Appendix F: Methodology for Rasch Analysis and Multiple Imputation

This study used a Rasch model (Wright & Stone, 1979) combined with multiple imputation (Rubin, 1987) as an alternate approach for handling missing responses to the Algebra Achievement test. A Rasch model can be thought of as a special case of a class of models called "Item Response Theory" (IRT) models. When implemented together with an IRT model, multiple imputations have often been referred to as "plausible values" (Mislevy, Beaton, Kaplan, & Sheehan, 1992). This appendix discusses the theory behind both multiple imputation/plausible values and Rasch models, and then describes how these concepts were implemented by the current study.

Multiple Imputation

As noted in Chapter 4, there were a number of students who participated in this study and completed 18 of the 19 items on Test 1, as well as a number of students who completed 13 of the 14 items on Test 2. If a student was missing a response to one item on a test, the analysis in Chapter 4 used OLS regression to impute the student's response to that particular item based on her or his responses to the other items on the test. Using this imputed response, the student's mean score on the test was computed and analyzed as part of the complete data set.

Rubin (1987) has demonstrated that this "single imputation" approach to missing data is not ideal because it underestimates the variance that would have been observed had all responses been available. Instead, it is better to use "multiple imputation." Multiple imputation creates the probability distribution of each student's possible responses, and then for each student draws a response (known as a "plausible value") from this probability distribution. OLS regression parameters are computed for this

"draw" of imputed plausible values. Then, the process is repeated. A second "draw" of plausible values is taken from the probability distribution, and a second set of parameters is computed. This can be repeated any number of times. The mean value of the parameter estimates from all the draws is the reported parameter estimate, which has a t-distribution with degrees of freedom and standard deviation that can be computed with formulas reported below. Even two "draws" of plausible values provide a relatively efficient parameter estimate and accurate confidence interval. Five draws are commonly used, and in most cases five draws provide efficiency and accuracy of confidence intervals approaching what could be accomplished with an infinite number of draws.

While theoretically superior to single imputation, using multiple imputation for a missing student response to one item would have been rather like swatting a mosquito with a sledge hammer. The missing response comprised, respectively, one-nineteenth of Test 1 or one-fourteenth of Test 2, and underestimating the variance of that one response was unlikely to impact final results. Multiple imputation was unlikely to be familiar to most readers of the current study. Implementing such a complex methodology to account for a missing response to a single item would thus provide little advantage, while potentially confusing readers of the study.

However, students who attended only the first day of testing, while responding to 18 of the 19 items on Test 1, responded to only 1 of the 14 items on Test 2. For this reason, there was very little data available to estimate their Test 2 score, and the Test 2 scores of these students were not used in the primary analysis reported in Chapter 4. Similarly, students who attended only the second day of testing responded to only 1 of

the 19 items on Test 1, and their Test 1 scores were not used in the primary analysis reported in Chapter 4.

Dropping these scores was a potential problem, because it seemed reasonable to suppose that students who missed one day of testing might have systematically different achievement than students who attended both days. Doing without these students' data could potentially bias the analysis.

Imputing student scores on Test 1 or Test 2 from responses to a single item is very different from imputing scores from responses to all but one item. Most of the data for these students is missing, and underestimating the variance in the missing data could bias results in important ways. Thus, if the scores of students who responded to a single item were to be included in OLS regression estimates, multiple imputation would be necessary.

Imputing plausible values from student responses to a single item was difficult for two reasons. First, using standard methods like OLS regression for imputing a mean score on a test composed of 14 or 19 items from a score on just one of those items would be likely to provide a very inaccurate estimate. Second, students who were missing data might be systematically different from students with complete data. In the terminology used by Rubin (1987), the non-response may be "nonignorible." (Readers may be more familiar with the concept "missing at random." Rubin has demonstrated that, in most cases, "missing at random" as commonly defined is equivalent to "nonignorable.") In such a situation, even when students with complete data are used to develop a probability distribution of complete scores given the score on just one item, this probability distribution may not be accurate for students with missing data. Such students are

perhaps systematically different from the complete-data students whose responses were used to develop the probability distribution. Thus, a method must be devised to modify the probability distribution to account for possible systematic characteristics of students with missing data.

Fortunately, a Rasch model solves both problems. It provides a good way to estimate student abilities based on just one response. Further, it provides a way to modify these estimates to account for systematic differences between responders and non-responders, that is, to adjust for nonignorable non-response. The next section describes how this is accomplished.

The Rasch Model

Arguments have raged over the relative merit of Rasch versus other IRT models. This study chose a Rasch approach over competing IRT models not based upon any theoretical argument, but rather due to the practical considerations that 1) other IRT models often require a larger sample size than was available for this study and 2) an excellent software package implementing a Rasch model was available. This section provides a brief description of the theory behind the Rasch approach.

The simplest Rasch model describes a test composed of dichotomous (right/wrong) items designed to measure a single, unidimensional "ability". A person's likelihood of answering any given item correctly is a function of the person's latent ability β and the item's difficulty δ. Specifically, the "odds of getting an item correct" is modeled as $e^{\beta-\delta}$, or equivalently, $\beta-\delta$ is the "log-odds" of a person with ability β getting an item with difficulty δ correct.

