5-1-2013

JMASM 32: Multiple Imputation of Missing Multilevel, Longitudinal Data: A Case When Practical Considerations Trump Best Practices?

Jennifer E. V. Lloyd
University of British Columbia

Jelena Obradović
Stanford University

Richard M. Carpiano
University of British Columbia

Frosso Motti-Stefanidi
University of Athens, Greece

JMASM 32: Multiple Imputation of Missing Multilevel, Longitudinal Data: A Case When Practical Considerations Trump Best Practices?

Algorithms & Code
JMASM 32: Multiple Imputation of Missing Multilevel, Longitudinal Data: A Case When Practical Considerations Trump Best Practices?

Jennifer E. V. Lloyd	Jelena Obradović	Richard M. Carpiano	Frosso Motti-Stefanidi
University of British Columbia, Vancouver, Canada	Stanford University, Stanford, CA, USA	University of British Columbia, Vancouver, Canada	University of Athens, Athens, Greece

A pedagogical tool is presented for applied researchers dealing with incomplete multilevel, longitudinal data. It explains why such data pose special challenges regarding missingness. Syntax created to perform a multiply-imputed growth modeling procedure in Stata Version 11 (StataCorp, 2009) is also described.

Key words: Missing data, longitudinal data, multilevel data, multiple imputation, growth modeling, Stata.

Introduction

One research challenge faced when conducting a longitudinal study is selecting a method for handling missing data. Incomplete assessment histories for longitudinal study participants are ubiquitous (Allison, 2002; Jeličić, Phelps & Lerner, 2009), and are due to multiple factors, such as participants' attrition, illness, unwillingness or inability to answer certain questions, and problems related to the methods of data collection.

When considering how longitudinal data are inherently structured – with repeated measurements (at level-one) clustered or nested within individual participants (at level-two) – such data are in effect multilevel or hierarchical

Jennifer E. V. Lloyd is a Research Associate at the Human Early Learning Partnership (HELP). Email her at: jennifer.lloyd@ubc.ca. Jelena Obradović is an Assistant Professor in the Graduate School of Education. Email her at: jelena.obradovic@stanford.edu. Richard M. Carpiano is an Associate Professor in the Department of Sociology. Email him at: richard.carpiano@ubc.ca. Frosso Motti-Stefanidi is a Professor in the Department of Psychology. Email her at: frmotti@psych.uoa.gr.

(complex) (Zaidman-Zait & Zumbo, 2013). Hence, incomplete assessment histories may affect the availability of values for both time-varying and time-invariant variables.

Ignoring missing data can significantly bias estimates of coefficients and standard errors, inflate Type I and II error rates, degrade confidence intervals and/or distort statistical power (Acock, 2005; Allison, 2002; Collins, Schafer, & Kam, 2011; Little & Rubin, 2002; Schafer & Graham, 2002; Zaidman-Zait & Zumbo, 2013). Therefore, missing data should be a focus of any longitudinal study, rather than being sidelined as a bother (Allison, 2002; Little & Rubin, 2002; Schafer & Graham, 2002).

Unfortunately, many of the strategies proposed to handle missingness tend to be primarily implementable in relatively rudimentary research contexts in which data lack the intricacy and "messiness" of real-life data (Carpenter, Goldstein & Kenward, 2011; Graham, 2009; Johnson & Young, 2011). To complicate matters, some best-practice studies of missing data imputation provide such stringent technical requirements for filling-in missing data that their recommendations cannot realistically be met in many real-life research contexts. Some of these requirements were generated from elementary simulated, single-level data sets (von Hippel, 2007, 2009). Furthermore, large percentages of longitudinal

researchers either did not comment on their studies' missingness or they utilized outdated and even incorrect methods to handle missingness. As a result, the longitudinal research literature is scattered with examples of bad missing data practices (Jeličić, et al., 2009).

Given the lack of a robust empirical research base centered on this problem, the lack of conclusive recommendations about precisely how to deal with the problem, and the dearth of statistical software resources that allow users to resolve this problem (Allison, 2002; Carpenter, et al., 2011; Graham, 2009), many longitudinal researchers find themselves at an impasse. To further complicate matters, graduate training in statistics, measurement and research methodology in North American universities has declined significantly in recent years, while there has been an increasing trend toward doctoral-level researchers with minimal knowledge of statistics who nonetheless conduct analyses (Aiken, West & Millsap, 2008; Aiken, et al., 1990; Merenda, 2003). For these reasons, dealing with incomplete complex data is a good idea in theory, but a challenging one in practice.

