Fall 10-19-2012

Multiple Imputation and Higher Education Research

Catherine A. Manly
University of Massachusetts, Amherst, camanly@gmail.com

Ryan S. Wells
University of Massachusetts, Amherst, rswells@educ.umass.edu

MULTIPLE IMPUTATION HIGHER EDUCATION

Multiple imputation and higher education research

Catherine A. Manly and Ryan S. Wells

University of Massachusetts, Amherst

A paper prepared for presentation at the annual meeting of the

Northeastern Educational Research Association

Rocky Hill, CT

October 18, 2012

Keywords: multiple imputation, survey research, missing data, higher education

Abstract

Higher education researchers using survey data often face decisions about handling missing data. Multiple imputation (MI) is considered by many statisticians to be the most appropriate technique for addressing missing data in many circumstances. However, our content analysis of a decade of higher education research literature reveals that the field has yet to make substantial use of this technique despite common employment of quantitative analysis, and that many recommended MI reporting practices are not being followed. We conclude that additional information about the technique and recommended reporting practices may help improve the quality of the research involving missing data. In an attempt to address this issue, we offer an annotated practical example focusing on decision points researchers often face.

Multiple imputation and higher education research

Introduction

Higher education researchers using large survey datasets frequently face decisions about how to handle missing data, which may influence their results and conclusions. The presence of missing data can potentially affect both the validity and reliability of research findings, and lead to problems with the generalizability of results (see McKnight et al (2007) for a clear elaboration of the potential consequences of missing data). While listwise (or casewise) deletion has often been higher education researchers' approach for addressing missing data in the past, newer statistical techniques have eclipsed traditional methods of handling missing data in appropriateness for most circumstances (Croninger & Douglas, 2005; Donders, van der Heijden, Stijnen, & Moons, 2006; Schafer & Graham, 2002).

In many practical social science research situations, the multiple imputation (MI) technique introduced by Rubin (1987) is considered the "gold standard" for handling missing data (Treiman, 2009, p. 185). The phrase "multiple imputation" refers to a process of proceeding reasonably with statistical analyses given the uncertainty caused by the presence of missing data. Despite entailing a relatively complicated array of steps, the process is approaching the realm of regular use by non-statistically-oriented researchers and is now implemented in most common statistical software packages (e.g. R, SAS, Splus, SPSS, and Stata) in addition to several special packages (e.g. Amelia and SOLAS).

Interested researchers should understand that MI is a "state of the art" technique (Schafer & Graham, 2002). This status means that it is constantly being revised and expanding into new areas, and there is uncertainty associated with aspects of the technique still in development or lacking consensus among statistical experts. This means that as researchers, we face subjective

decisions based on our data, and guidance for such questions has been scattered or slowly developed. As a result, while the technique has made significant inroads in certain disciplines such as health (Schafer, 1999), psychology (Graham, 2009), and biostatistics (White, Royston, & Wood, 2011) (also see the list of published research using MI by van Buuren and Groothuis-Oudshoorn (2011)), it is still being minimally implemented in many fields, including ones that use large data sets (Stuart, Azur, Frangakis, & Leaf, 2009), particularly in education (Peugh & Enders, 2004). The purpose of our paper is to understand the current state of MI in higher education research and based on that understanding, to recommend practical ways to improve its use and reporting grounded in the current state of MI development.

We achieve this purpose by: a) generating an aggregated list of recommended practices for reporting based on a review of the methodological MI literature, b) analyzing the content of published higher education research to examine how recommended practices are being used, and c) presenting an example analysis demonstrating recommended MI practices on a higher education topic, while providing a non-statistically-oriented discussion of some of the practical choices to be made when using MI. In doing so, we attempt to facilitate the use of the MI technique by researchers who are interested in following the growing recommendation within the statistical community of using MI and to improve the state of higher education research.

Why Is It Okay To Use Multiple Imputation?

As a prelude, we feel it important to address a fundamental concern we have heard in discussions with colleagues. Even experienced quantitative researchers in higher education can be highly skeptical of the advisability of using a technique like multiple imputation. It smacks too much of "making up data" for some people's comfort, and many simply dismiss it since it does not solely use data collected by the researcher (even though it uses *information* collected by

the researcher, a key distinction to which we will return). Thus, we begin here with a subjective decision point facing experienced skeptics and/or researchers new to MI: whether to continue using listwise deletion (or another traditional method) or learn about MI and implement it if appropriate. If one is not convinced of the merits of the MI approach then what follows in this article will be irrelevant and ineffective in its purpose.

There are several well-known problems fundamental to the use of listwise deletion (Acock, 2005; Ludtke, Robitzsch, Trautwein, & Koller, 2007; Peugh & Enders, 2004; Schafer & Graham, 2002). For researchers using large survey datasets, the most relevant of these is that often they will not know whether there are any systematic reasons why particular data are missing, and this means they do not know whether their statistical estimates are biased, potentially causing incorrect conclusions. Thus, the typical assumption that data are missing completely at random (MCAR) and that dropping cases will not affect the validity of results, may be convenient, but may not be correct.[1]

MI has been shown to be superior to listwise deletion in almost all circumstances (Allison, 2002), except possibly in cases with very small amounts of missing data and data that is missing completely at random. Even with such data, however, the loss of power that comes with dropping cases may still be problematic, and Graham (2009) argues that MI could always be reasonably used (even with <5% missing data). The bigger question for skeptical researchers is

[1] There are numerous clear explanations of the possible types of missing values, traditionally referred to using Rubin's nomenclature of MCAR, MAR, and MNAR (McKnight et al., 2007; Nakagawa & Freckleton, 2011; Peugh & Enders, 2004; Schafer & Graham, 2002). Essentially, MCAR data are a random sample where any missing data are due to random reasons related neither to observed nor unobserved variables. With data that are MAR, the missing data are either random or related to observed variables, and so will essentially be MCAR once a researcher controls for those observed variables in a statistical analysis.

whether multiple imputation actually offers a reasonable alternative to the known difficulties with listwise deletion and other traditional techniques.

Understanding appropriate and inappropriate ways to think about what is happening when data is multiply imputed may help those in doubt. When using MI, you are fundamentally considering the distribution of values in the entire population for certain variables. You are not considering the individual value of a variable for a particular person in a sample. This would indeed be "making up data," and is expressly not the point of multiple imputation. Schafer (1999) describes MI as "a device for representing missing-data uncertainty. Information is not being invented with MI any more than with… other well accepted likelihood-based methods, which average over a predictive distribution" (p. 8).

Recall that each variable in a study has some underlying distribution of expected values within the population as a whole, and the randomly sampled individuals in a survey sample are chosen in order to be representative of that population. Each variable measured has an associated underlying distribution function that represents the probability of particular possible values (with the probabilities across the distribution summing to 1 for a given variable). These variables may be correlated with each other, which means that knowing something about one variable gives you information about another correlate. Multiple imputation is using the fundamental distributional property of the measured variables and their correlations to produce reasonable average estimates of statistical inferences based on the collected information. Thus, it is not necessary to drop cases, thereby losing known correlation information (and the corresponding potentially substantial investment of resources put into collecting the data). Instead, all of the information collected is used to estimate statistical inferences addressing questions you have about relationships shown by the data.

In other words, the general point of MI is not to produce particular values for the data. The point is to produce valid statistical inferences. Using all of the information about your variables actually helps in producing better statistical results than other methods like listwise deletion, given the uncertainty that exists because of the missing data. This is because MI produces smaller standard errors and less bias than other typical methods. Put another way, if you know something about the missing data for a particular variable because of that variable's correlation with other variables in a model, why would you throw out that information and therefore generate less accurate statistical inferences?

