 are worth noting. First, there appears to be a level of consistency among the standard errors for each background item across all 4 years. For example, the standard error for Television is .03 for 1990–1994 and .04 for 1996. The standard error for the Private variable was somewhat high (i.e., exceeded a value of one) in both 1990 and 1994. All other standard errors ranged from .03 to .73. Second, there was little consistency in the significance of the variables across years. In fact, only the Suburban, Rural, Books, and Television variables remained significant across all 4 years. Variation in the statistical significance of predictors may not be limited to data from multiple years but may also occur for different subgroups within a year. Schenker and Parker (2003) indicated that their predictors performed differently for each multiple-race group (e.g., White/Asian, White/Black). Consequently, selecting the best predictors within a year

TABLE 2
NAEP Grade 4/Age 9 Years Reading Trend Logistic Regression Estimates and Standard Errors (in Parentheses)

	Year			
Variable	1990	1992	1994	1996
Intercept	−.10(.39)	−.85(.42)*	−.60(.31)	−1.76(.39)*
Southeast	.95(.36)*	.28(.41)	1.07(.38)*	1.16(.38)*
Central	−.45(.37)	−.48(.37)	−1.07(.40)*	.00(.37)
West	−.84(.33)*	−.75(.40)	−.48(.35)	−.27(.38)
Private	−14.43(1.27)*	−1.22(.36)*	−1.39(1.42)	−.74(.44)
Catholic	−.77(.72)	−1.63(.46)*	−.61(.56)	−.69(.55)
Suburban	−1.77(.38)*	−1.07(.32)*	−1.57(.34)*	−1.14(.27)*
Rural	−2.11(.46)*	−2.77(.49)*	−2.49(.55)*	−2.16(.44)*
Newspaper	−.28(.10)*	−.21(.13)	−.19(.15)	−.02(.12)
Books	−1.28(.14)*	−1.06(.11)*	−1.36(.18)*	−.95(.15)*
Television	.25(.03)*	.32(.03)*	.32(.03)*	.36(.04)*

*($p < .05$).

and even within a multiple-race group may be a better way to identify background items for bridging.

The final set of background items selected for the simulation study included Southeast, Suburban, Rural, Books, and Television. All but the last of these items were dummy coded versions of the responses to the original background items.

Method: Simulation Study

Only whole assignment methods (LR, PWA, DSM, and DLG) were used in the simulation study. Whole assignment is easier to implement when analyzing NAEP data because partial assignment methods would require an adjustment to the NAEP sampling weights.

Conditions of the simulation included (a) the percent biracial, and (b) the type of missing data mechanism. For the trend analysis a pattern of percent biracial condition was also included. The simulation was designed to answer four questions:

1. What bridging method results in subgroup mean and trend coefficient estimates that are most similar to the true single-race subgroup estimates?
2. How accurate are coverage values from the bridging methods?
3. How does the missing data mechanism affect the results?
4. How does the size of the biracial group affect the effectiveness of bridging?

Extant NAEP data served as the population from which random samples were taken. An examinee's score was the average of five plausible values. The simulation included independent random samples from the 1990, 1992, 1994, and 1996 NAEP reading long-term trend data for 9-year-olds. Each data set was restricted to only include examinees from the White or Black race subgroups, where race subgroup was indicated by the derived race variable (DRACE[2]). The Black subgroup comprised

TABLE 3
Unweighted Descriptive Statistics for NAEP Reading Trends

Year	Subgroup	N^{a}	Mean	*SD*
1990	White	3690	22.29	38.00
	Black	940	182.52	35.67
	Difference	–	37.77	–
1992	White	4680	22.95	32.70
	Black	1020	188.54	33.86
	Difference	–	32.41	–
1994	White	3250	22.64	33.19
	Black	730	188.81	34.16
	Difference	–	31.83	–
1996	White	2890	222.93	32.64
	Black	810	194.82	33.52
	Difference	–	28.11	–

[a] All sample size numbers are rounded to the nearest ten due to National Center for Education Statistics confidentiality policy.

about 20% of the data for each year. Across years the White-Black achievement gap ranged from 37 to 28 points, with the gap decreasing across years (see Table 3).

The percent biracial condition had five levels: 2%, 5%, 8% 10%, and 15%. Examinees were identified as biracial by two types of missing data mechanisms. MCAR was simulated by taking a random draw from a uniform distribution on the unit interval for each examinee using SAS's RANUNI(0) function (SAS Institute, 2006). If the examinee's draw (x 100) was less than or equal to the stated level of percent biracial the examinee's DRACE was ignored and treated as missing. A check on this part of the simulation confirmed that the data were MCAR. The average effect size, $\mathbf{w} = \sqrt{\chi^2/N}$ (Cohen, 1988), between DRACE and a missing data indicator was .03 or lower across all conditions suggesting a mild or small relationship (details not shown). The average effect size between the five background items and the missing data indicator was also around .03 with a standard deviation of .02 (details not shown). The exception was the Television item that commonly had an average effect size around .08. The extent to which these effect sizes are considered small or negligible determines whether the assumption of MAR or MCAR is more tenable. MCAR was assumed herein, but both assumptions imply an ignorable missing data mechanism.

MNAR was also simulated using random draws from a uniform distribution. However, For the MNAR-W condition, only White examinees were permitted to be included in the biracial subgroup. This qualification was designed to have the missing data mechanism be related to the single-race subgroup indicator variable. The average effect size between DRACE and a missing data indicator ranged from .07 to .21 across the levels of percent biracial. For the MNAR-B condition, only Black examinees were permitted to be included in the biracial subgroup. The average effect size for MNAR-B ranged from .28 to .84. These results suggest that the simulation established MNAR and that MNAR was more prominent in the MNAR-B than the

MNAR-W condition. The effect sizes were not only larger than those for the MCAR condition, but they increased in size as the percent multiracial increased.

The pattern of percent biracial was the last condition of the simulation. It was used in the analysis of subgroup trends. The two patterns were even and sudden. Even biracial patterns had the same percent biracial across all 4 years of data (e.g., 5%). Sudden patterns had 2 years of the same biracial percent with a large increase in the third year (2%, 2%, 8%, and 10%). The latter case was designed to mimic the realistic situation where the selection of multiple race categories is permitted mid-trend.

All ($5 \times 2 = 10$) simulation conditions were created independently for each year of NAEP data. Within years, each condition was replicated by taking 1,000 random samples of $N = 1{,}000$ examinees. The percent biracial was established in each random sample.

