

ONLINE SUPPLEMENT
to article in

AMERICAN SOCIOLOGICAL REVIEW, 2007, VOL. 72 (AUGUST: 487–511)

Diversity, Opportunity, and the Shifting Meritocracy in Higher Education

Sigal Alon
Tel Aviv University

Marta Tienda
Princeton University

SECTION A: MISSING DATA

We improve on prior evidence based on the two national surveys by tackling the issue of missing data using state-of-the-art multiple imputation (MI) techniques. Missing data increases the risk of reaching incorrect conclusions because it may bias parameter estimates, inflate type I and II error rates, degrade the performance of confidence intervals, and dramatically reduce statistical power (Allison 2001; Collins, Schafer, and Kam 2001). Little and Rubin (1987) have shown that ad hoc procedures such as listwise or pairwise deletion, substitution with constants, regression predictions, or other forms of single imputations perform poorly except under very restrictive conditions. Statisticians widely agree that MI is the best approach for dealing with missing data (e.g., Allison 2001; Little 1993; Little and Schenker 1995; Schafer 1997). MI is based on sound theory and is shown to produce efficient estimates and accurate measures of statistical uncertainty. Allison (2001:2) observes that MI has “statistical properties that are about as good as we can reasonably hope to achieve.” Notably, several NCES studies frequently use MI to fill missing data (see, for example, the NCES [2001] report on “Educational Achievement and Black-White Inequality,” which uses MI to fill missing data in NLS:72, HS&B, and NLSY).

For our analyses, MI offers several additional and important advantages. First, MI fills in test scores for those who took the test but lacked information in their records or those missing class rank information. MI also allows us to include high school graduates who did not pursue higher education and did not take the SAT. If their SAT scores and class rank had not been imputed, we would have had to restrict all analyses to students attending postsecondary institutions, which is a nonrandom segment of the population. However, this would not solve the missing data problem. Thus, a major improvement of our approach over prior studies is the ability to evaluate the postsecondary destination of *all* high school graduates.

The second advantage is the additional information gained in the imputation process for all the other covariates included in the analyses, which increases precision of estimates for the main independent variables. Yet, the most important advantage of using MI is related to the temporal analyses we perform: the ability to merge two complete data sets (for a set of variables) and to transform both SAT scores and class rank into percentile distributions, thereby enabling a direct comparison of the two variables over time.

ONLINE SUPPLEMENT
to article in

AMERICAN SOCIOLOGICAL REVIEW, 2007, VOL. 72 (AUGUST:487–511)

The procedure: MI is a Monte Carlo technique that substitutes missing values by $m > 1$ simulated versions, where m typically ranges from 3 to 10 (Rubin 1987). MI does not produce a unique set of numbers because random variation is deliberately introduced in the imputation process (Allison 2001). MI's random components adjust the standard errors upward to correct the downward bias in standard errors produced by deterministic imputation. Each of the simulated complete data sets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing data uncertainty (Rubin 1987). We use Schafer's MI software NORM (Schafer 1999) for incomplete multivariate data (free for download; see Schafer [1997] for computational routines).

included in the imputation procedure. As a rule, the most restrictive imputation model should consist of all variables included in the analytical models. Yet an inclusive strategy, such as modeling additional variables that are not included in the prospective analyses, is greatly preferred because it increases efficiency and reduces bias (Collins et al. 2001). We did not impute values for the dependent variables (college destination and graduation status) or race/ethnicity, although these covariates were included in the imputation model to derive unbiased estimates (Schafer 1997). Statistical analysis uses STATA's Clarify module (King, Tomz, and Wittenberg 2000; Tomz, Wittenberg, and King 2003) that incorporates Rubin's (1987) rules to produce one set of estimates and standard errors that take into account missing data uncertainty. Empirical estimates are based on five versions of complete data sets.