It may be the case that analyses using MI and a traditional method like listwise deletion produce similar results. In this situation, it may be fine to conduct and report study results using listwise deletion. However, researchers are advised to learn how to perform MI analyses in order to be able to carry out this comparison and identify situations where results differ between methods of handling missing data. Such discrepancies should be of interest not only to researchers but also to consumers of the research. With this introductory grounding in MI, we now turn to our investigation of this technique in higher education research.

Conceptual Framework

Our study of the higher education research literature is framed conceptually by the ideas of Silverman (1987) concerning the purposes and functions of higher education journals. He proposed, "journals both create and mirror their fields" (p. 40). By studying what has been published in our premier journals, we can learn what is currently happening in the field of higher education research, as well as what foundation is being laid for future research in the field. Knowing what methods are being used, and how they are implemented and reported, is an integral aspect of understanding "how we know what we know" (Silverman, 1987, p. 40). In our

case, we are interested in knowing how the field of higher education handles missing data, specifically if and how the field implements MI, and how this use compares to currently recommended practices in this area.

Methods

We begin by reviewing the seminal and current literature concerning multiple imputation, including both statistically-focused and discipline-oriented literature. Several recent accounts present introductions targeted toward MI users that also include reporting recommendations. We review this literature to identify recommended reporting practices, which form the basis for the categories of our content analysis coding, as well as to understand the steps users should take and the decisions that must be made in order to inform our practical example.

A content analysis (Berelson, 1952; Huckin, 2004) of the higher education literature was then employed to reveal the use (or lack) of recommended MI practices in higher education research. Our study corpus includes literature in four of the most prestigious higher education journals: *Journal of College Student Development, Journal of Higher Education, Research in Higher Education,* and *Review of Higher Education* (Bray & Major, 2011). However, this content is limited by its orientation toward researchers and by solely representing the field of higher education. To expand our content analysis to include journal articles that are more oriented toward higher education practitioners, we purposefully selected three additional higher education journals for our analysis that are used relatively frequently (Bray & Major, 2011) and which span various areas of professional higher education practice: *New Directions for Institutional Research, Journal of Student Affairs Research and Practice* (formerly *NASPA Journal*), and *Community College Journal of Research and Practice.*

In a broader preliminary search of the literature, it was apparent that disciplinary journals would be a necessary component of our analysis. Specifically, sociology journals were often addressing higher education issues and utilizing MI. To include this perspective in our analysis, we selected three prominent sociology journals that frequently include higher education issues in their content: *Sociology of Education, Social Forces,* and *American Sociological Review.* Given that very little research, especially in education (Peugh & Enders, 2004), was using MI over 10 years ago, we limit our review of these ten journals to the years 2001-2011.

After our journals were selected and articles using MI to study higher education issues were identified, we coded the articles using the recommended practice categories generated from our review of the methodological literature. Each article's content was analyzed for the presence of ten possible recommended reporting practices. In cases where subjective judgment was needed, we erred on the side of generosity. For example, when a researcher reported that "several" imputations were used in the MI procedure, we still coded this as reporting the number of imputations. When an author gave the rate of missing data on some variables, but not all, we still coded that article as having reported rates of missing data.

In some cases where items in our coding schema were not explicitly addressed by the authors, it was straightforward to determine what they had done implicitly. For example, when an author did not explicitly state the software that was used, but included a citation to an article about a specific command and/or software package, we coded this article as meeting the criteria for reporting software and algorithm/procedure information. Overall, our results should be a liberal estimate of the use recommended reporting practices.

Following this content analysis, we present a discussion of performing a secondary data analysis using MI along with using the recommended reporting practices. Using data from the

Educational Longitudinal Study (ELS:2002-2006) we examine how receiving undergraduate loans in a financial aid package is related to persistence at a student's first postsecondary institution via logistic regression. We use this primarily as a vehicle for discussing the imputation procedure and how to analyze multiply imputed datasets rather than discussing the empirical results of this particular analysis. In the appendices, annotated Stata code for this example walks a reader through the MI process, highlighting the recommended reporting practices as well as key decision-points.

Results

Review of Literature on Reporting the Multiple Imputation Technique

Reporting in peer-reviewed literature concerning how researchers address missing data has historically been sparse in education, although it has increased over time (Peugh & Enders, 2004). In 1999, the American Psychological Association argued that techniques for addressing missing data ought to be reported, although they did not provide explicit guidelines (Wilkinson & Task Force on Statistical Inference, 1999). Currently, there has been no definitive statement of recommended reporting practices for MI, and while suggestions do exist, van Buuren and Groothuis-Oudshoorn (2011) claim, "We need guidelines on how to report MI" (p. 55)

Such guidelines for reporting are scattered in a number of places. Several missing data statistical experts have provided reporting suggestions as part of their descriptions of the MI process (Allison, 2002; Enders, 2010; Little & Rubin, 2002), and a recent basic introduction to missing data issues targeted toward researchers who are less statistically-oriented also included reporting recommendations (McKnight et al., 2007). The discipline-oriented literature has a number of examples where reporting recommendations have been included: Burton and Altman (2004) in cancer research, Graham (2009) as well as Jelicic, Phelps, and Lerner (2009) in

psychology, Peugh and Enders (2004) in education, Sterne, et al. (2009) as well as Klebanoff and Cole (2008) in epidemiology.

We synthesized the recommendations presented in these sources to develop the list of recommended reporting practices presented in Table 1. We selected items for our list of recommended reporting practices which met two criteria: 1) they are intended to allow readers of a paper to understand how the results were obtained, and 2) they are intended to allow replication of results by others. In generating this list, we also aimed to balance considerations of clarity in reporting with conserving journal space since, as Enders (2010) recognizes, journal editors must agree to publish enough detail so that the key decision points are communicated. In addition to providing guidance for future researchers, these recommendations also formed the basis for the categories of our content analysis coding.

Content Analysis of Higher Education Research Literature Using Multiple Imputation

Our search of the 10 specified journals led us to 34 articles addressing higher education issues and utilizing MI (see Appendix A). We did not include articles that were instructional in nature rather than a true empirical study that employed MI as part of a statistical analysis to address a research question. We also did not include articles that referred to MI only as a check of the robustness of results from other ways of handling missing data, without actually presenting the MI results. The journal that contains the most articles of this type is *Sociology of Education*, representing 10 of the 34 articles. *Research in Higher Education* contained seven articles, *Social Forces* had five, and *Journal of Higher Education* had four.

The results of our content analysis reveal that higher education research is frequently not meeting either the standards of use or reporting that the most current methodological literature suggests. The recommended information that was most commonly reported was the number of

imputations used; 22 of 34 articles reported this. However, it was clear that most of the literature relied on the somewhat outdated notion that three to five imputations is recommended. (See the section on reporting the number of imputations later in this paper for more about updates to this common misconception.) The second-most commonly used reporting practice was to provide information about the software and/or commands used; 18 articles reported this.

Most of the recommended reporting practices were infrequently used. No articles compared observed versus imputed values, though six did commendably report a comparison of results using MI versus other methods, such as listwise deletion. Two articles reported special considerations such as MI in relation to longitudinal or multi-level data. Four articles reported at least some variables used in the imputation phase (though without mention of whether auxiliary variables were used in imputing) and another four reported the overall percent of missing data.

Of the 34 articles, 15 of them reported two of the 10 possible recommended reporting practices. Six articles reported three practices, four articles reported four of the practices, another four articles reported only one practice, and one article reported five of the 10 practices. Three articles only reported using MI, but did not include any of the recommended reporting practices.

The outlier of the group reported eight of the 10 recommended reporting MI practices (Alon & Tienda, 2007). This article is the best example in our study of how to report MI when publishing higher education research. Interestingly, however, the practices were all reported in an online supplement, and not in the main text of the article itself. This likely reveals a trade-off between reporting all of the recommended information about MI and using limited space in journals for this purpose, and may reveal the need for more online supplemental options to convey information necessary for readers to make sense of, and/or replicate, research.

