IMPUTATION OF TEST SCORES IN THE NATIONAL EDUCATION LONGITUDINAL STUDY OF 1988 (NELS:88)

Maxime C. Bokossa, Gary G. Huang, Synectics for Management Decisions, Inc.,
Michael P. Cohen, National Center for Education Statistics
Maxime C. Bokossa, Synectics, 1901 North Moore Street, Suite 900, Arlington, VA 22209

Key Words: PROC IMPUTE, Hot-deck imputation method, IRT theta scores, Simulation study

Abstract

This paper presents results from a project that consisted of two components: (1) evaluating two methods (a *model-based random imputation* method called PROC IMPUTE and a *within-class random hot-deck imputation*) for imputing missing cognitive test scores in the National Education Longitudinal Study of 1988 (NELS:88), and (2) using the best method to impute missing cognitive test scores in the second follow-up (F2) of NELS:88. After examining and selecting a range of auxiliary variables that are conceptually and empirically related to the F2 test scores, we conducted a simulation study to determine the "best" of two imputation methods for the purpose. Then we used that method to impute missing F2 test scores with cross-sectional F2 data and base-year through F2 panel data. The imputation covered the Item Response Theory (IRT) theta scores in math and reading. Other measurement scales (proficiency scores, standardized scores, and the number right scores) may be subsequently converted using the theta scores.

In the simulation, PROC IMPUTE provided better results than the random hot-deck imputation method for the math and reading cognitive test scores in the second follow-up (F2) of NELS:88.

Background

NELS:88 is the only current National Center for Education Statistics (NCES) dataset that contains scores from cognitive tests given to the same set of students across multiple points in time. The resulting longitudinal test data offer the possibility of researching cognitive gains from middle school through high school—an attractive feature.

In NELS:88, the respondents' cognitive ability and the growth (cognitive gains) from 8th through 12th grades at the group and individual levels were measured by a calibrated scale based on Item Response Theory (IRT). This calibration process requires that items are relatively unifactorial across grades in each subject area; that is, with the same dominant factor underlying all test forms in a given subject, say, math (Rock and Pollack 1995). There should be a common set of "anchor" items across adjacent forms, and most content areas should be represented in all grade forms. In NELS:88, the increasingly difficult levels from 8th through 12th grades were created by raising the problem-solving demands in the existing content areas and adding new content in the later forms, especially at 12th grade.

IRT assumes that a test taker's probability of answering an item correctly is a function of his or her ability and one or more characteristics of the test item itself. The three-parameter IRT logistic model uses the pattern of right, wrong, and omitted responses to the items administered in a test form, and the difficulty, discriminating ability, and "guess-ability" of each item, to place each test taker at a particular point, θ (theta), on a continuous ability scale. The probability of a correct answer (called the theta score) on item i can be expressed as:

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-1.702 a_i(\theta - b_i)}},$$

where θ is the ability of the test taker, a_i is discrimination of item i, or how well the item distinguishes between ability levels at a particular point, b_i is the difficulty of item i, and c_i is the "guess-ability" of item i.

A computer program is used to calculate the marginal maximum-likelihood estimates of the IRT parameters that best fit test takers' responses (Muraki and Bock 1991). To assess the models' match with the test data, one compares the IRT-estimated parameters with the actual proportion of correct answers to a test item for test takers grouped by ability. If the IRT-estimated curves and the actual data points match closely, then the theoretical model represents the data accurately. After the parameters for a set of test items are calibrated on the same scale as the test takers' ability estimates, a test taker's probability of a correct answer

to each item in the test battery can be estimated, even for items that were not administered to the test taker. Theta scores can be used to derive other test scores: the IRT-estimated number correct score in a subject area is the sum of the probabilities of correct answers for the items in the area. However, as is inevitable in any survey, some cases in the NELS:88 cognitive test data are missing in each round due to absence, nonparticipation, or results that were unscorable because of too many unattempted test items. This missingness problem is more severe for math theta scores in the second follow-up (22.9 percent missing scores) than in the earlier two rounds of tests (3.7 percent and 6.0 percent missing scores for the base-year (BY) and the first follow-up (F1), respectively), as shown in table 1.