Simulated data were analyzed using all four bridging methods. For the multiple imputation procedures (LR and PWA), $M = 10$ imputations were used. As in Schenker and Parker (2003), this study only conducted steps 1 and 2a of the multiple imputation procedure. NAEP reading score means and standard errors were estimated for the bridged subgroups. The bridged mean estimates were compared to their true estimates (i.e., subgroup means based on the DRACE variable) using the root mean squared error,

$$RMSE = \sqrt{\sum_{j=1}^{1000} (bridged_j - true_j)^2/1,000},$$

and bias,

$$bias = \sum_{j=1}^{1000} (bridged_j - true_j)/1,000.$$

In addition, coverage was evaluated for each bridging method. The bridged mean and standard errors were used to form 95% confidence intervals for each subgroup mean. Coverage was the proportion of confidence intervals that contained the subgroup mean population parameter (Table 3), where the proportion was computed across the 1,000 replications.

Linear and quadratic tests of race subgroup trends were evaluated using standard NAEP procedures. Subgroup means, standard errors, and the years of data collection are all the information necessary to conduct a trend analysis. Consequently, the bridged trend analysis was conducted by using the bridged means and standard errors. Formulas for computing the linear and quadratic tests of trend (Allen et al., 1999, pp. 374–375) resulted in linear coefficients of −.15, −.05, .05, and .15. The quadratic coefficients were .0625, −.0625, −.0625, and .0625. These coefficients are proportional to the coefficients commonly used for trend analysis when data are collected at equal intervals (Kirk, 1995, p. 814). Estimates of bridged subgroup trend coefficients and their standard errors were also compared to the true subgroup estimates using the RMSE.

TABLE 4
RMSE for Bridged Means, MCAR Condition

Percent		Logistic Regression[a]		PWA[a]		DSM		DLG	
Biracial	Year	White	Black	White	Black	White	Black	White	Black
2	1990	.15	.34	.16	1.57	.20	2.94	.23	.36
	1992	.12	.33	.13	1.56	.17	2.91	.17	.36
	1994	.12	.32	.12	1.46	.16	2.74	.18	.36
	1996	.12	.30	.13	1.06	.16	2.00	.19	.31
5	1990	.29	.55	.31	3.59	.31	6.37	.52	.56
	1992	.25	.52	.24	3.53	.26	6.21	.39	.59
	1994	.26	.56	.25	3.45	.27	6.03	.40	.61
	1996	.24	.48	.26	2.46	.27	4.38	.43	.52
8	1990	.44	.71	.46	5.41	.38	9.19	.79	.72
	1992	.37	.68	.35	5.30	.33	8.83	.61	.76
	1994	.37	.66	.37	5.12	.35	8.54	.60	.76
	1996	.37	.63	.40	3.74	.35	6.37	.66	.68
10	1990	.52	.79	.57	6.63	.44	1.84	.98	.85
	1992	.44	.76	.43	6.35	.38	1.24	.73	.82
	1994	.46	.74	.43	6.15	.38	9.96	.74	.85
	1996	.43	.69	.47	4.52	.39	7.50	.79	.77
15	1990	.75	.97	.82	9.24	.58	14.15	1.41	1.05
	1992	.63	1.03	.62	8.80	.48	13.19	1.07	1.11
	1994	.64	.89	.63	8.59	.49	12.97	1.07	1.02
	1996	.62	.88	.68	6.36	.50	9.84	1.17	.96

[a] Multiple imputation procedure based on $M = 10$ imputations.

Results: Simulation Study

Bridging When Data Are MCAR

Mean estimate RMSE. A few broad patterns were apparent among the RMSE values for the bridged means displayed in Table 4. For a given bridging method, the RMSEs increased as the percent biracial increased, and the RMSEs were frequently larger for the Black subgroup. A more specific look at the results reveals that the effectiveness of particular bridging methods varied depending on the percent multiracial, bridging procedure, and subgroup for which the estimates were obtained.

The LR procedure consistently resulted in the most accurate estimates for the Black subgroup. None of the Black subgroup RMSEs exceeded 1.03 for this method. Bridged estimates for the DLG method were also fairly close to the true estimates and commonly had the second smallest RMSEs for the Black subgroup. By comparison, the Black subgroup means were poorly estimated by the PWA and DSM methods. For example, in the 15% biracial condition with the 1990 data, the RMSE for the LR condition was only .97 but it was 9.24 for PWA and 14.15 for DSM. This amount of bias was notable given that the 1990 Black subgroup standard deviation was 35.67 (Table 3). The results of the DSM method were not unexpected given that

TABLE 5
Bias for Bridged Means, MCAR Condition

Percent Biracial		Logistic Regression[a]		PWA[a]		DSM		DLG	
	Year	White	Black	White	Black	White	Black	White	Black
2	1990	−.097	.012	−.096	1.461	.004	2.776	−.194	.005
	1992	−.081	−.040	−.075	1.450	−.010	2.751	−.141	−.001
	1994	−.082	−.003	−.068	1.361	.004	2.596	−.143	.004
	1996	−.074	.008	−.080	.983	−.001	1.881	−.159	.013
5	1990	−.233	.024	−.248	3.487	.005	6.217	−.479	.024
	1992	−.199	−.081	−.179	3.414	.002	6.052	−.349	.000
	1994	−.211	.011	−.191	3.333	−.007	5.877	−.360	.031
	1996	−.189	.009	−.202	2.371	.008	4.252	−.396	.020
8	1990	−.376	.026	−.394	5.303	−.007	9.053	−.751	−.011
	1992	−.321	−.145	−.293	5.201	.008	8.700	−.568	.027
	1994	−.321	−.030	−.298	5.018	−.004	8.414	−.561	.006
	1996	−.306	.035	−.331	3.650	−.007	6.247	−.624	.030
10	1990	−.453	−.012	−.505	6.524	−.003	1.711	−.936	−.015
	1992	−.387	−.195	−.367	6.244	.000	1.109	−.691	.018
	1994	−.408	−.028	−.357	6.047	.022	9.834	−.701	−.003
	1996	−.373	.059	−.403	4.436	−.003	7.402	−.756	−.000
15	1990	−.698	.077	−.754	9.146	−.008	14.040	−1.366	−.026
	1992	−.575	−.337	−.551	8.694	.025	13.073	−1.025	−.022
	1994	−.594	−.013	−.568	8.503	−.014	12.859	−1.031	.016
	1996	−.562	.027	−.617	6.268	.011	9.735	−1.131	−.015

[a] Multiple imputation procedure based on $M = 10$ imputations.

it assigns all biracial examinees to the smallest (Black) subgroup. Consequently, the Black subgroup mean was increased toward the White subgroup mean, which was at least 28 points higher in the population data sets.

For the White subgroup, the LR RMSEs were the smallest in the 2% and 5% conditions. With larger percentages of biracial examinees, the DSM bridged means were smallest for the White subgroup. However, DSM performed only slightly better than LR as evidenced by the slightly smaller RMSEs. Bridging with the DLG method frequently resulted in the largest RMSEs for the White subgroup. The latter finding was also not surprising given that all biracial examinees are placed into the White subgroup. The result was that the White mean decreased toward the Black subgroup mean. This finding reflects the role of bias in the RMSE.