Objectives

The goal of this article is to provide a pedagogical tool for applied longitudinal researchers dealing with incomplete complex data. The first objective is to explain and illustrate why complex data pose special challenges when it comes to missingness. Inspired by the work of the UCLA Academic Technology Services' Statistical Consulting Group (n.d.a), the second objective is to provide a step-by-step description of syntax created to perform a multiply-imputed individual growth modeling procedure in a real-life longitudinal research context (Obradović, Lloyd & Motti-Stefanidi, manuscript in preparation), using Stata Version 11 (StataCorp, 2009).

Strategies for handling missing complex data will be presented, although it is not claimed that they are a perfect solution to the problem. The complex structure of data in this study was not amenable to certain imputation-related recommendations offered in the general missing data literature. In short, missingness was dealt with in the best way possible given the unanswered questions that surround missing complex data. But it is precisely because of these unanswered questions that the implementation of a modern procedure, such as multiple imputation, to fill in incomplete multilevel, longitudinal data is, in principal, justifiable until the missing data literature provides conclusive recommendations for handling missing complex data in broadly-defined longitudinal research contexts.

Different types of missing data and missing data mechanisms were discussed by Allison (2002), Collins, et al. (2011), Graham (2009), Little (1995), Little and Rubin (2002), Schafer and Graham (2002) and Zaidman-Zait and Zumbo (2013). Applications, strengths, and limitations of assorted traditional and modern methods by which to handle missing data were discussed by Acock (2005), Allison (2002), Collins, et al. (2011), Little and Rubin (2002), Schafer and Graham (2002). Working knowledge of how to run individual growth models (multilevel models of change) is assumed (Raudenbush & Bryk, 2002; Singer & Willett, 2003), as is familiarity with Stata's programming language. That said, step-by-step syntax descriptions facilitate translating the Stata commands into other programming languages or platforms.

Why Complex Data Pose Special Challenges

Multiple imputation involves four steps: (1) replication, wherein multiple copies of an incomplete data set are created; (2) imputation, wherein missing values in each data set are replaced with plausible versions of the complete data derived from multivariate data; (3) analysis, wherein each imputed data set is analyzed separately using standard methods of statistical analysis; and (4) recombination, wherein the results of the separate analyses are combined or pooled (Rubin, 1987; Schafer, 1999; Schafer & Graham, 2002; von Hippel, 2007). The process of combining results of parameter estimates and their respective standard errors from several imputed data sets has been shown to yield valid statistical inferences that reflect the uncertainty due to the missingness (Yuan, 2011).

Unfortunately, it is challenging to begin the multiple imputation process when dealing with complex data. For example, von Hippel (2009) and Allison (2002) recommended

calculating transformations such as interactions and squared terms using the incomplete data and, in turn, imputing the transformations alongside the other regular variables. This transform-then-impute approach has been shown to yield better, less-biased regression estimates than when variables are imputed in their raw form and, in turn, transformations are calculated from the imputed data (the impute-then-transform approach; von Hippel, 2009). Although the transform-then-impute approach may be possible to heed in certain single-level or simulated contexts, it is difficult to implement when dealing with complex data.

Consider a hypothetical example in which a longitudinal study involves data collected across three waves, in which Y_{ti} is the observed score at time or wave t for individual participant i. Consider further that there is one time-varying predictor, X_{ti}, and two time-invariant predictors, V_i and W_i. Then imagine that the data have been entered into a spreadsheet in person or wide format – wherein all of the records collected for an individual participant are entered along one row of the spreadsheet. As indicated in Table 1, when data are formatted this way, there is no time or wave variable (i.e., a variable that explicitly denotes the particular period of data collection). Instead, individuals' scores for time-varying variables are represented by as many separate variable names (i.e., columns in the spreadsheet) as there are waves. For example, X_i for waves 0 through 2 are respectively denoted by variables X_0, X_1, and X_2.

Although the lack of a wave variable makes it possible to create single-level interactions between time-invariant variables (e.g., V_i*W_i at level-two), it is computationally difficult to automate the inclusion of cross-level interactions in the imputation model. This difficulty is lamentable because the ability to explore cross-level interactions is one of the primary advantages of performing an individual growth modeling analysis (Holt, 2008). A cross-level interaction refers the interaction between level-two variables and level-one variables, "that is, to modification of the effects of lower level variables by characteristics of the higher level units to which the lower level units belong (or vice versa)" (Diez-Roux, 2002, p. 589).

For example, suppose a study is designed to explore the cross-level interaction between time or wave (at level-one) and the time-invariant variable W_i (at level-two). In the absence of an explicit time or wave variable, the interaction cannot be computed. Similarly, suppose a study is designed to explore the cross-level interaction between the time-varying variable X_{ti} (at level-one) and the time-invariant variable V_i (at level-two). Because X_{ti}'s time-varying values are represented by as many variable names (columns) as there are waves, there is no way of creating a cross-level product term that takes into account the temporal nature of X_{ti} while also taking into account the constant nature of V_i.