Having identified that little higher education research literature reports fully about MI, we now turn to a discussion of using multiple imputation that highlights how to implement the recommended reporting practices as well as other decisions that must be made when using MI.

Discussion of an Illustrative Analysis

In the ELS data for our example, a nationally representative sample of high school seniors was surveyed by NCES in 2004, and then surveyed again in 2006 to find out about their postsecondary experiences. We look at the subpopulation of students who attended postsecondary education immediately after high school and who received financial aid, asking whether being offered loans as part of a financial aid package by a student's first institution influenced persistence at that institution within the first two years. We found no relationship between loans and persistence at a student's first institution, controlling for a host of other variables in the model, although the specifics of this result are not of interest here.

There are several methods for conducting MI, and so before discussing the choices we made during our illustrative analysis, we wish to orient readers to the method we use here. Within MI, there are two widely implemented methods for imputing–multiple imputation by chained equations (MICE or fully conditional specification, FCS) and multivariate normal imputation (MVN). Both produce reasonable results for categorical variables (Lee & Carlin, 2010). However, the MVN method assumes multivariate normality among variables, an assumption that does not hold for the binary, nominal, and ordinal variables common in social science survey research (Social Science Computing Cooperative, 2012), and so we will focus our discussion on implementing the popular MICE/FCS technique which has been shown to produce good results for these kinds of data (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). However, we will identify places where considerations for the chained and MVN methods differ.

We must also note that full information maximum likelihood (FIML), another newer technique that is theoretically an excellent choice for handling missing data, is practically difficult to implement unless you use structural equation modeling (SEM), and it is particularly problematic for situations involving complex survey data (Heeringa, West, & Berglund, 2010), including many datasets from the National Center for Education Statistics (NCES). Researchers interested in the maximum likelihood method should look at introductions to this technique (Allison, 2002; Baraldi & Enders, 2010; Enders, 2006, 2010), as we focus here on the MI technique, which is more commonly implemented in survey research.

We intend our example to provide practical guidance for researchers who wish to use MI by chained equations. For researchers just encountering MI, several good tutorials and introductions are available, both for those who wish to use MI by chained equations (Royston & White, 2011; Social Science Computing Cooperative, 2012; van Buuren & Groothuis-Oudshoorn, 2011), and for those who wish to use MI under the multivariate normal assumption (Allison, 2002; Enders, 2010; Graham, 2009; McKnight et al., 2007). We do not attempt to replicate these introductions. Instead, we focus our discussion on the implementation of the recommended reporting practices identified in Table 1 and the subjective decision points that researchers face when implementing MI, as well as on several issues specific to secondary analysis of large datasets. By doing so, we hope to aid MI novices in understanding how to report this complex process and moderately experienced MI-users in improving their practice.

Data Preparation

We will not linger on issues of data preparation, except for a few notes germane to MI and to understanding the Stata code for our example of data preparation in Appendix B. The data for our example are from the publicly released version of the Education Longitudinal Study

(ELS:2002-2006) from the NCES, which can be downloaded using the EDAT tool at http://nces.ed.gov/surveys/els2002/.

While several software programs now offer MI as part of their statistical packages, our example uses Stata. If Stata programming concepts such as loops to repeat commands and local macros to identify variables or text phrases[2] are new to you, we recommend Scott Long's (2009) book about improving workflow using Stata.

Subjective decision–do you decode prior imputations from other methods? NCES imputes single values for missing data for several commonly used variables such as gender and race. Since multiple imputation is a preferable method of handling missing data, for our example, we choose to decode the NCES imputations (NCES provides variables identifying imputed values) so that the missing data for gender and race can be multiply imputed. This could also be done for the socioeconomic status (SES) variable, which NCES also imputed, but since SES is composed of 20 different variables, we felt that doing so would unnecessarily complicate our example (and would increase the time for imputation to an unreasonable length–more about this issue later), and so we use the NCES-imputed SES variable.

Researchers implementing multiple imputation while using large publicly available data collected by other agencies or organizations will similarly face a subjective decision about whether to use imputations generated by those agencies. In general, unless the dataset was multiply imputed by the agency (as is done with NHANES data in the medical community), it is preferable to use multiple imputation unless this becomes impractical (as with the composite SES variable here).

---

[2] For example, Stata's '*foreach* { … }' command allows the commands within the braces to be repeated, and the command '*local* date "2012-10-18"' sets up a local macro, later referred to in the code as `date', and which Stata then converts to 2012-10-18).

Subjective decision–do you impute component variables or whole scales? This is another situation where reasonable people differ, and the answer may often depend on practical circumstances rather than theoretical considerations. It may be theoretically good to impute the individual components of a scale and then combine them after imputation (Ho, Silva, & Hogg, 2001; Schafer & Graham, 2002; Van Ginkel, 2010) However, practical guidelines for doing this are sparse (Enders, 2010), and many researchers who use techniques like principal components analysis do not report about missing data (Jackson, Gillaspy Jr., & Purc-Stephenson, 2009).

Graham (2009) provides some guidance about how to tell whether it is reasonable to impute at the scale level, but it may not be practical in many circumstances until computing power improves significantly. Scales are composed of multiple variables, each of which may require several dummy variables in the imputation. The number of variables added to the imputation model can quickly balloon, causing a corresponding balloon in the time required to impute, and making including the individual components of the scale impractical. For the purpose of our example, we have chosen to conduct a principal components analysis for two variables in the data preparation stage, imputing the scale variables rather than their component variables, as this seems to be the most practical solution for many circumstances. However, this appears to be an area that could benefit from further research and practical guidance.

Data Imputation

While we leave the heavy lifting of explaining the chained equations method and how to implement it to others (Royston & White, 2011; Social Science Computing Cooperative, 2012; van Buuren & Groothuis-Oudshoorn, 2011), we do wish to suggest how novice or intermediate MI users might think conceptually about the method. In essence, in the first iteration, Stata orders the variables in sequence from the least to the most amount of missing data and then

conducts an initial imputation (using the monotone imputation method) to get starting values for all missing data. Then in the second iteration, Stata looks at each variable in turn and uses the model specification you provided for that variable (hence the alternative term "fully conditional specification") to impute new values for the missing data using the imputed values from the previous iteration. This process repeats itself for some number of iterations (e.g. Stata's default is 10), which should converge such that he values produced by the imputation process settle into a random pattern with a reasonable amount of error. Then an imputed dataset is captured. The process then begins all over again, with Stata storing the number of imputed datasets you specified (the number of imputations, *m*). While the full process has more complexity, this outlines Stata's basic process. We will now turn to consideration of some specific issues that arise in practice and our recommendations for reporting during the data imputation phase.

Recommendation–Report rates of missing data. This includes reporting both the overall percentage of missing data, and the range of missing data rates across all variables, which involves checking the percentage of the missing cases in each variable that is to be imputed (see Appendix C). As general guidelines, the imputation results will be best if there is less than 10% missing data, and be very cautious about imputing any variables with over 50% missing data unless you know why[3] or unless you know that the uncertainty resulting from this missing data is small[4] (Barzi & Woodward, 2004; Royston, 2004).

---

[3] For example, we run over 50% missing data with our academic (57%) and social (59%) integration scale variables because not all of the cases with missing data will actually be part of our analysis subpopulation since not everyone who went to high school in 2004 actually attended postsecondary education by 2006. However, we cannot give individuals who have never gone to postsecondary education an actual postsecondary integration score without inappropriately affecting the range of imputed values.

[4] In order to check the impact of uncertainty from missing data, see the discussion of "missing information" in note 6 and the code for evaluating this in Appendix D (McKnight et al., 2007).

Recommendation–Report variables used in the imputation phase. In general, researchers want to identify all variables used for imputation. This will typically include all variables that the researcher intends to be part of the final analytical model used to investigate the main research question(s). Researchers may also wish to include variables that are highly correlated with missingness on variables to be imputed; such "extra" variables are called auxiliary variables (Enders, 2010).