The imputation models were estimated separately for each data set. Table S1 lists all variables that were

Table S1. Variables Included in the Imputation Model, High School and Beyond and National Education Longitudinal Survey

Category	High School and Beyond	National Education Longitudinal Survey
Individual Characteristics	race/ethnicity female athlete	race/ethnicity female
College Destination	selectivity by <i>Barron's</i> classification type of college (voc/2/4 yr)	selectivity by <i>Barron's</i> classification type of college (voc/2/4 yr) sector for first pse attended
Family Background	family income parental education SES intact family home ownership number of rooms in HH	family income parental education SES family size received grants for college (financial aid) composite measure for educational resources in HH (daily newspaper, magazine, encyclopedia, atlas, dictionary, computer, 50 books, calculator)
High School Characteristics	type of HS (private/public/catholic) region urbanicity	type of HS (private/public/catholic) region urbanicity HS % whites HS % of students in single parent HH
Academic Preparedness and Expectations	class rank (HS transcripts) cumulative HS GPA (HS transcripts) ^a revised SAT/ACT scores senior reading test quintile College 6-yr grad status	class rank (HS transcripts) cumulative HS GPA (HS transcripts) SAT math score SAT verbal score ACT score (composite) college 6-year grad status applied to college number of institutions applied educational expectations (highest level of education expected)

^a As noted by NCES, this variable is not standardized. Some values exceed 100 percent because of quality points awarded for advanced courses.

ONLINE SUPPLEMENT
to article in

AMERICAN SOCIOLOGICAL REVIEW, 2007, VOL. 72 (AUGUST:487-511)

Diagnostic analysis: We compare the estimates produced by the MI data sets to the most popular alternatives for handling missing data: listwise deletion and constant substitution with flags for cases with missing information. Allison (2001) concludes that MI offers substantial improvements over listwise deletion while the other methods are inferior to listwise deletion. We replicated the analysis presented in Table 4 in the text (determinants of college destinations) using the three approaches to missing data. Table S2 reports these results. The top panel shows the results based on multiply-imputed data sets (results identical to Table 4). The middle panel shows the results based on listwise deletion. We excluded from the sample any observations that have missing information on any variables in the model. This dramatic reduction in sample size is consequential for statistical power. The drop in statistical power is especially evident for the NO PSE destination. Yet, the results are strikingly similar to the MI results. Most importantly, using listwise

deletion we would have reached the same conclusions we reached based on the MI results.

The bottom panel reports results for the entire sample while substituting missing data with constant values and including dummies as indicators of missing data. The advantage of this method is that it uses all the information, but unfortunately it generally produces biased estimates of the coefficients (Allison 2001). This seems to be the case with our estimates. Some of the point estimates are different from those produced by either MI or listwise strategies. Still, despite the evident biases in point estimates, our general conclusions are unaltered. In sum, our results and conclusions are not an artifact of the use of MI. We think that the use of MI to fill missing information is one of the strengths of the article and a major improvement over prior research that either ignores the problem altogether or uses inferior methods to handle it.

Table S2. Alternatives for Handling Missing Data: Determinants of College Destinations Multinomial Logistic Odds Ratios

	High School and Beyond 1982				National Education Longitudinal Survey 1992			
	MI				LISTWISE			
	(1) No PSE	(2) Nonselect	(3) Select	(4) More-Select	(5) No PSE	(6) Nonselect	(7) Select	(8) More-Select
Black	.905	.877	1.428**	2.839**	.787**	1.076	1.264	4.081**
Hispanic	.870*	.942	1.146	1.329	.701**	.882	.945	1.548**
Asian	.669**	.851	1.640**	4.025**	.547**	.880	1.174	2.104**
SAT Percent	.992**	1.010**	1.017**	1.048**	.993**	1.014**	1.021**	1.061**
Class Rank Percent	.990**	1.013**	1.017**	1.034**	.984**	1.013**	1.020**	1.028**
N	12,754				12,861			
	LISTWISE				FLAGS			
	(9) No PSE	(10) Nonselect	(11) Select	(12) More-Select	(13) No PSE	(14) Nonselect	(15) Select	(16) More-Select
Black	1.033	.946	1.202	2.895**	.697	.977	1.285	4.137**
Hispanic	.944	.928	1.055	1.542*	.701	.977	1.365*	1.761**
Asian	1.017	.824	1.159	3.574**	.606	.805	1.100	1.798**
SAT	.887**	1.063	1.219**	1.738**	.914*	1.184**	1.306**	1.930**
Class Rank	.998	1.012**	1.014**	1.032**	.988**	1.006*	1.013**	1.025**
N	4,062				4,545			
	(17) No PSE	(18) Nonselect	(19) Select	(20) More-Select	(21) No PSE	(21) Nonselect	(22) Select	(23) More-Select
Black	1.000	.859	1.163	1.657**	.838*	.961	1.054	2.346**
Hispanic	.956	.869	.996	.938	.689**	.840	.885	1.188
Asian	.764	.790	1.530**	3.514**	.521**	.946	1.293*	2.275**
SAT	.878**	1.063*	1.254**	1.742**	.935*	1.110**	1.194**	1.810**
Class Rank	.991**	1.012**	1.016**	1.040**	.988**	1.014**	1.022**	1.033**
N	12,754				12,861			