Mean estimate bias. Bias was typically negative for the White subgroup and positive for the Black subgroup (Table 5). This pattern was due to the White mean being larger in the population by 28 points or more. Classifying an examinee into the wrong subgroup caused the subgroup mean to be shifted toward the mean of the examinee's true subgroup. Regarding the specific bridging methods, DSM resulted in a substantial amount of bias for the Black subgroup mean estimates; it was as large as 14 scale

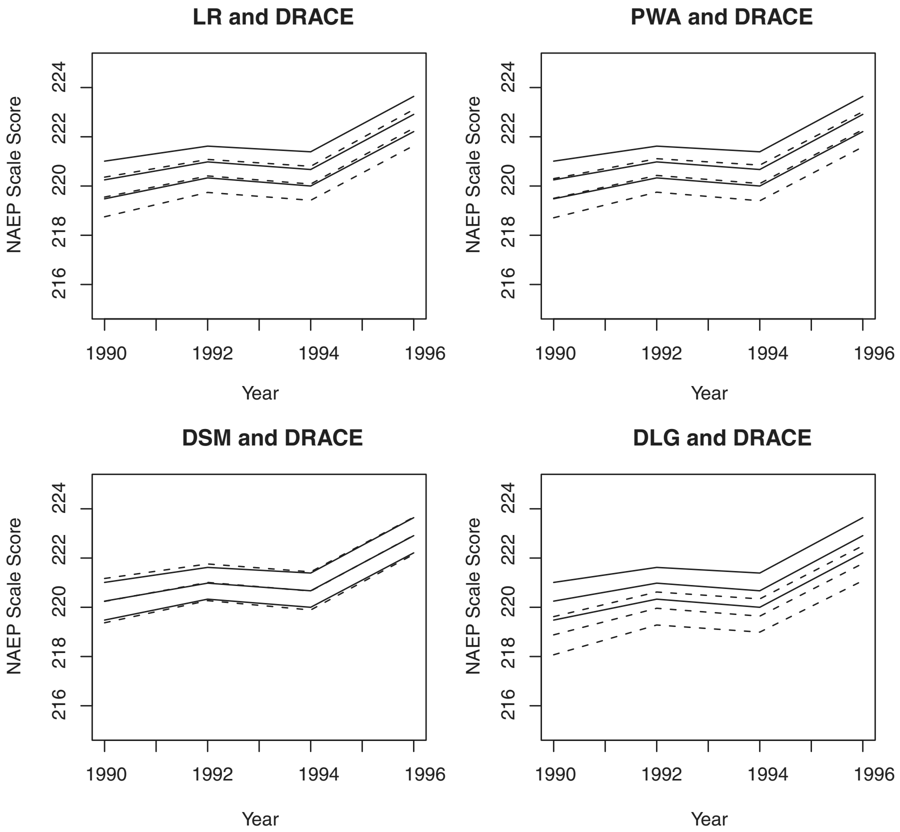

FIGURE 1. *Distribution of White subgroup means over the 1,000 replications: 15% biracial condition, MCAR. The solid line corresponds to DRACE. The dash line pertains to a bridging method. For each line type, the upper line is the third quartile, the middle line is the mean, and the lower line is the first quartile.*

score points (about .4 of a standard deviation) in one condition. LR was the only method in which bias remained low for both subgroups. Bias for the MCAR, 15% multiracial condition is displayed graphically in Figure 1 for the White subgroup and Figure 2 for the Black subgroup. These figures show that a particular bridging method may work well for one subgroup but poorly for the other. In addition, Figures 1 and 2 show that the pattern of means across years and within a race does not appear to be affected much by bridging.

Coverage. Coverage for LR remains close to the nominal level of .95 in all conditions (Table 6). This result stands in sharp contrast to the coverage for the other bridging methods. Coverage was no better than 92% for DSM estimates for the Black subgroup, and, in many cases, it was effectively zero. Coverage for the PWA Black subgroup estimates was also low, with values as small as 22.5%. Of the noncovariate

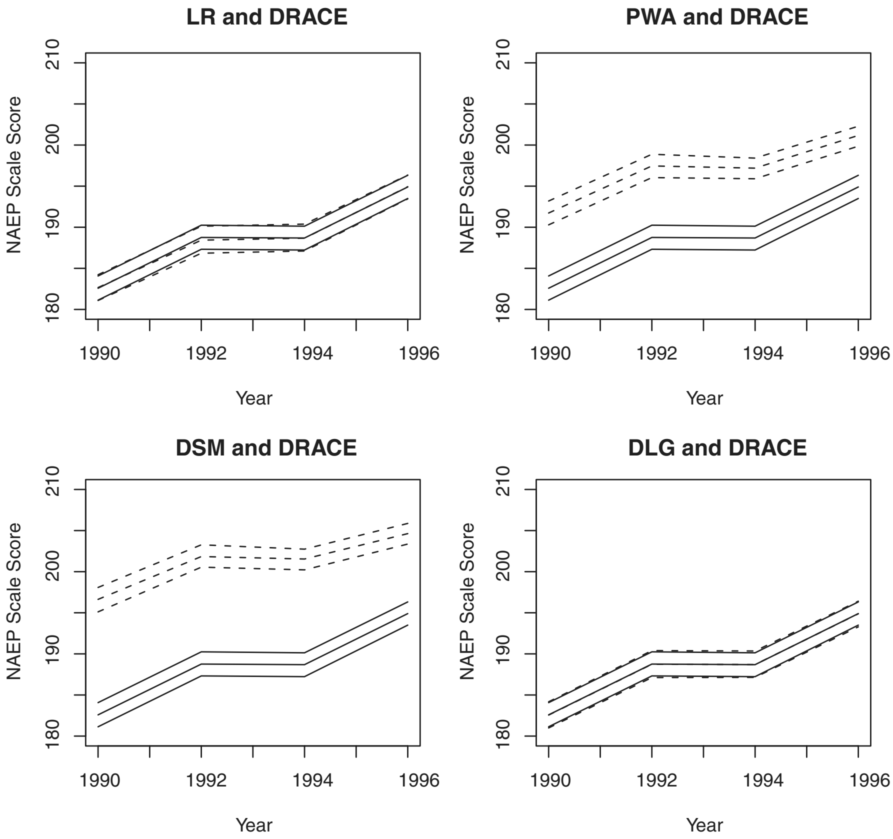

FIGURE 2. *Distribution of Black subgroup means over the 1,000 replications: 15% biracial condition, MCAR. The solid line corresponds to DRACE. The dash line pertains to a bridging method. For each line type, the upper line is the third quartile, the middle line is the mean, and the lower line is the first quartile.*

bridging methods, DLG was the best in terms of coverage. However, it still had values that substantially deviated from the nominal level.