Recommendation–Communicate the algorithm/procedure. Researchers should communicate how the imputation models were set up, and much of the relevant information can be communicated by specifying the basic software command used and any key options changed from their defaults. For example, if one chooses to do more than Stata's default of 10 burn-in iterations[5], this ought to be communicated (Enders, 2010). It is also sometimes possible to provide a software-specific citation that indicates the method chosen to implement, although even in this situation the researcher ought to pay attention to any non-default choices that were made. In any case, it is good practice to include a relevant citation for the procedure since there are several versions of MI that use different algorithms.

Subjective decision–Convergence. Convergence of imputed values ought to be checked with either the chained equations or the multivariate normal approach. MI using the multivariate normal assumption (e.g. Stata's *mi impute mvn* command) has been proven theoretically to converge (Allison, 2002; Enders, 2010), although whether convergence has been achieved in

[5] The number of burn-in iterations refers to the number of times the imputation process is iterated prior to actually saving the first complete dataset to memory (e.g. saving a dataset as *m*=1). For the multivariate normal (MVN) MI method, the researcher also may decide to select a different number of between-imputation iterations, which refers to the number of times the imputation process is iterated between saving one complete dataset to memory and the next (e.g. saving a dataset as *m*=2), and this convergence aspect should also be investigated (Enders, 2010). See the next section on convergence for information about evaluating this.

practice for a given number of imputations (*m*) should be assessed. While there is no equivalent theoretical justification for why convergence of the chained equations method should be achieved (e.g. Stata's *mi impute chained* command), the procedure has been shown to work well in practice (van Buuren et al., 2006). This means researchers using either MI method should investigate convergence, and the easiest way to do this is graphically (see Appendix C). One may decide after looking at plots of imputation results across iterations for different variables to alter the default number of burn-in iterations (e.g. the default for Stata is 10). For our example, we chose to iterate for a somewhat conservative burn-in of 30 times after evaluating the results of a chain with 50 burn-in iterations.

Complex survey design imputation considerations. There are several special considerations relevant for researchers using complex survey design. Heeringa et al. (2010) gives guidance for MI analysts using complex survey data in Stata. Stata 12 has an option to include weights in MI commands. Also, Azur, Stuart, Frangakis, and Leaf (2011) recommend including the primary sampling unit as a model predictor, which we have done in our example. After imputation, the dataset must be set up for complex survey design in Stata using the *mi svyset* command (see Appendix C).

Recommendation–Report the number of imputations. The traditional wisdom about the number of imputations to choose, based on the concept of efficiency[6] in Rubin's (1987) original work, was that around five imputations (*m*) was typically sufficient. More recently, Graham, Olchowski, and Gilreath (2007) argued that researchers should consider more

[6] The idea of efficiency is based on the amount of "missing information" in your data, a concept that is clearly explained by McKnight et al. (2007). It gives a measure of the influence of missing data on statistical results. To see how to view the rate of missing information (typically denoted by γ) in Stata, see Appendix D. If the fraction of missing information for variables is high (greater than 50%), then one should consider doing more imputations (since this rate is related to the number of imputations).

imputations (e.g. perhaps m=20 or m=40) in order to improve the power of their analysis. White, Royston, and Wood (2011) provided a very practical and helpful "rule of thumb that m should be at least equal to the percentage of incomplete cases" (p. 388) based on the desire to produce results that can be reproduced across repeated imputations, even with different starting random number seeds (which allows for exact duplication of results).

How long will it take to impute? Figuring out how long an imputation ought to take can be a helpful sanity check, since the time to impute can be a practical constraint on the number of imputations chosen. The estimation of imputation time is a combination of art and logic and depends on numerous factors, including the computer's processing capacity, the number of variables in the overall model specification, the types of models used for different variables (e.g. multinomial logistic regression takes noticeably longer than ordinary least squares regression or logistic regression (White et al., 2011)), and the number of iterations and imputations chosen. Don't wait days for an initial imputation attempt to complete. It is rare to specify an imputation model the first time without needing modification. When making model adjustments before developing a final model, the researcher will want trials to be short.

This leads us to answer the question of time to impute with a practical strategy. After using Stata's dryrun option to ensure that the command is structured correctly, count up the number of imputation model variables (including categorical dummy variables). Treiman (2009) suggests that adding variables to a model increases the imputation time faster than an arithmetic increase, finding that "approximately doubling the number of variables to be imputed increased the time by a factor of four" (p. 186). While we have read that models might practically go as high as 100 variables before imploding, we have encountered problems with more than 50 to 70 variables (particularly if most are binary/categorical variables).

Now set the number of imputations (what we call "*nummi*" in our Appendix example) to *m*=2 and impute, using timing code (see Appendix C. Debug the code and get the model right. Double the number of imputations to *m*=4, and impute again to check the time. Imputing time does not increase entirely linearly as m increases, but that can be a rough approximation of the order of magnitude of how long processing more imputations for a final analysis might take. Make any model adjustments needed, and run the imputation (perhaps overnight) with more imputations (perhaps *m*=20). If you now choose to set *m* even higher (e.g. *m*=75 or *m*=100), you are likely waiting for your final results instead of an error message.

Data Analysis

Pooling statistical analysis results. After the data imputation phase is complete, a researcher has multiple complete datasets and wants to conduct statistical analyses. Since the data comprising each imputation could be viewed as a complete dataset, each imputation can be analyzed using typical complete case methods (regression, logistic regression, etc.). These results can then be averaged, with the parameter standard errors being combined using "Rubin's rules" which incorporate both the within-imputation variance and the between-imputation variance (with an adjustment). McKnight et al. (2007) have a clear, step-by-step explanation of this pooling process. In Stata, this pooling can be accomplished with the *mi estimate* command.

Complex survey design analysis considerations. Sometimes a researcher does not want to conduct an analysis on the full survey sample. In this situation, it is preferable to identify a subpopulation for analysis rather than dropping cases (Heeringa et al., 2010). In Stata, this can be accomplished for multiply imputed datasets by using the regular Stata command for specifying survey subpopulations (*svy, subpop(:*) in conjunction with the *mi estimate* command to pool results and analyze only a specific set of cases if that is desired (see Appendix D).

Small sample adjustment. There is a "small sample" method developed by Barnard and Rubin (1999) for determining degrees of freedom (and thus confidence intervals) for analyses such as logistic regression on multiply imputed datasets. However, according to Heeringa et al. (2010) this method has also been shown to produce good results for large sample sizes as well, and so we conclude that this adjustment should usually be used. It is the default option in Stata, but an analyst needs to know not to turn it off even if one is working with a large sample.

Subjective decision–what fit statistics should I check/report? More and better guidelines are needed for fit statistics that are clearly presented. White, Royston and Wood (2011) indicate that statistics such as the likelihood ratio test statistic, model chi-squared statistic, and goodness-of-fit test statistic cannot be combined using Rubin's rules. Enders (2010) basically says there is no good choice for this yet, but suggests three possible multiparameter significance tests: D_1, which resembles a Wald statistic, but whose trustworthiness Enders says has not yet been tested "in realistic research scenarios" (p. 236); D_2, which pools Wald tests but which Enders says may not be trustworthy; and D_3, which pools likelihood ratio tests (but which White, Royston and Wood appear to say is not appropriate).

Our basic understanding of the fundamental problem is that MI approximates a model for each parameter separately while typical fit measures do simultaneous test of multiple parameters, and thus typical fit measures are not meaningful. However, journal editors may require that fit statistics be reported anyway, and it is not clear to us whether some researchers simply report the average (across imputed datasets) of typical fit tests like BICs anyway despite this not being technically appropriate. We have not yet found clarification of this issue, so it appears to be an area for future research.