Table 1. Number of students and mean math scores by test missing status

Test missing status	# of students	Mean math standardized scores		
		BY	F1	F2
Total BY-F2 panel	16,489	--	--	--
Completed all tests	11,832	46.16	51.53	54.80
Missing BY only	415	--	48.86	51.94
Missing F1 only	444	42.60	--	49.40
Missing F2 only	3,117	43.96	48.62	--
Missing BY and F1	23	--	--	44.63
Missing BY and F2	130	--	44.73	--
Missing F1 and F2	486	40.09	--	--
Missing all tests	42	--	--	--

The sample weighting adjustment cannot fully solve the problem resulting from survey non-response, neither in theory nor in practice (Rubin 1996). Specifically, the bias generated by missing cognitive scores cannot be corrected by the NELS:88 sampling weights because the weights were constructed to remedy unit non-response, not item non-response (Ingels et al. 1994, p. 70). In fact, the joint impact of item non-response to cognitive tests and unit non-response on NELS:88 tends to damage the data quality to a potentially dangerous extent. The weighted percentage of students who took all four cognitive tests in all three waves of the survey was 65 percent of the eligible core panel sample (see Rock and Pollack 1995, table 1.1, p. 2).

In addition, Rock and Pollack (1995, pp. 53-56) demonstrated that the missingness pattern of the F2 test scores across demographic subgroups was not completely at random. Our tabulation of the BY-F2 panel data confirms this. Table 2 presents a comparison of the rate of missing F2 test scores for some basic demographic subgroups of students in the BY-F2 panel who completed all three tests and those who missed the F2 test. It shows that minority students and students in the lowest socioeconomic (SES) quartile were more likely than others to miss the test. Thus, NELS:88 estimates of academic performance based on the available cases could be biased.

Table 2. Number of students and mean math theta scores by sex, race, and SES quartile

		# of students	Missing rate	Math Mean
TOTAL		16,489	22.9%	54.5
Sex	Female	8,349	23.0%	53.9
	Male	8,144	22.8%	55.1
Race[1]	Wh/As.[2]	12,657	21.5%	56.1
	Minority	3,823	27.5%	48.6
SES quar-tile	1st	4,121	27.5%	47.8
	2nd	4,095	22.2%	52.2
	3rd	4,147	21.4%	55.5
	4th	4,125	20.5%	61.8

[1] There are, respectively, 6 and 9 cases with missing data on race/ethnicity for the F2 and BY-F2 panels.
[2] "Wh/As." stands for White/Asian combined.

The gain measure, which is of critical utility in NELS:88 longitudinal research, is thus built upon test data with high levels of item non-response. To assure NELS:88 data quality, strategies other than weighting are needed to address the item non-response problem. Imputation of missing test scores is one viable strategy. Our approach to NELS:88 cognitive test score imputations could be applicable to similar problems likely to arise in the Early Childhood Longitudinal Studies (ECLS), conducted by NCES, which will also include multiple rounds of cognitive tests.

It is feasible to impute F2 cognitive test scores because a great deal of information is available to reasonably predict the missing scores. This information includes student sociodemographic background, school experience (e.g., coursework, ability and curriculum program placements, and enrichment activity participation), self-reported achievement level, and available scores in other subjects. Furthermore, the general pattern in which such predictive variables relate to achievement is known in the educational research literature. We developed our imputation models based on such knowledge.

Approach

We decided to impute the IRT-estimated theta scores (in two F2 subject areas, math and reading) since theta scores are the original estimates of the test takers' probability of correctly answering items in a given set

of test items.

We found that in F2 of NELS:88 the missing test scores were not "missing completely at random" (or MCAR as defined by Little and Rubin 1987). That is, the cases that did not have scorable tests in the second follow-up were systematically different from the cases that had completed the three tests in a variety of auxiliary variables, including background and schooling (see table 2 and Rock and Pollack 1995, pp. 53-56). Such non-MCAR missingness patterns call for imputation based on information for a subsample that had completed test scores but shared attributes with the missing cases. Our approach included three steps:

1. Examine a range of candidate variables in order to select the best auxiliary variables;
2. Conduct a simulation study to determine the "best" of two imputation methods used by NCES; and
3. Impute missing F2 test scores with cross-sectional F2 data.

Selection of Auxiliary Variables

We examined a group of candidate variables to identify those which were related to test missingness. The candidate variables were race, sex, SES, coursework in the target subject areas, advanced academic program placement, F1 and F2 dropout status, early graduation status, and BY and F1 cognitive test scores. To determine their utility in the imputation model, we examined bivariate correlation between these variables and the cognitive test scores in two subject areas (math and reading). We then selected variables that correlated highly with theta scores. Next we identified important predictors of the cognitive test outcome by fitting regression models. The final regression model reflected test scores that were homogeneous within the imputation classes defined by the covariates.