Trend coefficient estimates. Table 7 shows the RMSEs for the trend coefficients and their standard errors. Overall these RMSEs are noticeably smaller than the RMSEs for the mean estimates. Figures 1 and 2 may explain the reason for this finding. The pattern of bridged mean estimates across years does not deviate from the pattern of true mean estimates. Only the intercept of the trend changes and this coefficient was not evaluated. In terms of the other coefficients, LR consistently resulted in the lowest RMSE for linear and quadratic trend coefficients, although on a few occasions other methods had similar RMSEs. Conversely, DSM consistently had the largest RMSEs regardless of type of trend coefficient or subgroup. There was only one instance where DSM did not have the largest RMSE.

TABLE 6

95% Confidence Interval Coverage for Bridging Methods, MCAR Condition

Percent		Logistic Regression[a]		PWA[a]		DSM		DLG	
Multiracial	Year	White	Black	White	Black	White	Black	White	Black
2	1990	.968	.980	.968	.976	.967	.857	.971	.977
	1992	.964	.968	.968	.956	.970	.809	.961	.967
	1994	.969	.972	.973	.974	.974	.852	.971	.973
	1996	.982	.983	.980	.981	.979	.924	.977	.982
5	1990	.957	.976	.963	.907	.962	.264	.942	.974
	1992	.972	.964	.980	.886	.972	.280	.964	.962
	1994	.967	.978	.974	.900	.973	.309	.962	.972
	1996	.975	.976	.977	.936	.974	.512	.971	.970
8	1990	.958	.973	.969	.759	.971	.021	.930	.967
	1992	.959	.969	.968	.748	.969	.028	.949	.967
	1994	.970	.979	.975	.757	.977	.033	.954	.976
	1996	.970	.975	.971	.834	.972	.133	.953	.975
10	1990	.969	.973	.976	.628	.976	.008	.923	.964
	1992	.965	.978	.973	.583	.970	.003	.938	.974
	1994	.969	.976	.978	.631	.970	.008	.948	.974
	1996	.967	.982	.985	.746	.981	.025	.924	.976
15	1990	.942	.969	.962	.257	.980	.000	.857	.959
	1992	.950	.972	.972	.225	.971	.000	.884	.965
	1994	.962	.972	.970	.302	.966	.000	.909	.971
	1996	.960	.980	.978	.444	.979	.000	.869	.970

[a] Multiple imputation procedure based on $M = 10$ imputations.

Bridging for the MNAR-W Condition

Mean estimate RMSE. In the MNAR-W condition, LR frequently resulted in bridged estimates that were closer to the true estimates (Table 8) than when data were MCAR (Table 4). Moreover, LR performed notably better than the other two methods regardless of the percentage of examinees assigned to the biracial group. Interestingly, PWA had the second smallest RMSEs across conditions, unlike the findings for PWA when data were MCAR. Taken together, these results suggest that multiple imputation methods work well in the MNAR-W condition, at least as compared to a deterministic single imputation method.

Mean estimate bias. Table 9 shows that the pattern of negative bias for the White subgroup and positive bias for the Black subgroup does not hold in the MNAR-W condition. Rather, it appears that there is a predominantly positive bias, particularly for the LR method. The reason for this effect may be due to the relationship between test scores, background items, and single-race group membership, even though test scores are not explicitly part of the bridging method. Table 3 indicates that there is a positive relationship between test scores and the White subgroup. This relationship is likely affecting the bridging. When the LR method indicates that an examinee has a high probability for the White subgroup, there is a good chance that the examinee

TABLE 7

RMSE for Bridged Trend Coefficients, MCAR, MNAR-W, and MNAR-B Conditions

			MCAR				MNAR-W			MNAR-B		
Condition	Group	Coefficient	LR	PWA	DSM	DLG	LR[a]	PWA[a]	DSM	LR[a]	PWA[a]	DSM
2%	White	Linear	.02	.02	.04	.03	.01	.02	.05	.05	.03	.07
		Quadratic	.01	.01	.02	.01	.01	.01	.02	.03	.02	.04
	Black	Linear	.07	.13	.24	.07	.04	.15	.27	.15	.09	.18
		Quadratic	.04	.07	.12	.04	.02	.08	.13	.08	.05	.10
5%	White	Linear	.04	.04	.07	.04	.02	.04	.07	.08	.06	.10
		Quadratic	.02	.02	.03	.03	.01	.02	.04	.04	.03	.06
	Black	Linear	.11	.25	.42	.12	.07	.27	.47	.25	.14	.31
		Quadratic	.07	.12	.18	.07	.04	.12	.20	.15	.08	.18
8%	White	Linear	.05	.05	.08	.06	.03	.04	.09	.12	.08	.14
		Quadratic	.02	.03	.04	.03	.01	.02	.05	.06	.05	.09
	Black	Linear	.15	.34	.54	.15	.11	.37	.61	.36	.18	.44
		Quadratic	.08	.15	.21	.09	.06	.15	.22	.21	.10	.25
10%	White	Linear	.05	.06	.09	.06	.03	.05	.10	.14	.09	.17
		Quadratic	.03	.03	.05	.04	.01	.03	.06	.07	.06	.10
	Black	Linear	.17	.40	.61	.18	.11	.43	.69	.47	.20	.56
		Quadratic	.09	.16	.22	.10	.06	.17	.25	.27	.11	.32
15%	White	Linear	.06	.07	.12	.08	.04	.06	.13	.20	.12	.21
		Quadratic	.03	.04	.06	.05	.02	.04	.07	.12	.08	.13
	Black	Linear	.20	.52	.75	.22	.15	.56	.83	.87	.22	.96
		Quadratic	.12	.19	.25	.13	.09	.19	.26	.49	.12	.52
Sudden	White	Linear	.07	.07	.07	.11	.03	.04	.08	.04	.26	.48
		Quadratic	.02	.02	.04	.03	.01	.02	.04	.07	.05	.08
	Black	Linear	.12	.65	1.01	.13	.10	.72	1.14	.40	.15	.38
		Quadratic	.06	.11	.16	.07	.04	.11	.18	.18	.08	.20

Note. DLG RMSE for MNAR-W not reported because it perfectly matches the missing data mechanism. DSM RMSE for MNAR-B not reported because it perfectly matches the missing data mechanism.

[a] Multiple imputation procedure based on $M = 10$ imputation.

also has a high test score. As a result the bridged means for the White and Black subgroups are increased. Stated differently, only the low scoring White subgroup members (but high scoring relative to the Black subgroup) are being placed into the Black subgroup so the bridged White and Black means are higher than the true values.