Alternatively, imagine that the same data have been entered into a spreadsheet in person-period or long format – wherein each individual's records are entered into as many rows as there are waves of data collection (in the case of the example presented, three rows per individual).

Although it is shown in Table 2 that there is a wave variable, most statistical software programs, including Stata, require data to be in person format during imputation. If not, the software erroneously views separate rows as representing separate individuals. Hence, when exploring cross-level interactions, it is not possible to take into account the within-individual covariance – the inherently nested or clustered structure of the data – whether the data are entered in person format or in person-period format (Han, 2008).

von Hippel (2009) recommended any centering of the scores of a given variable – a practice aimed at reducing collinearity and improving interpretation of the intercept (Raudenbush & Bryk, 2002) – be carried out prior to imputation. What is unclear is which problems are introduced if the mean that is being subtracted is being skewed by the variable's missing values. The benefit of centering pre-imputation is also unclear, given that centering simply linearly transforms a variable's scores into those different metric. This problem is not endemic to complex data sets alone, but highlights the questions surrounding multiple imputation and the transform-then-impute approach specifically.

Table 1: Hypothetical Example of Data Entered into a Spreadsheet in Person (Wide) Format

ID	Y_0	Y_1	Y_2	X_0	X_1	X_2	V	W
1	4	7	10	13	16	19	22	25
2	5		11	14		20	23	26
3	6	9	12	15	18	21		27

Table 2: Hypothetical Example of Data Entered into a
Spreadsheet in Person-Period (Long) Format

ID	Y	X	Wave	V	W
1	4	13	0	22	25
1	7	16	1	22	25
1	10	19	2	22	25
2	5	14	0	23	26
2			1	23	26
2	11	20	2	23	26
3	6	15	0		27
3	9	18	1		27
3	12	21	2		27

Stata Tutorial

A step-by-step description of syntax used to perform a multiply-imputed growth modeling procedure in a longitudinal research context is presented (Obradović, et al., manuscript in preparation). Although this tutorial describes specific imputation and analytic choices made with respect to the data at hand, there is no 'one size fits all' approach to addressing the problem of missing data (Johnson & Young, 2011; Yuan, 2011).

Software

Stata Version 11 (StataCorp, 2009) was used for the tutorial due to its versatile ability to perform data management tasks, multiple imputation and complex analyses. Schafer (2001) developed a statistical program, PAN, which accounts for the clustered nature of longitudinal data as part of S-Plus (Schafer, 2001; Schafer & Yucel, 2002). An imputation macro for MLwiN, REALCOM-IMPUTE, was developed for multilevel data (Carpenter, et al., 2011). Although these are exciting advancements, PAN's limited availability and accessibility (Graham, 2009) and REALCOM-IMPUTE's relative newness (with documentation focused only on non-growth model examples of nested data) indicate neither has made its way into routine use by applied longitudinal researchers.

Missing Data Procedure

Two choices of modern missing data procedures were available to implement in this study: full information maximum likelihood (FIML) or multiple imputation (MI). As Collins, et al. (2001) wrote, FIML "chooses parameter values that assign the highest possible probability or probability density to the data values actually seen, under a well-defined family of parametric probability models" (p. 334).

FIML treats the missing data as random variables to be removed from the likelihood function. This, in a sense, treats the missing data as if they were never sampled, rather than deleting or filling in the missing cases (Schafer & Graham, 2002). Little and Rubin (2002) provided detail on FIML estimation. By contrast, MI "attempts to handle the missing data aspect in advance of the substantive analysis" (Collins, et al., p. 335) by combining results of parameter estimates and their respective standard errors from several imputed data sets, pre-analysis.

According to Carpenter, et al. (2011), MI is the leading approach to handling data that are missing at random (MAR). It is implementable with a larger variety of data and statistical models than FIML (Allison, 2002; Johnson & Young, 2011). Although FIML yields more efficient estimates, it is a more case-specific (less general) approach to missingness, and is computationally more difficult (Allison, 2002; Schafer, 1999). Allison (2002) and Johnson and Young (2011) provided overviews of the advantages, disadvantages, and applications of both types of modern missing data procedures.