Simulation Study

We studied two imputation techniques, namely, a *model-based imputation* method implemented by computer software called PROC IMPUTE and a *within-class random hot-deck imputation* method that has been used by NCES in other surveys. The study included simulating a few levels and patterns of missingness (about 20 percent of the data were made missing) in the NELS:88 BY-F2 panel cases where the BY, F1, and F2 test scores are all non-missing. We compared statistics derived from the incomplete data with the data after imputing simulated missing cases. Three criteria were used to compare the accuracy of the two types of imputations: the average imputing error, the bias of the variance, and the mean bias. The method with the "best" scores based on these criteria was used in the next step (i.e., the method with the least average imputing error and mean bias and with the least distortion in variance).

The relative bias of variance estimate is defined as

$$Relative\ \ Bias = \frac{(Estimated\ \ Var)\ -\ (True\ \ Var)}{True\ \ Var}$$

and the average imputation error is defined as

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i^* - y_i)^2}$$

where m is the number of missing values, y_i is the true value which is intentionally set to missing, and y_i^* is the imputed value for the i-th missing case. That an imputation method has smaller average imputation errors only implies that the method provides imputations on average closer to the real values. This does not necessarily mean that it gives more accurate estimates for all types of statistics, although that is true in many situations.

Description of Imputation Methods

Within-class random hot-deck imputation: Since we understand reasonably well the factors related to F2 test non-response and have data on such factors, we could assume model-based approaches would probably produce more accurate imputation than randomization-based approaches if the model assumptions were satisfied (Hu and Salvucci 1999). Thus, we imputed the IRT-estimated number of the right score in each subject using F2 cross-sectional data on student sociodemographic and socioeconomic background, academic coursework, self-reported grade average point, and available test scores on subjects other than the one to be imputed.

For the implementation of the within-class random hot-deck imputation method, we first sorted the dataset by the auxiliary variables in order to obtain homogeneous cells called imputation classes. To impute a missing value in a given imputation class, we randomly selected an observed value of the target variable in that class to fill-in for the missing value.

PROC IMPUTE: To overcome the underestimation of variance which is typical in a hot-deck imputation method or a regression-based imputation method, we also added disturbance by using the software package PROC IMPUTE (McLaughlin 1991).

PROC IMPUTE combines the procedures of regression-based and data sampling (often called "hot-deck") methods. Regression involves generating a function, $\hat{y} = f(x_1, x_2, ..., x_p)$, that relates a "target" variable (cognitive test score) to auxiliary variables, then uses the function along with the existing values of the auxiliary variables to compute $\hat{y}$ whenever it is missing. Data sampling involves subsetting the data on the basis of relevant variables and randomly selecting a value for the target variable from an available target variable within the same subset.

PROC IMPUTE considers each variable on the file in turn as a target variable whose missing values are to be filled in, and it uses information on other variables to minimize the error in imputing each target variable. Three steps are taken to impute each variable in PROC IMPUTE.

First, stepwise regression analyses are performed "simultaneously" for each variable. During these analyses, an ordered list of the imputation variables is constructed. The regression analysis for each variable uses as predictors all the complete variables, including the previously imputed variables. The process terminates when there are no more permissible predictors that provide a significant improvement of fit in the prediction of any of the target variables. Second, homogeneous cells (imputation classes) are created for records that have close predicted regression values. Finally, two donors are drawn from the adjacent cells. Each missing record in a given cell is imputed with a weighted average of these two donors with probability proportional to the observed frequencies within the two cells.

PROC IMPUTE runs all the imputation procedures automatically and generates a dataset in which all the records are complete. Imputed data flags are also automatically created by the software and set for each variable; a value of "I" corresponds to imputed values, "R" to reported values, and "A" to skip missing values.

Simulation Results

Math Theta Score

We used the F2 panel sample members that had non-missing math theta scores and non-missing information for the following auxiliary variables: sex, race, socioeconomic status, units in foreign languages, units in physics, base-year grade composites, and teacher's opinion about student attending college. We generated 1,996 cases, about 20 percent, from the F2 panel members and set their math theta scores as missing. To simulate the actual missingness pattern, the rate of missingness across sex, race/ethnicity, and SES quartiles mimicked that of the actual F2 test missing cases. We used PROC IMPUTE and random hot-deck to impute these simulated missing cases. The mean and variance for the math scores were calculated for the following four groups:

1. A group of 10,248 cases in the F2 panel that reported the math theta scores and auxiliary variables specified above;
2. A group that included the 8,252 cases with actual math theta scores and 1,996 cases with imputed scores using PROC IMPUTE;
3. A group that included the 8,252 cases with actual math theta scores and 1,996 cases with imputed scores using the hot-deck method; and
4. A group of 8,252 cases with actual math theta scores (the 1,996 cases were deleted as "missing"). This group simulates the current scenario in NELS:88 where there are missing test scores, but no imputation has been used.