The two noncovariate bridging methods (PWA and DSM) typically had the smallest amount of bias for the White subgroup. However, this positive result was overshadowed by the large amount of bias these two methods produced for the Black subgroup. In the 1990 data, 15% multiracial condition, bias was 10 scale score points for PWA and 16 points for DSM. Only LR produced low bias for both subgroups. Figures 3 and 4 illustrate this result. Bias was rather small for the White subgroup (smaller than the White subgroup bias in the MCAR condition). However, bias was quite prominent for the Black subgroup. Only LR had a small amount of bias for both subgroups.

Coverage. As in the MCAR condition, LR was the only method that had coverage values that were near the nominal level for all conditions (Table 10). PWA and DSM

TABLE 8

RMSE for Bridged Means, MNAR-W, and MNAR-B Conditions

		MNAR-W						MNAR-B					
		LR[a]		PWA[a]		DSM		LR[a]		PWA[a]		DLG	
Percent Biracial	Year	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
2	1990	.06	.23	.11	1.87	.22	3.55	.63	.75	.50	.40	.98	.82
	1992	.05	.14	.09	1.85	.18	3.52	.54	.82	.37	.45	.74	.90
	1994	.04	.16	.10	1.78	.19	3.35	.51	.72	.37	.43	.73	.86
	1996	.05	.19	.10	1.27	.19	2.42	.53	.67	.41	.37	.80	.76
5	1990	.12	.43	.17	4.23	.34	7.60	1.54	1.47	1.17	.66	2.27	1.44
	1992	.09	.23	.15	4.15	.30	7.36	1.32	1.51	.89	.67	1.73	1.48
	1994	.08	.28	.15	3.94	.29	7.03	1.25	1.32	.89	.65	1.73	1.46
	1996	.10	.36	.15	2.86	.29	5.16	1.32	1.31	.97	.59	1.88	1.34
8	1990	.16	.73	.21	6.34	.43	1.85	2.57	2.26	1.85	.81	3.51	1.96
	1992	.13	.31	.18	6.07	.37	1.24	2.14	2.36	1.36	.86	2.62	2.15
	1994	.11	.39	.20	5.90	.40	9.96	2.01	2.00	1.38	.85	2.64	2.10
	1996	.14	.53	.19	4.32	.39	7.51	2.15	2.01	1.51	.74	2.86	1.88
10	1990	.20	.78	.25	7.55	.51	12.59	3.32	2.90	2.28	.93	4.30	2.62
	1992	.16	.41	.21	7.30	.43	11.94	2.75	2.86	1.69	.92	3.22	2.53
	1994	.14	.47	.21	7.10	.42	11.62	2.57	2.50	1.70	.91	3.23	2.57
	1996	.16	.66	.21	5.16	.43	8.70	2.77	2.54	1.86	.82	3.49	2.27
15	1990	.29	1.23	.30	1.26	.64	16.14	5.43	4.76	3.30	.98	6.07	4.22
	1992	.24	.60	.26	9.85	.55	15.12	4.33	4.81	2.47	1.09	4.57	4.55
	1994	.21	.72	.26	9.47	.55	14.60	4.12	4.57	2.48	1.02	4.59	4.33
	1996	.25	1.01	.26	7.09	.56	11.34	4.51	4.54	2.69	.93	4.91	3.99

Note. RMSE for DLG not reported for MNAR-W because it perfectly matches the missing data mechanism. RMSE for DSM not reported for MNAR-B because it perfectly matches the missing data mechanism. RMSEs were 0.

[a] Multiple imputation procedure based on $M = 10$ imputations.

TABLE 9
Bias for Bridged Means, MNAR-W and MNAR-B Conditions

		MNAR-W						MNAR-B					
		LR[a]		PWA[a]		DSM		LR[a]		PWA[a]		DSM	
Percent Biracial	Year	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
2	1990	.023	.109	.003	1.776	.003	3.385	−.592	−.340	−.473	−.000	−.938	.009
	1992	.023	.028	−.004	1.762	−.008	3.367	−.498	−.317	−.351	−.010	−.693	−.018
	1994	.018	.040	.002	1.683	.004	3.191	−.472	−.218	−.351	.004	−.694	.008
	1996	.020	.074	.001	1.194	.001	2.292	−.485	−.273	−.387	.005	−.758	−.007
5	1990	.071	.245	.001	4.137	−.001	7.463	−1.507	−.938	−1.144	−.026	−2.226	−.026
	1992	.059	.075	.003	4.049	.004	7.207	−1.280	−.843	−.866	.007	−1.686	.011
	1994	.043	.123	.003	3.847	.002	6.887	−1.215	−.551	−.874	.037	−1.700	.079
	1996	.052	.219	.001	2.782	.004	5.049	−1.279	−.718	−.950	.028	−1.839	.048
8	1990	.110	.456	−.003	6.252	−.004	1.730	−2.533	−1.588	−1.826	.019	−3.476	.010
	1992	.101	.132	.010	5.988	.015	1.117	−2.101	−1.420	−1.344	−.037	−2.585	−.089
	1994	.077	.217	.005	5.813	.006	9.831	−1.970	−1.012	−1.361	.023	−2.605	.021
	1996	.090	.367	.003	4.246	.000	7.398	−2.107	−1.293	−1.484	−.001	−2.820	.008
10	1990	.143	.582	−.005	7.461	−.009	12.466	−3.280	−2.000	−2.261	.027	−4.264	.050
	1992	.129	.219	−.002	7.211	−.006	11.822	−2.713	−1.687	−1.672	−.025	−3.186	−.015
	1994	.103	.286	−.004	7.014	−.008	11.506	−2.532	−1.343	−1.684	−.006	−3.199	−.008
	1996	.113	.496	−.000	5.091	.006	8.602	−2.736	−1.638	−1.838	.036	−3.454	.126
15	1990	.233	1.021	.005	1.181	.009	16.037	−5.401	−3.038	−3.283	.018	−6.043	−.057
	1992	.210	.376	.004	9.766	.007	15.015	−4.293	−1.769	−2.448	−.007	−4.544	−.138
	1994	.169	.521	.000	9.387	.001	14.491	−4.082	−2.573	−2.460	.011	−4.560	−.048
	1996	.190	.825	.010	7.018	.016	11.241	−4.480	−2.530	−2.674	.001	−4.884	−.015

Note. DLG bias in MNAR-W not reported because it perfectly matches the missing data mechanism. DLG bias for MNAR-B not reported because it perfectly matches the missing data mechanism.

[a] Multiple imputation procedure based on $M = 10$ imputations.