Recommended practice–Describe any notable imputation results. It is good practice to compare observed and imputed values, particularly for variables with high rates of missing data. Tabulating values for the original data (imputation m=0 in Stata) and imputed values (m>0) is a straightforward way of comparing values. In addition to tabulations, graphical methods can be helpful. van Buuren (along with colleagues) has offered several nice visual data comparison methods. In one paper (van Buuren et al., 2006), a histogram approach for comparing data is shown that might be adapted to show original and imputed data distributions with different bar darkness. A different paper (van Buuren & Groothuis-Oudshoorn, 2011) shows several possible methods of viewing differences, for example with observed data in blue and imputed data in red plotted separately for each imputation. White et al. (2011) offer another type of visual comparison, using boxplots of all imputations (where m=0 is the original data, and m>0 shows each imputed dataset).

Finally, a researcher might decide to investigate statistical results as determined under different approaches to handling missing data, perhaps comparing results obtained via listwise deletion to those obtained via MI (see Appendix D). If these different approaches produce discrepancies in results that would affect interpretation, they should be discussed.

Conclusions

We conclude that higher education research is using multiple imputation infrequently given the field's common use of quantitative research (Hutchinson & Lovell, 2004; Wells et al., 2012). Higher education research using MI typically does not follow most of the recommended reporting practices we identified. If higher education journal content is mean to "both create and mirror" (Silverman, 1987, p. 40) the field, these results suggest that the field could benefit from suggestions for improvement. As a mirror, these results reflect a slow adoption of current

techniques and practices related to missing data. As an influence on the creation of our field, adopting the recommended practices will not only improve the content of journals, but will also allow for readers to gain a better understanding of the techniques and to be able to replicate studies. Our findings may also reveal a need for more advanced statistical training for researchers and graduate students, supporting prior recommendations (Hutchinson & Lovell, 2004).

We doubt that researchers infrequently report use of MI because they have compared the results of listwise deletion (or other methods) with multiple imputation and concluded that the results were similar. When this happens, however, we recommend that researchers mention this comparison and conclusion in their paper. It is more probable that the lack of evidence of MI in higher education research represents a combination of a dearth of understanding of the technique and skepticism about MI, both issues we have addressed here. It is also probable that the inconsistent reporting of MI when it has been used is partly due to the lack of guidelines for reporting the technique. We hope our synthesis of recommended reporting practices provides such guidance to researchers.

We recognize that authors and editors may be wary of using significant space to report on MI at the expense of other information. As the procedures commonly used to handle missing data become more complex, as they are with MI, more authors and editors concerned about journal space may take advantage of making appropriate additional detail available online (e.g. see van Buuren and Groothuis-Oudshoorn's (2011) introduction to MI or Alon and Tienda's (2007) research article for examples of online supplemental material). While authors should strive for conciseness and efficiency in including the recommended information, the use of

online supplements should be considered by more journals to provide the information needed for replication and a complete understanding of the analyses that were conducted.

Peugh and Enders (Peugh & Enders, 2004) found evidence of a gap between recommendations in the statistical literature and applied researchers' use of missing data handling techniques in education. Our findings support and extend this conclusion for the use of MI in higher education. van Buuren and Groothuis-Oudshoorn (2011) identified a gap in reporting guidelines for use of MI as well as the need for "entry-level texts that explain the idea and that demonstrate how to use the techniques in practice" (p. 57). As researchers who are neither statisticians nor MI developers, we agree with them and feel confident we are not the only researchers using large education datasets who have encountered MI and found the existing guidance in the literature lacking in clarity.

Overall, our findings imply the need to convey to applied researchers in higher education that the newest state of the art includes MI. This paper outlines the components necessary to clearly report use of the MI technique and highlights the moments when a researcher's subjective sense is involved in the decision-making process, including decisions that even statistical experts do not wholly agree upon. We hope this will lead to more investigation of the technique by higher education researchers and more accurate and appropriate implementation when it is selected to address missing data. We also hope that others knowledgeable about MI will continue the effort to communicate MI in more accessible terms for applied researchers in the future.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*, 1012-1028. doi: 10.1111/j.1741-3737.2005.00191.x

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

Alon, S., & Tienda, M. (2007). Diversity, opportunity, and the shifting meritocracy in higher education. *American Sociological Review, 72*(4), 487-511. doi: 10.1177/000312240707200401

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40-49. doi: 10.1002/mpr.329

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. doi: 10.1016/J.Jsp.2009.10.001

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*(4), 948-955.

Barzi, F., & Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology, 160*(1), 34-45. doi: 10.1093/Aje/Kwh175

Berelson, B. (1952). *Content analysis in communication research*. Glencoe, IL: The Free Press.

Bray, N. J., & Major, C. H. (2011). Status of journals in the field of higher education. *Journal of Higher Education, 82*(4), 479-503.

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer, 91*, 4-8. doi: 10.1038/sj.bjc.6601907

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research. *New Directions for Institutional Research, 2005*(127), 33-49.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*, 1087-1091.

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 313-342). Greenwich, CT: Information Age Publishing.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. doi: 10.1146/annurev.psych.58.110405.085530

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206-213.

Heeringa, S., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Ho, P., Silva, M. C. M., & Hogg, T. A. (2001). Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the ageing of port. *Chemometrics and Intelligent Laboratory Systems, 55*, 1-11.

Huckin, T. (2004). Content analysis: What texts talk about. In C. Bazerman & P. Prior (Eds.), *What writing does and how it does it: An introdution to analyzing texts and textual practices* (pp. 13-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Hutchinson, S. R., & Lovell, C. D. (2004). A review of methodological characteristics of research published in key journals in higher education: Implications for graduate research training. *Research in Higher Education, 45*(4), 383-403.

Jackson, D. L., Gillaspy Jr., J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6-23. doi: 10.1037/a0014694

Jelicic, H., Phelps, E., & Lerner, R. A. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45*(4), 1195-1199. doi: 10.1037/A0015665

Klebanoff, M. A., & Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology, 168*(4), 355-357. doi: 10.1093/Aje/Kwn071

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology, 171*(5), 624-632. doi: 10.1093/Aje/Kwp425

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, TX: Stata Press.

Ludtke, O., Robitzsch, A., Trautwein, U., & Koller, O. (2007). Handling of missing data in psychological research: Problems and solutions. *Psychologische Rundschau, 58*(2), 103-117. doi: 10.1026/0033-3042.58.2.103

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

Nakagawa, S., & Freckleton, R. P. (2011). Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behavioral Ecology and Sociobiology, 65*, 103-116. doi: 10.1007/s00265-010-1044-7

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556. doi: 10.3102/00346543074004525

Royston, P. (2004). Mutliple imputation of missing value. *Stata Journal, 4*(3), 227-241.

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: Wiley.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. doi: 10.1037/1082-989X.7.2.147

Silverman, R. J. (1987). How we know what we know: A study of higher education journal articles. *Review of Higher Education, 11*(1), 39-59.

Social Science Computing Cooperative. (2012, August 24). Multiple imputation in Stata: Introduction. University of Wisconsin, Madison. Retrieved September 27, 2012, from http://www.ssc.wisc.edu/sscc/pubs/stata_mi_intro.htm

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal, 339*. doi: 10.1136/bmj.b2393

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology, 169*(9), 1133-1139. doi: 10.1093/Aje/Kwp026

Treiman, D. J. (2009). *Quantitative data analysis: Doing social research to test ideas*. San Francisco: Jossey-Bass.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064. doi: 10.1080/10629360600810434

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67.

Van Ginkel, J. R. (2010). Investigation of multiple imputation in low-quality questionnaire data. *Multivariate Behavioral Research, 45*(3), 574-598. doi: 10.1080/00273171.2010.483373

Wells, R., Kolek, E., Williams, L. & Saunders, D. (2012, Nov.). *Methodological characteristics of research in* Review of Higher Education*: Implications for knowledge production in the field of higher education.* Paper presented at the Association for the Study of Higher Education (ASHE) Annual Meeting, Las Vegas, NV.