Group 1 estimates served as the "true scores." Group 2, Group 3, and Group 4 estimates were compared with the true Group 1 estimates to examine if Group 2 (with PROC IMPUTE imputation) did better than Group 3 (with hot-deck imputation) and Group 4 (with no imputation). Table 3a provides the results for average imputation error for the math theta score. Then figure 1 compares the results for the bias of the mean, while table 3b presents the relative bias of the variance for the math theta score. Note that in the race variable, "Wh/As." or "White/Asian" stands for Whites and Asians combined. These groups were combined because preliminary results had shown that both Whites and Asians have on average higher math scores than the other minority groups.

Table 3a. Percentage of missing values and average imputation error for math score

		# of students	% imputed	Average imputation error	
				Hot-deck	PROC IMP.
TOTAL		10,248	19.5%	14.50	13.56
Sex	Female	5,139	20.2%	14.51	13.23
	Male	5,109	18.8%	14.49	13.90
Race	Wh/As.	8,196	19.0%	14.32	13.58
	Minority	2,052	21.3%	15.10	13.49
SES quar -tile	1st	2,176	20.3%	14.34	13.82
	2nd	2,596	19.7%	14.98	14.16
	3rd	2,734	19.3%	14.18	12.77
	4th	2,742	18.8%	14.47	13.51

About 20 percent of the math scores were imputed using first PROC IMPUTE, and then the random hot-deck imputation method. The average imputation error is consistently lower for PROC IMPUTE than it is for hot-deck in every single category of the sociodemographic group of study, and overall.

Figure 1. Comparison of bias of the mean for math score

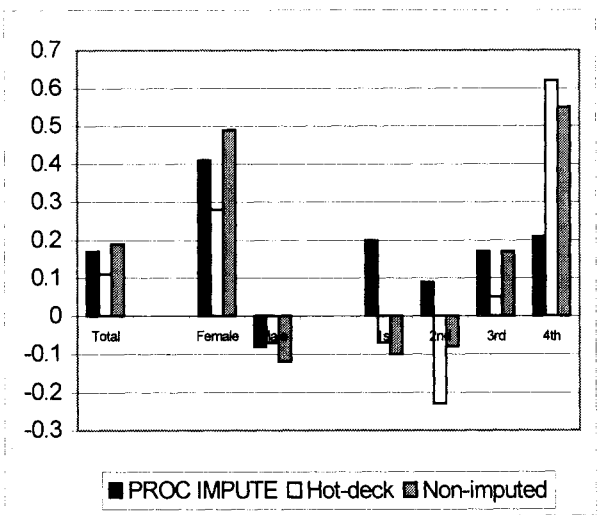


Figure 1 above shows the bias of the mean after using PROC IMPUTE and the random hot-deck imputation method. Figure 1 also presents the bias of the mean for the incomplete math score without any imputation. It turns out that none of the three groups compared shows a consistent improvement over the other two groups, across all the sociodemographic categories.

Table 3b. Comparison of relative bias of variance for math score

		Relative bias of variance		
		Non imputed	Hot-deck	PROC IMP.
TOTAL		0.055	0.060	0.001
Sex	Female	0.053	0.069	-0.005
	Male	0.061	0.056	0.010
Race	White/Asian	0.059	0.068	0.018
	Minority	0.046	0.076	0.021
SES quar-tile	1st	0.036	0.051	-0.003
	2nd	0.053	0.049	0.009
	3rd	0.062	0.076	0.005
	4th	0.002	-0.009	-0.021

Table 3b shows that the relative bias of the variance is consistently smaller for PROC IMPUTE than it is for hot-deck and the non-imputed group, in each group of the sociodemographic group of study, and overall, with the exception of the fourth quartile of the socioeconomic status category.

Reading Theta Score

For the reading cognitive test score simulation study, we used the F2 panel sample members that had non-missing reading theta scores and non-missing auxiliary variables. The auxiliary variables considered here were sex, race, socioeconomic status, units in foreign languages, units in reading, units in chemistry, grade composites from base-year, and teacher's opinion about student attending college. We selected 2,017 cases, about 20 percent, from the F2 panel members and set their reading theta scores as missing. We used PROC IMPUTE and random hot-deck to impute these simulated missing cases. The mean and variance for the reading scores were calculated for the following four groups: (1) group of 10,249 cases in the F2 panel that reported the reading theta scores and auxiliary variables specified above; (2) group of 8,232 cases with actual reading theta scores and 2,017 cases with imputed scores using PROC IMPUTE; (3) group of 8,232 cases with actual reading theta scores and 2,017 cases with imputed scores using the hot-deck method; and (4) group of 8,232 cases with actual reading theta scores.