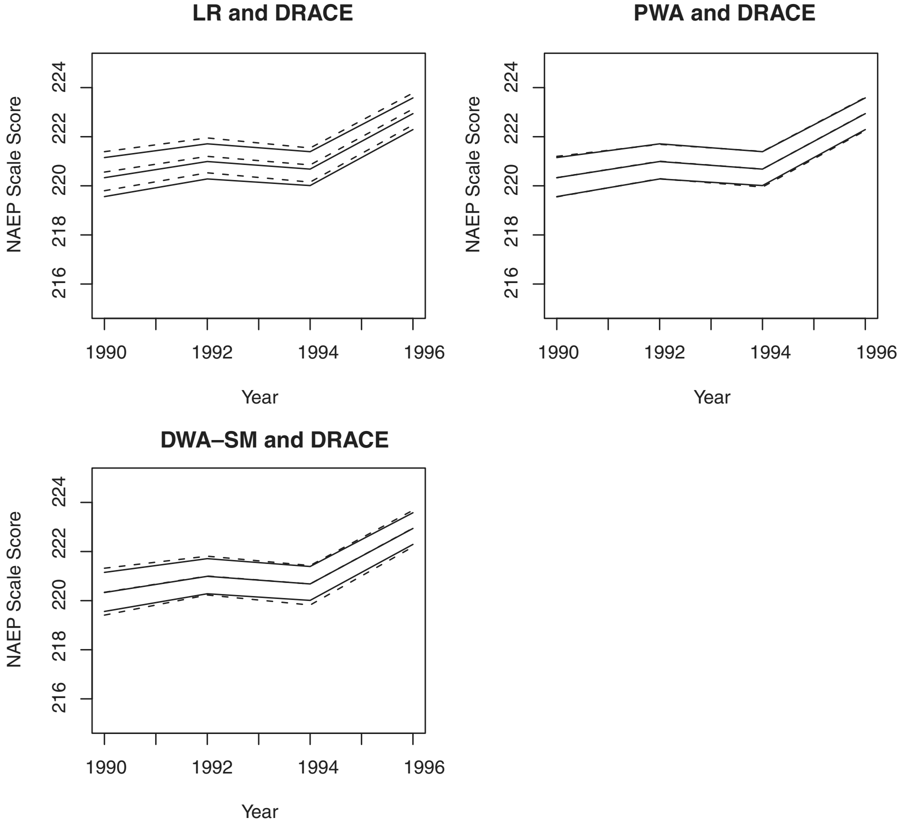

FIGURE 3. *Distribution of White subgroup means over the 1,000 replications: 15% biracial condition, MNAR-W. The solid line corresponds to DRACE. The dash line pertains to a bridging method. For each line type, the upper line is the third quartile, the middle line is the mean, and the lower line is the first quartile.*

had good coverage for the White subgroup, but rather poor coverage for the Black subgroup. Once the percent multiracial was larger than 5%, the coverage for PWA and DSM departed substantially from the nominal level. Coverage for DSM, in particular, was no better than .86, and it was close to zero in many cases.

Trend coefficient estimates. The RMSEs for the trend coefficients and their standard errors for the MNAR-W condition (Table 7) were very similar to, albeit slightly larger than, those for the MCAR condition. Moreover, Figures 3 and 4 reinforce the notion that bridging does not affect the pattern of means across years. Rather, the intercept is shifted and may be shifted severely depending on the bridging method used. These results and their similarity to the trend coefficient results for the MCAR condition suggest that the missing data mechanism does not affect estimates of the trend coefficients. However, this finding may simply be due to the simplicity of the simulation. If missingness were related to the test scores in a way that varied

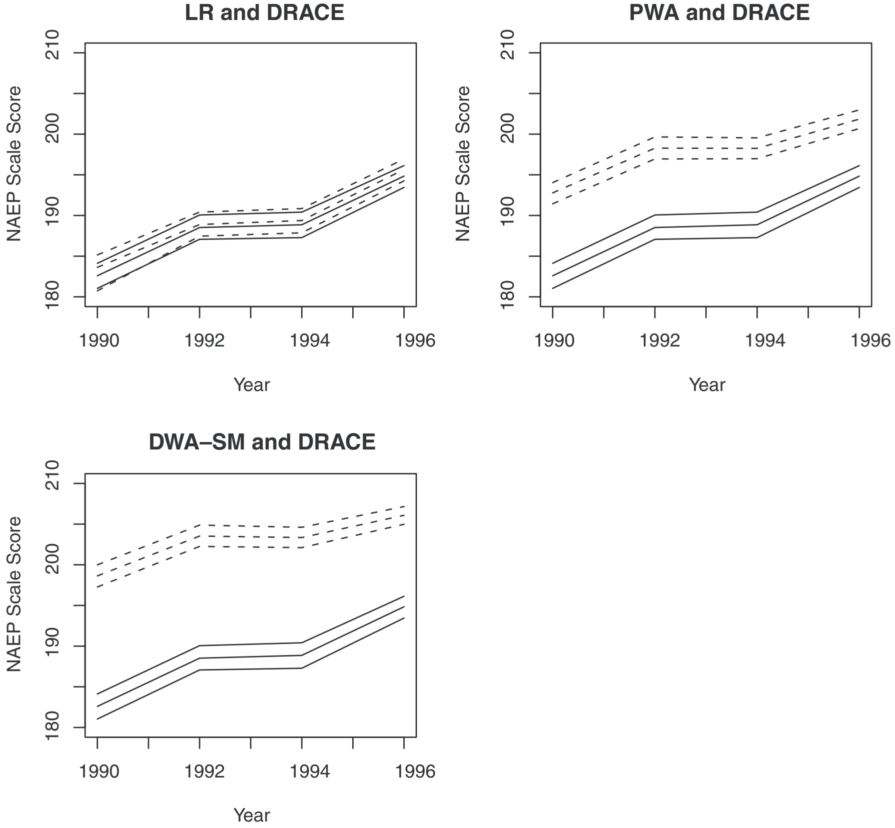

FIGURE 4. *Distribution of Black subgroup means over the 1,000 replications: 15% biracial condition, MNAR-W. The solid line corresponds to DRACE. The dash line pertains to a bridging method. For each line type, the upper line is the third quartile, the middle line is the mean, and the lower line is the first quartile.*

across years, the trend estimates would likely be more (and unpredictably) affected by bridging.

Bridging for the MNAR-B Condition

Mean estimate RMSE. In the MNAR-B condition, RMSE was larger for all methods as compared to the MNAR-W condition (Table 8). RMSE was often largest for the DLG method and White subgroup. It was typically the second largest for the LR method and White subgroup. The opposite was true for the Black subgroup; the LR method commonly had the largest RMSE and the DLG method often had the second largest. The smallest RMSEs were produced by the PWA method, regardless of subgroup.