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. doi: 10.1002/sim.4067

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. doi: 10.1037/0003-066X.54.8.594

Table 1. Recommended MI Reporting Practices

Describe the nature and structure of any missing data
• Overall percentage of missing values
• Range of missing data rates across variables
• Reasons data is missing, if identifiable, e.g.
o Description of any planned missing data
o Description in terms of other variables if relevant
• Evidence of ignorable patterns or assumptions made, e.g.
o Missing completely at random (MCAR) or missing at random (MAR)
o Mean comparisons of missing and complete cases when identifying auxiliary variables (correlates of missingness) to make the MAR assumption more plausible
o Sensitivity analysis to detect nonrandom missing data (MNAR)
Describe the imputation model and procedures
• Variables used in imputation phase, including auxiliary variables, interactions, etc.
• Software, version, and command used in order to communicate the algorithm/procedure chosen, e.g. "*mi impute chained* in Stata v.12"
o Key non-default model options, e.g. burn-in and between-imputation iterations
o Cite appropriate reference(s) for the procedure chosen
• Other relevant special considerations, e.g. scales, multilevel data
• Number of imputations
Describe any notable imputation results
• Compare observed and imputed values, particularly with a high rate of missing data
• Discuss any discrepancies in results if multiple methods for handling missing data were employed

Appendix A. Literature addressing higher education and using MI, included in the text analysis

Alexander, K., Bozick, R., & Entwisle, D. (2008). Warming up, cooling out, or holding steady? Persistence and change in educational expectations after high school. *Sociology of Education*, *81*(4), 371 –396.

Alon, S. (2009). The evolution of class inequality in higher education competition, exclusion, and adaptation. *American Sociological Review*, *74*(5), 731–755.

Alon, S. (2010). Racial differences in test preparation strategies: A commentary on shadow education, American style: Test preparation, the SAT and college enrollment. *Social Forces*, *89*(2), 463–474.

Alon, S., & Tienda, M. (2007). Diversity, opportunity, and the shifting meritocracy in higher education. *American Sociological Review*, *72*(4), 487–511.

Attewell, P. A., Domina, T., Lavin, D. E., & Levey, T. (2006). New evidence on college remediation. *The Journal of Higher Education*, *77*(5), 886–924.

Bennett, P. R., & Lutz, A. (2009). How African American is the net black advantage? Differences in college attendance among immigrant blacks, native blacks, and whites. *Sociology of Education*, *82*(1), 70–100.

Bobbitt-Zeher, D. (2007). The gender income gap and the role of education. *Sociology of Education*, *80*(1), 1 –22.

Bozick, R. (2007). Making it through the first year of college: The role of students' economic resources, employment, and living arrangements. *Sociology of Education*, *80*(3), 261–284.

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow education, American style: Test preparation, the SAT and college enrollment. *Social Forces*, *89*(2), 435–461.

Chen, R., & DesJardins, S. L. (2010). Investigating the impact of financial aid on student dropout risks: Racial and ethnic differences. *The Journal of Higher Education*, *81*(2), 179–208.

Chen, R., & John, E. P. S. (2011). State financial policies and college student persistence: A national study. *The Journal of Higher Education*, *82*(5), 629–660.

Crisp, G., & Nora, A. (2010). Hispanic student success: Factors influencing the persistence and transfer decisions of Latino community college students enrolled in developmental education. *Research in Higher Education*, *51*(2), 175–194.

Doyle, W. R. (2010). Changes in institutional aid, 1992-2003: The evolving role of merit aid. *Research in Higher Education*, *51*(8), 789–810.

Engberg, M., & Allen, D. (2011). Uncontrolled destinies: Improving opportunity for low-income students in American higher education. *Research in Higher Education*, *52*(8), 786–807.

Goldrick-Rab, S., & Pfeffer, F. T. (2009). Beyond access: Explaining socioeconomic differences in college transfer. *Sociology of Education*, *82*(2), 101–125.

Hahs-Vaughn, D. (2004). The impact of parents' education level on college students: An analysis using the Beginning Postsecondary Students Longitudinal Study 1990-92/94. *Journal of College Student Development*, *45*(5), 483–500.

Harding, D. J. (2011). Rethinking the cultural context of schooling decisions in disadvantaged neighborhoods from deviant subculture to cultural heterogeneity. *Sociology of Education*, *84*(4), 322–339.

Hill, D. H. (2008). School strategies and the "college-linking" process: Reconsidering the effects of high schools on college enrollment. *Sociology of Education*, *81*(1), 53–76.

Kalogrides, D., & Grodsky, E. (2011). Something to fall back on: Community colleges as a safety net. *Social Forces*, *89*(3), 853–877.

Kim, D. H., & Schneider, B. (2005). Social capital in action: Alignment of parental support in adolescents' transition to postsecondary education. *Social Forces*, *84*(2), 1181–1206.

Klasik, D. (2012). The college application gauntlet: A systematic analysis of the steps to four-year college enrollment. *Research in Higher Education*, *53*(5), 506–549.

Kugelmass, H., & Ready, D. (2011). Racial/ethnic disparities in collegiate cognitive gains: A multilevel analysis of institutional influences on learning and its equitable distribution. *Research in Higher Education*, *52*(4), 323–348.

Marti, C. N. (2008). Dimensions of student engagement in American community colleges: Using the Community College Student Report in research and practice. *Community College Journal of Research and Practice*, *33*(1), 1–24.

Mattanah, J. F., Ayers, J. F., Brand, B. L., Brooks, L. J., Quimby, J. L., & McNary, S. W. (2010). A social support intervention to ease the college transition: Exploring main effects and moderators. *Journal of College Student Development*, *51*(1), 93–108.

Morrison, E., Rudd, E., Picciano, J., & Nerad, M. (2011). Are you satisfied? PhD education and faculty taste for prestige: Limits of the prestige value system. *Research in Higher Education*, *52*(1), 24–46.

Owens, A. (2010). Neighborhoods and schools as competing and reinforcing contexts for educational attainment. *Sociology of Education*, *83*(4), 287–311.

Reynolds, J. R., & Baird, C. L. (2010). Is there a downside to shooting for the stars? Unrealized educational expectations and symptoms of depression. *American Sociological Review*, *75*(1), 151–172.

Reynolds, J. R., & Johnson, M. K. (2011). Change in the stratification of educational expectations and their realization. *Social Forces*, *90*(1), 85–109.

Rockey, D. L., Beason, K. R., Howington, E. B., Rockey, C. M., & Gilbert, J. D. (2005). Gambling by Greek-affiliated college students: An association between affiliation and gambling. *Journal of College Student Development*, *46*(1), 75–87.

Scott, M., Bailey, T., & Kienzl, G. (2006). Relative success? Determinants of college graduation rates in public and private colleges in the U.S. *Research in Higher Education*, *47*(3), 249–279.

Strayhorn, T. (2007). Factors influencing the academic achievement of first-generation college students. *Journal of Student Affairs Research and Practice*, *43*(4).

Torche, F. (2005). Privatization reform and inequality of educational opportunity: The case of Chile. *Sociology of Education*, *78*(4), 316–343.

Turley, R. N. L. (2009). College proximity: Mapping access to opportunity. *Sociology of Education*, *82*(2), 126 –146.

Wells, R. S., Seifert, T. A., Padgett, R. D., Park, S., & Umbach, P. D. (2011). Why do more women than men want to earn a four-year degree?: Exploring the effects of gender, social origin, and social capital on educational expectations. *The Journal of Higher Education*, *82*(1), 1–32.