Notes: Base category = two-year institutions. PSE = postsecondary education.

* $p < .05$; ** $p < .01$.

Section B: Enrollment Versus Admission Data

We map college destinations for all three cohorts according to selectivity tier using enrollment rather than admission data for several reasons. First, detailed information on college application is available for the NELS data, but not for the HS&B and C&B data. Second, admission data in national surveys, which is based on students' self-reports, is of questionable quality and suffers from very large amounts of missing data (see Kane 1998, note 9). Our diagnostic assessment based on the NELS data reveals that about 33 percent of students who attended a postsecondary institution failed to report admission information. Acceptance rates are suspiciously high as well. The bias stemming from this data is worrisome, especially for the more selective destinations where the sample sizes are very small. Missing admission data is correlated with minority group status (Long 2004).

A third problem stems from inconsistent college destinations—for about 16 percent of the NELS students, the selectivity level of actual college destinations exceeds that of the “best” school to which they reported having been accepted.¹ Because the data about admission is suspect for *about half of the students* (33 percent missing and an additional 16 percent with inconsistent destinations), statistical analyses are likely to be severely biased. Most importantly, our concern with group differences in access to college cannot be answered with available data on admissions.

Finally, attendance patterns by selectivity tiers reveal a great deal about admission decisions because the difference between admission and tier-specific enrollment is not large.

Tier-specific yields (enrollment/admission ratios) are much higher than institution-specific yields. For example, when students are admitted to both Princeton and Yale, the tier-specific yield will be high regardless of their actual institutional choice. Using the NELS admission data we find that the yield for the more selective institutions is 82 percent. Moreover, diagnostic analyses based on NELS data produce strikingly similar multivariate results for both admission and enrollment destinations (results are available upon request). Similar conclusions about the relative weight of SAT, class rank, and minority students' advantages reinforces our claim that inferences based on tier-specific enrollment reasonably reflect the factors that govern admissions decisions.

¹ For example, a student reported Rutgers University to be the best college that admitted her, but eventually attended Yale University.

ONLINE SUPPLEMENT
to article in

AMERICAN SOCIOLOGICAL REVIEW, 2007, VOL. 72 (AUGUST:487–511)

References

- Allison, Paul D. 2001. *Missing Data*. Thousand Oaks, CA: Sage Publications, Inc.
- Collins, Linda M., Joseph L. Schafer, and Chioming Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6:30–51.
- Kane, Thomas J. 1998. *Misconceptions in the Debate Over Affirmative Action in College Admissions*. Cambridge, MA: Harvard Education Publishing Group.
- King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44:347–61.
- Little, Roderick J. A. 1993. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88:125–34.
- Little, Roderick J.A. and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- Little, Roderick J. A. and Nathaniel Schenker. 1995. "Missing Data." Pp 55–66 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by G. Arminger, C. C. Clogg, and M. E. Sobel. New York: Plenum Press.
- Long, Mark. 2004. "Race and College Admissions: An Alternative to Affirmative Action." *The Review of Economics and Statistics* 86:1020–33.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London, UK: Chapman & Hall.
- . 1999. *NORM: Multiple Imputation of Incomplete Multivariate Data Under a Normal Model*, Version 2. Software for windows 95/98/NT.
- Tomz, Michael, Jason Wittenberg, and Gary King. 2003. "Clarify: Software for Interpreting and Presenting Statistical Results." Retrieved April 22, 2007 (<http://gking.harvard.edu/clarify/docs/clarify.html>).