Approach to Multiple Imputation

The syntax features commands related to Stata 11's Imputation by Chained Equations (ICE) add-on program. Consider a dataset in which some or all of the variables, $X_1, \ldots, X_k$, have missing data:

Initially, all missing values are filled in at random. The first variable with at least one missing value, X_1 say, is then regressed on the other variables, $X_2, \ldots, X_k$. The estimation is restricted to individuals with observed X_1. Missing values in X_1 are replaced by simulated draws from the posterior predictive distribution of X_1, an important step known as proper imputation. The next variable with missing values, say X_2, is regressed on all the other variables, X_1; X_3; $\ldots, X_k$. Estimation is restricted to individuals with observed X_2 and uses the imputed values of X_1. Again, missing values in X_2 are replaced by

draws from the posterior predictive distribution of X_2. The process is repeated for all other variables with missing values in turn: one such round is called a cycle. To stabilize the results, the procedure is [repeated] to produce a single imputed dataset. (Royston & White, 2011, p. 2)

The posterior predictive distribution refers to the predictive distribution of unobserved scores, conditional on the observed data (Kelly & Smith, 2011). The process begins with each variable with missing values being imputed using a univariate regression model conditional on all of the other variables. The process cycles iteratively through the variables containing missing values until the procedure is stable – a process called regression switching (UCLA Academic Technology Services' Statistical Consulting Group, n.d.b).

Generally, ten to twenty repetitions of this cycle are required to produce an imputed data set. The procedure is repeated m times to yield m imputed data sets (White, Royston & Wood, 2011). Because variables may be of different types (binary, continuous, etc.), a suitable model must be identified for each variable. For example, logistic regression is used to predict a binary variable's values and ordinary least squares (OLS) regression is used if the variable is continuous (Johnson & Young, 2011). ICE has been lauded for its wide-reaching capabilities and different estimation methods depending on the type of variable (e.g., Acock, 2005; White, et al., 2011). In fact, ICE is now available by default in the multiple imputation module in Stata Version 12 (StataCorp, 2011). It is described by Royston (2004, 2005) and Royston and White (2011). White, et al. (2011) provided a tutorial using real and simulated datasets.

The chained equations approach is one of two multiple imputation approaches for handling missingness. In the second approach, called the multivariate normal model approach, the joint distribution of all variables in the imputation model is assumed to be multivariate normal (Little & Rubin, 2002). Information from the variables is used to impute all other variables based on a single model. In contrast, the chained

equation approach is based on each conditional density of a variable given other variables. It is the multivariate normal model approach used in Stata's mi impute mvn command (UCLA Academic Technology Services' Statistical Consulting Group, n.d.b).

In the context of this study, the chained equations approach was implemented because it does not assume a multivariate joint distribution and, therefore, can accommodate variables of different types. It also has lower sample size requirements than the multivariate normal approach (UCLA Academic Technology Services' Statistical Consulting Group, n.d.b).

Data

The data in the tutorial are a subset of the Athena Studies of Resilient Adaptation (AStRA) project, which focuses on the adaptation of immigrant youth living in Greece (Motti-Stefanidi, et al., 2008ab; Motti-Stefanidi & Asendorpf, 2012; Motti-Stefanidi, Asendorpf & Masten, 2012). This subset contained records for 793 youth participants across nine schools in Athens, Greece. Participants either were of Albanian origin (306 or 38.6%) or were native Greek youth (487 or 61.4%).

As shown in Table 3, participants were measured on five outcomes across three annual waves: self-esteem, self-efficacy, behavior problems, school grades, and school engagement – all of which are continuous variables. Also collected were five time-invariant predictors: participants' immigrant status (0 = non-immigrant, 1 = immigrant), sex (0 = male, 1 = female), initial adversity, initial socioeconomic risk, and initial adaptability – the latter three of which were continuously scored. In addition, information for a time-varying variable, adaptability, which was also continuously scored, was collected. Note that in some of the growth models, adaptability was treated as a time-invariant predictor (initial adaptability at Wave 0), whereas, in other analyses it was treated as time-varying predictor (adaptability). Because the interest in the study was in exploring differences between immigrants and their native peers, growth models were stratified by immigrant status. The three waves were coded as 0, 1 and 2, respectively.

Three waves of data were obtained for most of the 793 participants. For 165 participants (20.8%) in two schools, however, data collection stopped after Wave 0; therefore, 68 immigrants (22.2% of 306) and 97 non-immigrants (19.9% of 487) were missing data for Waves 1 and 2.

Because the missingness had to do with administrative reasons and not with the participants themselves, the missingness was treated as MAR – which is an assumption of MI (Schafer & Graham, 2002). Rather than deleting the participants missing Waves 1 and 2 from the sample, MI was performed to fill in the missingness, first, in order to avoid selection bias and, second, because these participants were considered to be a part of the population of interest (J. W. Graham, personal communication, January 14, 2011).

Methodology

Data were analyzed using two linear individual growth models. Although growth models allow for time-unstructured data (different data collection schedules for different individuals) and unbalanced data (different numbers of waves for each individual) (Holt, 2008), if it is suspected that growth curves are non-linear, large amounts of missing data may prohibit departures from linearity – even though intercepts and slopes can still be estimated (Bickel, 2007). Either complete or imputed data are required at higher levels of the analyses (Holt, 2008). An assumption underlying growth analyses is that there is a correctly specified level-one submodel.