Appendix B. Stata 12 code illustrating use of MI – Step 1, recoding the NCES ELS variables and preparing the dataset for imputation

```
capture log close
log using example-data01-prep, replace text

//  program:    example-data01-prep.do
//  task:       multiple imputation example - data preparation
//  project:    multiple imputation in higher education, NERA 2012 Conference
//  author:     cathy manly and ryan wells \ 2012-12-15

//     program setup, date and tag

version 12
set linesize 80
clear all
macro drop _all
set mem 500m
set more off

local date "2012-12-15"
local tag "example-data01-prep.do cam `date'."

//  load data

use els2002, clear                              // source dataset downloaded from NCES

//     keep only selected variables

keep F2PTN1PS F2B29A F2PS1AID F2PS1NTY F2PS1LN F1SEX F1SEXIM F1RACE F1RACEIM F1SES2
F2B18B F2B18G F2B18A F2B18E F2B18C F2B18F F2B18D F2PSPPLN STU_ID PSU STRAT_ID G12COHRT
F2F1WT F2QSTAT

//     value definitions

*                           "1234567890"
label define Limflag  0     "orig_data", modify       // original data-not imputed
label define Limflag  1     "BY_impute", modify       // value imputed in BY
label define Limflag  2     "F1_impute", modify       // value imputed in F1
*                             "1234567890"
label define Lyesno     0     "0No", modify
label define Lyesno     1     "1Yes", modify
*                             "1234567890"
label define raceall    1     "1AmerIndian", modify   // Alaska Native, non-Hispanic
label define raceall    2     "2Asian", modify        // Hawaii/Pac. Islander,non-
Hispanic
label define raceall    3     "3Black", modify        // or African-American, non-
Hispanic
label define raceall    4     "4HispNoRace", modify   // Hispanic, no race specified
label define raceall    5     "5HispRace", modify     // Hispanic, race specified
label define raceall    6     "6>1Race", modify       // non-Hispanic
label define raceall    7     "7White", modify        // non-Hispanic
*                             "1234567890"
label define gender     1     "1Male", modify
label define gender     2     "2Female", modify
*                             "1234567890"
label define Lfrq_nso   1     "1Never", modify
label define Lfrq_nso   2     "2Sometimes", modify
label define Lfrq_nso   3     "3Often", modify
*                             "1234567890"
label define f2f1wt     0     "{Zero}", modify
```

```
*                           "1234567890"
label define ofraid     1   "1App&Aid", modify      // app for aid, aid offered
label define ofraid     2   "2NApp&Aid", modify     // no aid app, aid offered
label define ofraid     3   "3App&NAid", modify     // app for aid, no aid offrd
label define ofraid     4   "4NApp&NAid", modify    // no aid app, no aid offrd
label define ofraid     5   "5NAdmitApp", modify    // no admission app
*                           "1234567890"
label define ofrtypes   0   "0NAid", modify         // Not offered aid by PS1
label define ofrtypes   1   "1AidOfferd", modify    // Offrd 1 type aid by PS1
label define ofrtypes   2   "2AidOfferd", modify    // Offrd 2 types aid by PS1
label define ofrtypes   3   "3AidOfferd", modify    // Offrd 3 types aid by PS1
label define ofrtypes   4   "4AidOfferd", modify    // Offrd 4 types aid by PS1
label define ofrtypes   5   "5NAdmAppPS", modify    // no admit/no aid app/no ps
label define ofrtypes   97  "97NAdmitApp", modify   // no admission app 1st inst
label define ofrtypes   98  "98NApp", modify        // no aid application
label define ofrtypes   99  "99NPS", modify         // No PS attendance as of F2
*                           "1234567890"
label define pspipeline  0  "0StillHS", modify            // Still in hs as of F2
label define pspipeline  1  "1NPip&NPS", modify     // Never entered pipe, no PS
label define pspipeline  2  "2PPip&NPS", modify     // Partial pipeline, no PS
label define pspipeline  3  "3PNPip&<4yr", modify   // Partial/no; 1st att <4yr
label define pspipeline  4  "4PNPip&4yr", modify    // Partial/no; 1st att 4yr
label define pspipeline  5  "5Pip&NPS", modify            // Completed pipeline; no
PS
label define pspipeline  6  "6Pip&<4yr", modify     // Compl pipe; 1st att <4yr
label define pspipeline  7  "7Pip&4yr", modify            // Compl pipe; 1st att 4yr
*                           "1234567890"
label define pspattern  1   "1_4yrNTran", modify    // 4yr-4yr, no transfer
label define pspattern  2   "2_4yrYTran", modify    // 4yr-4yr, w/transfer
label define pspattern  3   "3_4yr-<4yr", modify    // 4yr-<4yr
label define pspattern  4   "4_4yr-NEnr", modify    // 4yr-not enrolled
label define pspattern  5   "5_<4yrNTran", modify   // <4yr-<4yr, no transfer
label define pspattern  6   "6_<4yrYTran", modify   // <4yr-<4yr, w/transfer
label define pspattern  7   "7_<4yr-4yr", modify    // <4yr-4yr
label define pspattern  8   "8_<4yr-NEnr", modify   // <4yr-not enrolled
label define pspattern  9   "9_NEnr-4yr", modify    // Not enrolled-4yr
label define pspattern  10  "10NEnr-<4yr", modify   // Not enrolled-<4yr
label define pspattern  11  "11NEnrWPS", modify     // Not-not enrld, w/some PSE
label define pspattern  12  "12NoPS", modify        // No PSE as of January 2006
label define pspattern  13  "13StillHS", modify     // Still in hs in Jan 2006
*                           "1234567890"
label define f2qstat    0   "0NIntervw", modify     // Interview not complete
label define f2qstat    1   "1YIntervw", modify     // Completed an interview
label define f2qstat    2   "2PartIntvw", modify    // Partial interview
*                           "1234567890"
label define g12cohrt   0   "0NSrCohrt", modify     // Not senior cohort member
label define g12cohrt   1   "1YSrCohrt", modify     // F1 identified sen cohort
label define g12cohrt   2   "2F2SrCohrt", modify    // F2/trnscpt ident sr cohrt
*                           "1234567890"
label define psu        1   "PSU1", modify
label define psu        2   "PSU2", modify
label define psu        3   "PSU3", modify
*                           "1234567890"

//  rename and relabel variables

       local vin    F2PTN1PS
       local vout   Cpspattern
       local vval   pspattern
       local vlab   "F2 ps attendance pattern"
       rename `vin' `vout'
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': rename based on `vin' \ `tag'
```

```
local vin    F2B29A
local vout   Bdegdone
local vval   Lyesno
local vlab   "F2 29A degree complete-done"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2PS1AID
local vout   Cofraid
local vval   ofraid
local vlab   "F2 financial aid offered"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2PS1NTY
local vout   Cofrtypes
local vval   ofrtypes
local vlab   "F2 # of aid types offered"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2PS1LN
local vout   Bofrloan
local vval   Lyesno
local vlab   "F2 loan 1st inst"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F1SEX
local vout   Bncesgender
local vval   gender
local vlab   "F1 student gender"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F1SEXIM
local vout   Cimgender
local vval   Limflag
local vlab   "F1SEX imputation flag"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F1RACE
local vout   Cncesraceall
local vval   raceall
local vlab   "F1 race/ethnicity-all groups"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F1RACEIM
local vout   Cimraceall
```

```
local vval   Limflag
local vlab   "F1RACE imputation flag"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F1SES2
local vout   sesnces
local vval   ses
local vlab   "F1 socioeconomic status"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18B
local vout   Cintadvis
local vval   Lfrq_nso
local vlab   "F2 meet with advisor"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18G
local vout   Cintextra
local vval   Lfrq_nso
local vlab   "F2 extracurriculars"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18A
local vout   Cintfac
local vval   Lfrq_nso
local vlab   "F2 talk with faculty"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18E
local vout   Cintintra
local vval   Lfrq_nso