This study used the ConQuest computer program (Wu, Adams, & Wilson, 1998) to perform its analyses. Like other Rasch modeling software ConQuest simultaneously estimates item difficulties and person abilities, based on person's responses to a set of items. ConQuest generalizes the Rasch model for dichotomous data in a number of ways. First, it accommodates polytomous responses like those used on the Algebra Achievement test, implementing an adaptation proposed by Masters (1982). Second, the approach of ConQuest is multivariate. For example, in the current analysis ConQuest estimated student scores on Test 1 and Test 2 simultaneously, computing a student's

likely ability on each test while accounting for the correlation between Test 1 and Test 2. Finally, ConQuest uses a Bayesian approach to simultaneously estimate multivariate ability scores and to perform "latent regression", that is, to estimate regression parameters that predict those responses. The approach is iterative: each change in a regression parameter estimate causes changes in estimates of each item's difficulty and of each person's ability, which in turn cause changes in the regression parameter estimates. This process is repeated until the program converges on a solution.

A Rasch model like the one used by ConQuest can handle missing responses more appropriately than other approaches. This is true because the model adjusts a person's ability score based on the difficulty of the items answered. Thus, someone who responds to a few easy items correctly while not having the opportunity to answer the rest of the test will not be assigned as high an ability as someone who responds to a few difficult items correctly while not having an opportunity to answer the rest of the test. Also, the latent regression model used by ConQuest inherently weights a person's ability estimate by the error of measure of that estimate. An ability estimated based upon a few responses will usually not receive as much weight as an ability estimated by many responses, due to the inherent uncertainty in measuring the former.

If non-response were ignorable, ConQuest's iterative approach could provide a valid estimate of regression parameters and standard deviations directly, that is, without needing to run an OLS regression on plausible values or on any other explicit estimate of student ability. See Mislevy (1984) for more details on how this is accomplished. Unfortunately, the non-response in the current study was not ignorable, that is, student responses could not be assumed to be missing at random. For the current study, while some students with missing data scored very high on the part of the test they did complete, on average those with missing data scored noticeably lower on the parts of the test they completed. Thus, it was reasonable to assume that lower-ability students were more likely to be missing responses. For this reason, "plausible values" were utilized for the analysis.

The advantage of having the Rasch model compute plausible values for the current study was as follows. It was possible to estimate plausible values based on all the information in the model, including the partial responses of students with missing information, their Grade 6 ability, and an indicator that they had missed a day of testing. By including the indicator, individuals who missed a day had their estimated "ability" shrunken towards the mean of all students who missed a day. That is, the model accounted for the fact that missing a day predicted lower ability, over and above what was recorded by a person's actual responses. In short, the Rasch model used a different probability distribution for students with missing responses than it used for students with complete responses. It shrunk the estimated score of students who missed a day of testing towards the mean score of all students who missed that day of testing, while shrinking the score of students with complete responses towards the (higher) mean score of all students who attended both days of testing. In this way, the model generated plausible values that accounted for the nonignorable nature of non-response.

Implementation Details

Five draws of plausible values were estimated using as inputs student item responses, sixth grade ability scores, an indicator of whether a student had missed the first day of testing, and an indicator of whether a student had missed the second day of testing. However, the two indicator variables were not utilized in the subsequent five OLS regression runs.

OLS regression parameters estimated from a model including indicators of "absence" would not be appropriate, because they would estimate the effects of Treatment after "controlling for" absence. Because of different testing conditions, students in the Reform cohorts were slightly more likely to have missed a day of testing than were students in the Traditional cohort. "Conditioning out" the absentee variable by including it in the model would bias the results in favor of the Reform cohorts.

However, draws of plausible values estimated while conditioning on the "absentee" variable could legitimately be used in a series of Optimum Least Squares (OLS) regression that did not use the "absentee" variables. This what the current study did.

For each draw of plausible values, OLS regression was used to compute the effect on achievement of being in a Reform cohort, controlling for prior ability but not for the "absentee" indicator variables. For each measure (Test 1 and Test 2) the final estimate of the effect of being in a Reform cohort was computed as the average of the five OLS results. That is, if the output estimate of the Reform cohort effect from the i^{th} draw of plausible values is p_i, then the estimated Reform cohort effect Q reported in Chapter 4 is $Q = \Sigma p_i/5$. Following the formula of Rubin (1987) variance for Q was computed as follows:

Let p_i be a parameter estimate from the i^{th} run of plausible values and σ_i be its standard deviation. Then the estimated variance of Q is the sum of two values:

1) U = the mean of the five estimates, or mean(σ_i^2), plus
2) $\frac{}{}$ (n+1)/n * B where n = number of plausible values drawn, and B= the variance of the n parameter estimates, that is Variance(p_i). In this case, since there were five draws, the B was multiplied by 6/5.

The estimated parameter then has a *t*-distribution, with
Mean = $\overline{Q}$,
Variance = $(\overline{U} + (n+1)/n * B), = \overline{U} + 6/5*B$
Degrees of Freedom= $(n-1)*(1+r^{-1})^2 = 4*(1+r^{-1})^2$ where
$r = (n+1)/n*B/\overline{U} = 6/5*B/\overline{U}$.

As reported in Chapter 4, results of this analysis led to conclusions little different from the simpler OLS method of dealing with missing data. This confirms that the results reported in Chapter 4 are reasonable, despite the missing data.

References Used in Appendix F

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-74.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49(3), 359-381.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 133-61.

Rubin, Donald B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Wright, Benfamin D. & Stone, Mark H. (1979). Best test design: Rasch measurement. Chicago, IL: Mesa Press.

Wu, Margaret L. , Adams, Raymond J., & Wilson, Mark R. (1998). ACER ConQuest: Generalised item response modeling software. Melbourne, Australia: Australian Council for Educational Research.