Model specification refers to the process of choosing an appropriate functional form for, and variables to include in, the growth models. If the model is not correctly specified, growth models lose their ability to handle missing data well. A growth model's ability to handle incomplete data rests, in part, on the model's being correctly specified (B. D. Zumbo, personal communication, March 18, 2012). If the model is not correctly specified, conclusions may be distorted by the various missing data mechanisms (Zaidman-Zait & Zumbo, 2013). As with any type of analysis, "the nature and number of missing data may badly compromise

Table 3: Variable Descriptions and Names

Variable Description	Outcome Variable Name	Time-Invariant Variable Name	Time-Varying Variable Name
Self-esteem	Esteem		
Self-efficacy	Efficacy		
Behavior	Behavior		
School grades	Grades		
School engagement	Engage		
Immigrant status (0 = non-immigrant, 1 = immigrant)		Immigrant	
Sex (0 = male, 1 = female)		Female	
Initial adversity (Adversity at Wave 0)		Advers0	
Initial socioeconomic risk (SES risk at Wave 0)		SESRisk0	
Initial adaptability (Adaptability at Wave 0)		Adapt0	
Adaptability (Adaptability across Waves 0, 1, 2)			Adapt
Period of data collection			Wave
Participant identification number		ID	

[the] analysis, so that inferences from sample to population become dubious" (Bickel, 2007, p. 301). It is therefore necessary to pay heed to data missingness even when using methods of analysis that otherwise allow for some degree of time-unstructured and unbalanced data.

Models

Each of the growth models in this study was stratified by immigrant/non-immigrant status to allow for comparisons between immigrants and their native peers. Two growth models per outcome (esteem, efficacy, behavior, grades, engage) were run. Model 1 was designed to examine main effects of adaptability on initial levels and rate of change of adaptation, over and above sex, initial adversity, and initial SES risk. Model 2 was designed to examine whether changes in adaptability across the three annual waves were associated with changes in the participants' adaptation, which is known as

dynamic covariation (Long & Pellegrini, 2003; Murray-Close, Ostrov & Crick, 2007).

Model 1, Level 1:

$$Y_{ti} = \pi_{0i} + \pi_1 (Wave)_{ti} + e_{ti}$$

Model 1, Level 2:

$$\pi_{0i} = \beta_{00} + \beta_{01} (Female)_i + \beta_{02} (Adapt0)_i + \beta_{03} (Advers0)_i + \beta_{04} (SESRisk0)_i + r_{0i}$$

and

$$\pi_{1i} = \beta_{10} + \beta_{11} (Adapt0)_i + \beta_{12} (Advers0)_i + \beta_{13} (SESRisk0)_i + r_{1i}$$

Model 2, Level 1:

$$Y_{ti} = \pi_{0i} + \pi_1 (Wave)_{ti} + \pi_2 (Adapt)_{ti} + e_{ti}$$

Model 2, Level 2:

$$\pi_{0i} = \beta_{00} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

and

$$\pi_{2i} = \beta_{20}$$

Tutorial

A step-by-step description of syntax used to perform a multiply-imputed growth modeling procedure is provided in panels 1 - 5. Syntax commands conveniently outlined by UCLA Academic Technology Services' Statistical Consulting Group (n.d.a) for longitudinal data are used as a framework around which to organize the syntax used in this study. Stata command language is identified in bold face.

Conclusion

Due to the well-documented problems associated with missing data, researchers have long been cautioned to investigate data missingness closely and to carefully select a missing data technique that will assist in filling in their data's missing values. Even so, there continues to be uncertainty about how to deal with the problem of incomplete complex data.

There is a paucity of empirical studies centered on this problem, a lack of conclusive recommendations about precisely how to deal with the problem, and limited statistical software resources that allow users to resolve the problem (Allison, 2002). These factors, combined with recent decline in statistics, measurement and research methodology training in North American universities (Aiken, et al., 2008; Aiken, et al., 1990; Merenda, 2003) means that finding a solution to dealing with the problem of incomplete complex data is not an easy task.

This article served as a pedagogical tool for applied longitudinal researchers who are dealing with this problem in their own research contexts. By explaining why complex data pose special challenges with respect to missingness, as well as providing readers with a step-by-step description of syntax created to perform a multiply-imputed individual growth modeling procedure in a real-life longitudinal research

context, it is hoped that readers have a clearer sense of the methodological challenges and realities posed by incomplete complex data.

Once again, it is not claimed that the syntax outlined herein fully takes into account the within-individual dependencies among the study variables; however, the conspicuous lack of information available on this topic means that, from a practical perspective, researchers have little choice but to simply deal with missingness as best they can with available resources (Collins, et al., 2001). After all, it is likely better to fill in missing complex data using a modern missing data technique than it is to do nothing at all.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*(*4*), 1012-1028.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*, 32-50.