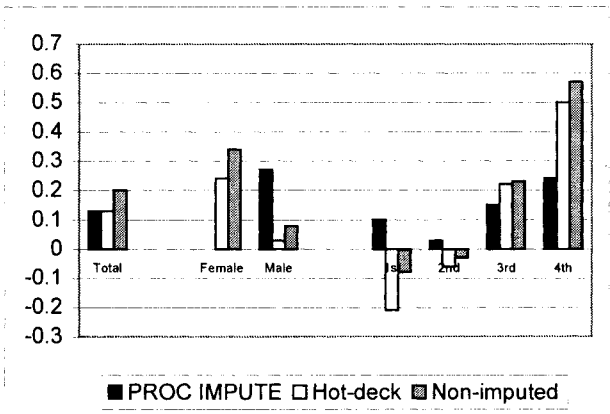
Table 4a provides the calculated average imputation error for the reading theta score, figure 2 displays the calculated bias of the mean, and table 4b presents the calculated relative bias of the variance for the imputed reading theta score using PROC IMPUTE, random hot-deck and no imputation. Note that, unlike the math test score, the race variable here is categorized by Whites on one hand and all minority groups on the other hand.

Table 4a. Percentage of missing values and average imputation error for reading score

		# of students	% imputed	Average imputation error	
				Hot-deck	PROC IMP.
TOTAL		10,249	19.7%	14.70	13.86
Sex	Female	5,144	20.0%	14.50	13.86
	Male	5,105	19.4%	14.90	13.85
Race	White	7,594	19.3%	14.48	13.63
	Minority	2,655	20.8%	15.27	14.44
SES quar-tile	1st	2,178	20.0%	14.69	14.36
	2nd	2,594	19.5%	15.66	14.14
	3rd	2,738	20.2%	14.27	13.51
	4th	2,739	19.1%	14.19	13.51

As in the simulation of math theta scores, around 20 percent of the reading scores were set to missing and imputed using first PROC IMPUTE and then random hot-deck imputation method. The average imputation error is consistently lower for PROC IMPUTE than it is for hot-deck, in every single category of the sociodemographic group of study, and overall.

Figure 2. Comparison of bias of the mean for reading score



Note that the bias of the mean for female reading theta score is zero for PROC IMPUTE.

The bias of the mean does not show that any particular method is consistently better across all sociodemographic categories.

However, the relative bias of the variance is consistently smaller for PROC IMPUTE than it is for hot-deck and the non-imputed group, in each category of the sociodemographic group of study, and overall, with the exception of the third and fourth quartile of the socioeconomic status category (see table 4b).

Table 4b. Comparison of relative bias of variance for reading score

		Relative bias of variance		
		Non imputed	Hot-deck	PROC IMP.
TOTAL		0.034	0.037	-0.009
Sex	Female	0.035	0.031	0.005
	Male	0.028	0.039	-0.015
Race	White	0.035	0.038	-0.001
	Minority	0.038	0.035	0.004
SES quar-tile	1st	0.024	0.030	0.021
	2nd	0.035	0.021	-0.003
	3rd	0.018	0.029	-0.036
	4th	-0.011	-0.002	-0.038

Conclusion

We found that PROC IMPUTE was the preferred method for imputing missing cognitive test scores in the National Education Longitudinal Study of 1988 (NELS:88), because it produced better results than both the random hot-deck imputation method and no imputation in the simulation study that we conducted using the math and reading cognitive test scores in the NELS:88 second follow-up (F2) data.

References

Hu, M., and Salvucci, S. (1999), *Evaluation of Some Popular Imputation Algorithms*, 1998 Proceedings of the Section on Survey Research Methods, pp. 308-313, Alexandria, VA: American Statistical Association.

Ingels, S. J., Dowd, K. L., Baldridge, J. D., Stipe, J. L., Bartot, V. H., and Frankel, M. R. (1994), *National Education Longitudinal Study of 1988: Second Follow-Up: Student Component Data File User's Manual* (NCES 94-374), Washington, DC: NCES.

Mclaughlin, D. H. (1991), *Imputation for Non-Response Adjustment*, Internal Report, American Institute for Research: Palo Alto, California.

Muraki, E. J., and Bock, R. D. (1991), *PARSCLE: Parameter scaling of rating data* (computer software), Chicago, IL: Scientific Software, Inc.

Rock, D. A., and Pollack, J. M. (1995), *Psychometric Report for the NELS:88 Base Year Through Second Follow-Up* (NCES 95-382), Washington, DC: NCES.

Rubin, D. B. (1996), *Multiple imputation after 18+ years*, Journal of the American Statistical Association 91, 473-489.