TABLE 10

95% Confidence Interval Coverage for Bridging Methods, MNAR-W and MNAR-B Conditions

		MNAR-W						MNAR-B					
		LR[a]		PWA[a]		DSM		LR[a]		PWA[a]		DLG	
Percent Biracial	Year	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
2	1990	.965	.981	.965	.959	.960	.762	.946	.974	.957	.984	.928	.970
	1992	.976	.980	.979	.947	.975	.746	.936	.973	.948	.985	.924	.972
	1994	.977	.981	.976	.958	.976	.762	.963	.979	.975	.993	.938	.976
	1996	.974	.990	.974	.977	.973	.855	.963	.981	.975	.989	.935	.977
5	1990	.974	.969	.980	.837	.974	.116	.805	.970	.894	.995	.613	.977
	1992	.970	.964	.976	.823	.972	.120	.826	.957	.928	.990	.713	.965
	1994	.982	.977	.984	.841	.981	.135	.864	.964	.935	.986	.723	.964
	1996	.986	.969	.988	.870	.981	.313	.836	.973	.919	.992	.677	.972
8	1990	.979	.975	.986	.586	.976	.002	.535	.951	.806	.998	.222	.956
	1992	.970	.970	.980	.556	.977	.002	.540	.948	.831	.996	.348	.953
	1994	.974	.975	.981	.607	.977	.010	.630	.965	.836	.998	.367	.968
	1996	.981	.979	.982	.695	.977	.037	.578	.959	.835	.998	.268	.963
10	1990	.963	.967	.972	.401	.966	.000	.312	.934	.661	.997	.072	.956
	1992	.959	.975	.966	.376	.957	.000	.331	.955	.737	.998	.153	.963
	1994	.972	.985	.986	.400	.978	.000	.408	.970	.771	.998	.178	.960
	1996	.977	.969	.983	.561	.971	.006	.352	.955	.725	.998	.104	.961
15	1990	.964	.968	.979	.081	.955	.000	.025	.962	.367	1.00	.001	.964
	1992	.963	.979	.979	.076	.962	.000	.024	.976	.485	1.00	.004	.949
	1994	.969	.984	.982	.076	.974	.000	.044	.974	.469	1.00	.012	.958
	1996	.976	.968	.990	.197	.982	.000	.016	.968	.414	1.00	.001	.950

Note. DLG coverage for MNAR-W not reported because it perfectly matches the missing data mechanism. DSM coverage for MNAR-B not reported because it perfectly matches the missing data mechanism.

[a] Multiple imputation procedure based on $M = 10$ imputations.

Mean estimate bias. The pattern of bias was similar to the pattern of RMSEs for the bridging methods. However, Table 9 shows that there was a predominant negative bias for all bridging methods. The reason for this negative bias may be the relationship between test scores, background items, and single-race group membership that was described in the MNAR-W condition. Black examinees misclassified into the White subgroup are likely the higher scoring Black examinees but still lower scoring than most White examinees. As a result, both the White and Black subgroup means were lowered. The negative bias was most obvious for the LR method. For the other two methods, the Black subgroup mean sometimes had a negative bias and sometimes a positive one. Given that the PWA method was a random imputation method, it was not affected by the relationship between test scores, background items, and single-race subgroup membership, which may explain its superior performance in this condition.

Coverage. In the MNAR-B condition, coverage for the White subgroup was low for all bridging methods (Table 10). For the Black subgroup, coverage was near the nominal level for LR and DLG. Coverage tended to be too high for the Black subgroup and PWA bridging method.

Trend coefficient estimates. The RMSEs for the trend coefficients were again small, and smallest for the PWA method, but they were larger than the trend coefficient RMSEs found in the MCAR and MNAR-W conditions (Table 7). The implications are again that the pattern of means does not change much, but the intercept does.

Discussion

Considering the results of this study and the findings from previous research (Meyer & Huynh, 2008; Schenker & Parker, 2003), covariate bridging methods frequently perform better than the noncovariate methods proposed by the OMB (2000). However, the success of covariate bridging rests on the selection of covariates or background items and the type of missing data mechanism. The first study demonstrated that a particular background item may be a good predictor one year, but a less useful predictor the next. Consequently, background items should be selected independently for each year when data from multiple years are to be analyzed. Moreover, the choice of background items should be specific to each multiple-race subgroup (Schenker & Parker). Although the background items selected in this study may not have been the best, they did often lead to better bridging when compared to the noncovariate bridging methods.

The multiple imputation logistic regression bridging method frequently produced the most accurate mean estimates and best coverage for both race subgroups. The percent multiracial did not appear to have much of an effect on the LR method. The type of missing data mechanism did not seem to have an adverse effect on this method when the mechanism was MCAR or MNAR to a moderate extent. When MNAR was severe, the LR method performed rather poorly. The likely reason was the relationship between test scores, background items, and single-race group, even though only the latter two variables were part of the bridging method. Indeed, if the two single-race groups do not differ on the test scores, all bridging methods will probably perform similarly. Bias really only becomes a problem when groups substantially differ on the test scores.

Note that multiple imputation alone was not sufficient for good bridging. The PWA method represented a multiple imputation procedure with a random imputation model and it frequently did not perform as well as the logistic regression method. The remaining two noncovariate methods tended to perform reasonably well for one subgroup, but quite poorly for the other. Given that a bridging method must be applied to both single-race subgroups, it must be judged by its worst subgroup performance. As such, the two deterministic whole assignment methods are not recommended.

When bridging was applied to linear and quadratic tests of trend, no alarming results were found. The pattern of means across years remained stable. Bridging primarily seemed to affect the trend intercept. The bias was small for the multiple imputation logistic regression methods in the MCAR and MNAR-W conditions, but it was rather large for the noncovariate methods, particularly when the percentage of multiracial examinees was large.

The type of missing data mechanism must be carefully considered when bridging in practice. If evidence suggests the data are MCAR, MAR, or slightly MNAR then the LR method is recommended. Otherwise, the PWA is preferred. In order to gather evidence to make this decision, two questions could be asked of examinees. One question would allow them to select all race subgroups and another question could ask “please select the one race subgroup that best describes you.” Cross tabulating these two questions would indicate whether all biracial examinees select the same single-race subgroup (in which case a condition like MNAR-W or MNAR-B would apply) or whether responses would be evenly distributed over the single-race subgroups.

Bridging is not the only method available for analyzing data with people classified into multiple race categories. An alternative method places all multiracial examinees into a “Two or More Races” group. A problem with this method is that any changes to the single-race subgroup test performance (i.e., subgroup mean proficiency) may be due to true changes in the single-race group proficiency or due to removing multiracial examinees from the single-race groups. In this case, bridging may be useful for actual reporting or in research that investigates the cause of changes in single-race subgroup proficiency.