PANEL 1

```
set seed 123
```

Data were first entered in person (wide) format. Records for both immigrant and non-immigrant participants were included in the one dataset so that Stata had the fullest information possible about the data prior to imputation.

(1)

A command that set the seed for the random number generator was then written, so that the results of the imputation could be replicated if needed.

```
ice immigrant sex esteem0 esteem1 esteem2 efficacy0 efficacy1
efficacy2 behavior0 behavior1 behavior2 grades0 grades1 grades2
engage0 engage1 engage2 adapt0 adapt1 adapt2 advers0 sesrisk0,
saving(imputed_dataset) m(5) cmd(sesrisk0:regress)
```

The ICE procedure began with this step. All time-varying and time-invariant predictors and outcomes in the two growth models were included in this imputation. To ensure that the imputation model had the most information possible, participants' immigrant flag was also included, rather than running separate imputations for each of the two groups.

Although the within-individual covariation among the variables could not be accounted for perfectly, as previously noted, an attempt was made to partially deal with the collinearity of the repeated measures nested within individuals by including all variables in the analytic models in the imputation model. This decision was made in an effort to account for as much variation as possible within and between individuals. A similar approach was taken by Han (2008).

(2)

With respect to certain segments of this command:
- `m(5)` = the number of imputations

- `saving(imputed_dataset)`= the name for the final outputted data set (containing all five imputation datasets, plus the original data, merged into one master file)

  - `cmd(sesrisk0:regress)=` ICE automatically decides what type a variable is, based on the variable's number of values it takes on. Stata's default was overridden to treat *SESRisk0* as an ordinal variable, so it could instead be treated as continuous.

PANEL 2

```
use imputed_dataset, clear
```

After imputation, ice saved a copy of the new dataset (imputed_dataset) in the current working directory. This command told Stata to open the new file. The new dataset contained all of the variables, plus two new variables: _mi, an identifier for each observation, and _mj, which indicated which imputed data file each row of the data belongs to (0 for the original data, and 1-5, respectively for each of the five new imputed data sets). (3)

```
drop if _mj==0
```

This command instructed Stata to drop the original data that still contained missing values (_mj ==0), keeping only the five newly-imputed data sets. Before running this command, check the descriptive statistics generated for the imputed data sets against the original data. Doing so will ensure that the imputed data indeed have no missing cases and that the descriptive statistics for the each of the variables in the imputed datasets make sense. (4)

```
gen adapt0b = adapt0
```

Because a later step involved restructuring the data from person (wide) format to person-period (long) format, a copy (*adapt0b*) of the initial adaptability variable (*adapt0*) was created.

The adaptation variables were tricky in that they served either as time-invariant or time-varying variables, depending on the growth model. It was therefore necessary to ensure that, during the restructuring, the initial adaptability would be preserved and, in turn, treated as a time-invariant predictor alongside the time-varying adaptability variables. (5)

```
reshape long adapt esteem efficacy behavior grades engage, i(id
_mj)
```

The data were restructured to person-period (long) format, because such a format is required for the growth modeling analyses (described in a later step).

With respect to a certain segment of this command:
- **i(id _mj)**= here, *id* and *_mj* served as our index variables. As UCLA Academic Technology Services' Statistical Consulting Group (n.d.a) notes, "Returning the data to long format has an added complication: we already have [multiple] rows of data for each [participant], one for each of the imputations. As a result, the variable *id* no longer uniquely identifies an observation. However, including both *id* and *_mj* as identifiers will uniquely identify each case." (6)

```
recode _j (3=2) (2=1) (1=0)
```

In restructuring the data to person-period (long) format, Stata automatically assigned the codes 1, 2, and 3 to represent each of the waves of data collection. This command allowed the recoding of waves as 0, 1, and 2, respectively. (7)

PANEL 3

```
rename _j wave
```

In restructuring the data to person-period (long) format, Stata automatically named the wave (8)
variable _j; however, the variable name *wave* was used for this study.

```
* Immigrants
summarize advers0 sesrisk0 if _mj == 1 & immigrant==1
summarize advers0 sesrisk0 if _mj == 2 & immigrant==1
summarize advers0 sesrisk0 if _mj == 3 & immigrant==1
summarize advers0 sesrisk0 if _mj == 4 & immigrant==1
summarize advers0 sesrisk0 if _mj == 5 & immigrant==1
gen ci_advers0 = (advers0 - 5.665472) if _mj==1 & immigrant==1
replace ci_advers0 = (advers0 - 5.552334) if _mj==2 & immigrant==1
replace ci_advers0 = (advers0 - 5.475183) if _mj==3 & immigrant==1
replace ci_advers0 = (advers0 - 5.494476) if _mj==4 & immigrant==1
replace ci_advers0 = (advers0 - 5.399577) if _mj==5 & immigrant==1
gen ci_sesrisk0 = (sesrisk0 - 1.11306) if _mj==1 & immigrant==1
replace ci_sesrisk0=(sesrisk0 - 1.115054) if _mj==2 & immigrant==1
replace ci_sesrisk0=(sesrisk0 - 1.127323) if _mj==3 & immigrant==1
replace ci_sesrisk0=(sesrisk0 - 1.13331) if _mj==4 & immigrant==1
replace ci_sesrisk0=(sesrisk0 - 1.126777) if _mj==5 & immigrant==1

* Non-Immigrants
summarize advers0 sesrisk0 if _mj == 1 & immigrant==0
summarize advers0 sesrisk0 if _mj ==== 2 & immigrant==0
summarize advers0 sesrisk0 if _mj ==3 & immigrant==0
summarize advers0 sesrisk0 if _mj ==4 & immigrant==0
summarize advers0 sesrisk0 if _mj ==5 & immigrant==0
gen cn_advers0 = (advers0 - 4.600149) if _mj==1 & immigrant==0
replace cn_advers0 = (advers0 - 4.554) if _mj==2 & immigrant==0
replace cn_advers0 = (advers0 - 4.492291) if _mj==3 & immigrant==0
replace cn_advers0 = (advers0 - 4.600816) if _mj==4 & immigrant==0
replace cn_advers0 = (advers0 - 4.477417) if _mj==5 & immigrant==0
gen cn_sesrisk0 = (sesrisk0 - .6087394) if _mj==1 & immigrant==0
replace cn_sesrisk0=(sesrisk0 - .6195851) if _mj==2 & immigrant==0
replace cn_sesrisk0=(sesrisk0 - .6162503) if _mj==3 & immigrant==0
replace cn_sesrisk0=(sesrisk0 - .6182393) if _mj==4 & immigrant==0
replace cn_sesrisk0=(sesrisk0 - .6137863) if _mj==5 & immigrant==0
```

(9)

von Hippel (2009) recommended that centering of scores for a given variable be conducted prior to imputation in order to reduce collinearity and improve interpretation of the intercept. What is unclear is which problems, if any, are introduced if the mean that is being subtracted from the given value of a variable is being skewed by the variable's missing values.

For this reason, the scores of the moderating variables (initial adversity and initial SES risk) were grand-mean centered post-imputation, rather than pre-imputation. For brevity, specifics of each line of command in this step are not presented; the commands demonstrate that the respective variables' scores were centered for each immigrant group (x 2) and each imputed data file (x 5), separately.

<u>PANEL 4</u>

```
gen waveadapt0b = wave*adapt0b
gen waveadvers0 = wave*advers0
gen wavesesrisk0 = wave*sesrisk0
```
(10)

Cross-level product terms were computed for subsequent growth modeling (i.e., Model 1).

```
* Model 1 / IMMIGRANTS
mim: xtmixed esteem wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed efficacy wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed behavior wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed grades wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed engage wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==1, || id: wave,
covariance(un) variance

* Model 2 / IMMIGRANTS
mim: xtmixed esteem wave adapt if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed efficacy wave adapt if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed behavior wave adapt if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed grades wave adapt if immigrant==1, || id: wave,
covariance(un) variance
mim: xtmixed engage wave adapt if immigrant==1, || id: wave,
covariance(un) variance

* Model 1 / NON-IMMIGRANTS
mim: xtmixed esteem wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed efficacy wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed behavior wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed grades wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed engage wave female adapt0b advers0 sesrisk0
waveadapt0b waveadvers0 wavesesrisk0 if immigrant==0, || id: wave,
covariance(un) variance
```

(11)

PANEL 5

```
* Model 2 / NON-IMMIGRANTS
mim: xtmixed esteem wave adapt if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed efficacy wave adapt if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed behavior wave adapt if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed grades wave adapt if immigrant==0, || id: wave,
covariance(un) variance
mim: xtmixed engage wave adapt if immigrant==0, || id: wave,
covariance(un) variance
```
(11)

These commands pertain to the two growth models run for each of the outcomes (x 5), stratified by immigrant status (x 2).

With respect to certain segments of this command:
- **mim** = a Stata prefix that pools the results of the five imputed data files
- **xtmixed** = linear mixed-effect module of Stata
- **id: wave** = *id* is the clustering variable; adding *wave* immediately afterwards indicated an associated random effect

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, *45*, 721-734.

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression.* New York, NY: Guilford Press.

Carpenter, J. R., Goldstein, H., Kenward, M. G. (2011). REALCOM-IMPUTE Software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software*, *45*(*5*), 1-14.