Implications for NCLB and Other Federally Mandated Reporting

A recent *Federal Register* Notice (Proposed Guidance on Maintaining, Collecting, and Reporting Data on Race and Ethnicity to the U.S. Department of Education, 2006) indicated that all K-12 and higher education institutions will be required to use the multiple-race reporting format by 2009. As mentioned in that notice, the new race reporting format will impact school report cards and adequate yearly progress (AYP) reports published pursuant to the No Child Left Behind Act of 2001 (2002). It will also impact Integrated Postsecondary Education Data systems (IPEDS) and Rehabilitation Services Administration (RSA) reporting. The notice recommends a two question format when collecting data on race and ethnicity. One question will ask whether or not a person is Hispanic and another will ask the person to select all (non-Hispanic) race categories that apply. Although certain single- and multiple-race reporting categories are required, the notice indicates that bridging methods are permitted, if there is evidence that the new reporting format has impacted the reported statistics. According to the notice, if bridging is used, an educational institution must document the bridging method and apply it consistently. The institution may not switch bridging methods year to year or even use a different method for each grade. Of concern is that the notice recommends the OMB (2000) (noncovariate) bridging methods. This study suggested that the amount of bias possible by using one of the noncovariate bridging methods could have a detrimental effect on a school's AYP for a race subgroup or an institution's report of subgroup graduation rates, particularly in the first year of bridging. An institution should carefully study the implementation of the new multiple-race format as well as the use of any bridging method. For example, an institution might consider collecting data under both the single- and multiple-race format for a few transition years. This additional data collection would allow an institution to evaluate changes due to the new race format as well as whether or not bridging would be useful and effective.

Limitations of the Study

One weakness that is not evident in the paper concerns the use of logistic regression for bridging. The original design of the study included White/Asian and White/American Indian Alaskan Native multiracial groups. These groups were dropped from the study because the logistic regression procedure frequently failed to converge because of quasi-complete separation. Details of complete and quasi complete separation may be found in the SAS User's Guide (SAS Institute, 2006). When the problem of quasi-complete separation in this study was investigated, the likely cause was extremely small sample sizes for some race groups. For example, in some simulated data sets, the American Indian/Alaskan Native subgroup included only six examinees. Even though the multiple imputation LR method worked well for the two larger groups studied in the simulation, caution is recommended when using it with small race subgroups. Moreover, SAS does not provide a warning for quasi-complete separation when sampling weights are included in an analysis (SAS Institute). Of course, the problem may also be avoided by wisely selecting background items.

Another limitation of the study concerns the manner in which missingness was simulated. The simulation did not explicitly account for any relationship between missingness and NAEP scale scores. Moreover, the interaction between this relationship and the NAEP assessment year was not manipulated. The possibility of such an interaction should be considered when bridging in practice.

Finally, multiple imputation has the advantage of allowing the imputed race subgroup values to be included in public-use data sets, which facilitates the secondary use of NAEP data. However, applying multiple-race subgroup imputations to the multiple NAEP plausible values results in a substantial increase in the number of estimates that must be pooled in an analysis. For 10 imputations of race subgroup and five plausible values, there are 50 analyses that must be summarized. The logical way to reduce the amount of work would be to use fewer, say five, race subgroup imputations. Even if only a single imputation were used, covariate bridging methods work better than noncovariate methods (Meyer & Huynh, 2008). However, the standard error for single imputation bridging would be a bit optimistic and not reflect error due to imputation.

Acknowledgments

The authors thank Christine E. DeMars, Huynh Huynh, and three anonymous reviewers for their comments and recommendations. We also thank the NAEP Secondary analysis grant program for supporting this research. This research was funded by Institute of Education Sciences NAEP Secondary Analysis grant number R902B050026 awarded to J. Patrick Meyer. Points of view or opinions stated herein do not necessarily reflect the position or policy of the Institute of Education Sciences and no official endorsement should be inferred.

Notes

[1] The terms White and Black are used herein to be consistent with NAEP reporting categories.

[2] DRACE is based on self-reported race when this information is provided by the student. It is based on the student's observed race, as reported by a school official, when it is not reported by the student.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 Technical Report*, NCES 1999–452. Washington, D C: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Campbell, J. R., Hombo, C. M., & Mazzeo, J. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance*, NCES 2000–469. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Cohen, J. (1988). *Statistical power analysis for the social sciences*. New York: Lawrence Erlbaum Associates.

Jones, N. A., & Smith, A. S. (2001, November). *The two or more races population 2000: Census 2000 brief*. Retrieved November 5, 2002, from http://www.census.gov/prod/2001pubs/c2kbr01–6.pdf

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Lopez, A. M. (2003). Mixed-race school-age children: A summary of Census 2000 Data. *Educational Researcher*, *32*, 25–37.

Meyer, J. P., & Hombo, C. M. (2003, April). *The impact of multiracial response data on NAEP scores*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Meyer, J. P., & Huynh, H. (2008). On the use of covariates in bridging single-race and multiple-race categories. *Journal of Experimental Education*, *77*, 69–94.

Meyer, J. P., McClellan, C. M., & Huynh, H. (2004, April). *Background information methods for the analysis of mixed groups with application to multi-racial NAEP data*. Paper presented at the meeting of the American Educational Research Association, San Diego.

Office of Management and Budget (2000, December). *Provisional guidance on the implementation of the 1997 Standards for Federal Data on Race and Ethnicity*. Retrieved

November 5, 2002, from http://www.whitehouse.gov/omb/inforeg/r&e˙guidance2000update.pdf

Proposed Guidance on Maintaining, Collecting, and Reporting Data on Race and Ethnicity to the U.S. Department of Education, 71 Fed. Reg. 44,865, fr07au06–150 (Aug. 7, 2006).

Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity, 62 Fed. Reg. 58,781, fr30oc97–141 (Oct. 30, 1997).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: John Wiley & Sons.

SAS Institute (2006). *Documentation for SAS products and solutions*. Retrieved November 16, 2006 from http://support.sas.com/documentation/onlinedoc/

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schenker, N., & Parker, J. D. (2003). From single-race reporting to multiple-race reporting: Using imputation methods to bridge the transition. *Statistics in Medicine*, *22*, 1571–1587.

The No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2002).

Authors

J. PATRICK MEYER is Assistant Professor, University of Virginia, Curry School of Education, 405 Emmet Street South, P.O. Box 400277, Charlottesville, VA 22903; meyerjp@virginia.edu. His primary research interests are educational measurement and statistics.

J. CARL SETZER is a psychometrician, GED® Testing Service, One Dupont Circle NW, Washington, DC 20036; carl_setzer@ace.nche.edu. His primary research interests include psychometrics.