Collins, L. M., Schafer, J. L., & Kam, C-M. (2001). A comparison inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(*4*), 330-351.

Diez-Roux, A. V. (2002). A glossary for multilevel analysis. *Journal of Epidemiology and Community Health*, *56*, 588-594.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.

Han, W. J. (2008). The academic trajectories of children of immigrants and their school environments. *Developmental Psychology*, *44*(*6*), 1572-1590.

Holt, J. K. (2008). Modeling growth using multilevel and alternative approaches. In *Multilevel Analysis of Educational Data. Volume 3 of the Quantitative Methods in Education and the Behavioral Sciences: Issues, Research and Teaching Series*, A. A. O'Connell & D. B. McCoach (Eds.), 111-159. Charlotte, NC: Information Age Publishing.

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(*4*), 1195-1199.

Johnson, D. R., & Young, R. (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family*, *73*, 926-945.

Kelly, D., & Smith, C. (2011). *Bayesian Inference for Probabilistic Risk Assessment: A Practitioner's Guidebook.* London: Springer-Verlag.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data, 2nd Ed.* New York, NY: Wiley.

Long, J. D., & Pellegrini, A. D. (2003). Studying change in dominance and bullying with linear mixed models. *School Psychology Review*, *32*, 401-417.

Merenda, P. F. (2003). Measurements in the future: Beyond the 20th century. *Psychological Reports*, *92*, 209-217.

Motti-Stefanidi, F., & Asendorpf, J. B. (2012). Perceived discrimination of immigrant youth living in Greece: How does group discrimination translate into personal discrimination? [Special Issue], *European Psychologist*, *17*(*2*), 93-104.

Motti-Stefanidi, F., Pavlopoulos, V., Obradović, J., Dalla, M., Takis, N., Papathanasiou, A., & Masten, A. (2008a). Immigration as a risk factor for adolescent adaptation in Greek urban schools. *European Journal of Developmental Psychology*, *5*(*2*), 235-261.

Motti-Stefanidi, F., Pavlopoulos, V., Obradović, J., & Masten, A. S. (2008b). Acculturation and adaptation of immigrant adolescents in Greek urban schools. *International Journal of Psychology*, *43*(*1*), 45-58.

Motti-Stefanidi, F., Asendorpf, J. B., & Masten, A. S. (2012). The adaptation and psychological well-being of adolescent immigrants in Greek schools: A multilevel, longitudinal study of risks and resources. [Special Issue], *Development and Psychopathology. 24*(*2*), 451-473.

Murray-Close, D., Ostrov, J. M., Crick, N. R. (2007). A short-term longitudinal study of growth of relational aggression during middle childhood: Associations with gender, friendship intimacy, and internalizing problems. *Development and Psychopathology*, *19*, 187-203.

Obradović, J., Lloyd, J. E. V., & Motti-Stefanidi, F. (in preparation). Adaptation of immigrant and non-immigrant youth living in Greece: The role of family adaptability.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd Ed.)*. Newbury Park, CA: Sage.

Royston, P. (2004). Multiple imputation of missing values: update. *Stata Journal, 5*(2), 188-201.

Royston, P. (2005). Multiple imputation of missing values: Update of ICE. *Stata Journal 5*(*4*), 527-536.

Royston P., & White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*, *45*(*4*), 1-20.

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

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3-15.

Schafer, J. L. (2001). Multiple imputation with PAN. In *New methods for the analysis of change*, A. G. Sayer and L. M. Collins (Eds.), 355-377. Washington, DC: American Psychological Association.

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

Schafer, J. L., Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*, 437-457.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

StataCorp. (2009). Stata Statistical Software: Release 11. College Station, TX: StataCorp LP.

StataCorp. (2011). Stata statistical software: Release 12. College Station, TX: StataCorp LP.

UCLA Academic Technology Services' Statistical Consulting Group (n.d.a). Stata FAQ: How can I perform multiple imputation on longitudinal data using ICE? Retrieved March 12, 2012, from http://www.ats.ucla.edu/stat/stata/faq/mi_longitudinal.htm.

UCLA Academic Technology Services' Statistical Consulting Group (n.d.b). Stata library: Multiple imputation using ICE introduction. Retrieved March 12, 2012, from http://www.ats.ucla.edu/stat/stata/library/ice.htm

von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply-imputed data. *Sociological Methodology*, *37*(*1*), 83-117.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265-291.

White, I.R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377-399

Yuan, Y. (2011). Multiple imputation using SAS software. *Journal of Statistical Software*, *45*(*6*), 1-25.

Zaidman-Zait, A., & Zumbo, B. D. (2013). Can multilevel (HLM) models of change over time adequately handle missing data? *Journal of Educational Research and Policy Studies*, *13*(*1*), 18-31.