local vlab   "F2 intramural/nonvarsity sport"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18C
local vout   Cintlib
local vval   Lfrq_nso
local vlab   "F2 study at library"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18F
local vout   Cintsport
local vval   Lfrq_nso
local vlab   "F2 varsity/intercollege sports"
```

```
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2B18D
local vout   Cintweb
local vval   Lfrq_nso
local vlab   "F2 library via web for classes"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2PSPPLN
local vout   Cpspipeline
local vval   pspipeline
local vlab   "F2 postsecondary ed pipeline"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    STU_ID
local vout   id_stu
local vval   id_stu
local vlab   "student id"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    PSU
local vout   psu
local vval   psu
local vlab   "primary sampling unit"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    STRAT_ID
local vout   strat_id
local vval   strat_id
local vlab   "stratum"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    G12COHRT
local vout   g12cohrt
local vval   g12cohrt
local vlab   "F1 senior cohort"
rename `vin' `vout'
label var `vout' "`vlab'"
label val `vout' `vval'
notes `vout': rename based on `vin' \ `tag'

local vin    F2F1WT
local vout   f2f1wt
local vval   f2f1wt
local vlab   "F2 weights"
rename `vin' `vout'
label var `vout' "`vlab'"
```

```
        label val `vout' `vval'
        notes `vout': rename based on `vin' \ `tag'

        local vin    F2QSTAT
        local vout   f2qstat
        local vval   f2qstat
        local vlab   "F2 participants"
        rename `vin' `vout'
        label var `vout' "`vlab'"
        label val `vout' `vval'
        notes `vout': rename based on `vin' \ `tag'

//      mvdecode NCES imputed values to sysmiss

** subjective decision: decide whether to decode prior to imputations

* NCES imputations in: Bncesgender Cncesraceall
* NCES imputation flags in: Cimgender Cimraceall
* create clones with "data" in varnames to identify vars without imputations

* change values of Bdatagender imputed by NCES to sysmiss
        clonevar Bdatagender = Bncesgender
        mvdecode Bdatagender if Cimgender!=0, mv(1 2=.)       // decode cases to sysmiss
        note Bdatagender: NCES imputed cases to sysmiss for Bncesgender clone \ `tag'

* change values of Cncesraceall imputed by NCES to sysmiss
        clonevar Cdataraceall = Cncesraceall
        tab Cdataraceall Cimraceall, miss
        mvdecode Cdataraceall if Cimraceall!=0, mv(0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
            15 16 17 18 19 20=.)
        note Cdataraceall: NCES imputed cases coded as sysmiss for Cdataraceall \ `tag'

//      special recoding of legitimate skips: -3 {Item legitimate skip/NA}

* recode Bdegdone
        * recode Bdegdone -3 = 0 because still enrolled (Cpspattern==enrolled)
              foreach i in  1 2 3 5 6 7 {
                    recode Bdegdone (-3 = 0) if Cpspattern==(`i')
              }
              note Bdegdone: Legit item skips (-3) coded as 0 for Bdegdone if still
                 enrolled \ `tag'

        * recode Bdegdone -3 = 0 because no ps or still in high school
              recode Bdegdone (-3 = 0) if Cpspattern==12|Cpspattern==13
              note Bdegdone: Legit item skips (-3) coded as 0 for Bdegdone if still in
                 high school \ `tag'

        * recode Bdegdone -3 = 0 because never enrolled in ps (Cofrtypes==99)
              recode Bdegdone (-3 = 0) if Cofrtypes==99
              note Bdegdone: Legit item skips (-3) coded as 0 for Bdegdone if still
                 enrolled \ `tag'

* recode Cofraid -3 = 4 because never enrolled in postsecondary (Cofrtypes==99)
        * orig label 4 => Did not apply for aid, no aid offered
        recode Cofraid (-3 = 4) if Cofrtypes==99 // 4: no aid application/no aid offer
        note: Cofraid: Legit item (-3) skips coded as 4 for Cofraid if no ps \ `tag'

* recode Bofr*
        * recode Bofrloan -3 = 0 because did not attend postsecondary ed or still in hs
            recode Bofrloan (-3 = 0) if Cpspattern==12 | Cpspattern==13
            note Bofrloan: Legit item (-3) skips coded as 0 for Bofrloan if no ps or
               still in hs \ `tag'

        * recode Bofrloan -3 = 0 because no aid offered or no application
            recode Bofrloan (-3 = 0) if Cofraid==4 | Cofraid==5
```

```
          note Bofrloan: Legit item (-3) skips coded as 0 for Bofrloan if no aid
            offered or no application \ `tag'

//     recode missing as sysmiss

* recode missing in binary/categorical vars as sysmiss
       *create an alphabetized list of the binary and categorical vars
              save x-temp, replace
              keep B* C*
              drop Bdata* Cdata* Bnces* Cnces* Cim* // exclude NCES imputation vars
              aorder
              unab varlist : _all
              display "`varlist'"
              use x-temp, clear

       foreach varname in `varlist' {
           recode `varname' (-9/-1 = .) (missing = .) // for bin/categorical vars
           note `varname': Missing all coded as sysmiss for `varname' \ `tag'
       }

recode sesnces (-8 = .) (missing = .)                        // recode ses separately

//     recode variables

* Benroll - from Cpspattern and Cofrtypes
       local vin    Cpspattern
       local vout   Benroll
       local vval   Lyesno
       local vlab   "F2 any initial ps enrollment?"
       recode `vin' (1/8=1) (9/13=0), gen(`vout')
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': binary based on `vin' \ `tag'

       * Cofrtypes==99 also identifies individuals with no postsecondary ed
       recode `vout' (.=0) if Cofrtypes==99    // not enrollee if no ps
       notes `vout': also based on Cofrtypes \ `tag'

       * consider individual a postsecondary enrollee if completed a degree
       recode `vout' (0 .=1) if Bdegdone==1    // enrollee if completed a degree
       notes `vout': also based on Bdegdone \ `tag'

* Baidofferd - from Cofraid
       local vin    Cofraid
       local vout   Baidofferd
       local vval   Lyesno
       local vlab   "F2 aid offered by 1st inst?"
       recode `vin' (1 2=1) (3/5=0), gen(`vout')
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': binary based on `vin' \ `tag'

* Bpersistps - from Cpspattern and Bdegdone
       local vin    Cpspattern
       local vout   Bpersistps
       local vval   Lyesno
       local vlab   "F2 persist in any ps ed?"
       recode `vin' (1/3 5/7=1) (4 8 9/13=0), gen(`vout')
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': binary based on `vin' \ `tag'

       * consider individual as a postsecondary persister if completed a degree
       recode `vout' (0 .=1) if Bdegdone==1    // persister if completed a degree
       notes `vout': also based on Bdegdone \ `tag'
```

```
* Bfemale - from Bdatagender
       label def    female       0 0_Male 1 1_Female
       local vin    Bdatagender
       local vout   Bfemale
       local vval   female
       local vlab   "F1 is student female?"
       recode `vin' (1=0) (2=1), gen(`vout')
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': binary based on `vin' \ `tag'

* Crace - from Cdatarace (white=reference group #1)
       label def    race         1 1White     2 2Asian     3 3Black     ///
                                 4 4Hisp             6 "6>1race"   8 8AmerIndian
       local vin    Cdatarace
       local Xvout Xracetmp
       local vout   Crace
       local vval   race
       local vlab   "F1 race/ethnicity-hisp combined, white ref"
       recode `vin' (1=8) (2=2) (3=3) (4/5=4) (6=6) (7=7), gen(`Xvout')
       recode `Xvout' (2=2) (3=3) (4=4) (6=6) (7=1) (8=8), gen(`vout')
       drop `Xvout'
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': based on `vin' \ `tag'

* Bpspipe - from Cpspipeline
       local vin    Cpspipeline
       local vout   Bpspipe
       local vval   Lyesno
       local vlab   "F2 stud completed ps pipeline?"
       recode `vin' (0/4=0) (5/7=1), gen(`vout')
       label var `vout' "`vlab'"
       label val `vout' `vval'
       notes `vout': binary based on `vin' \ `tag'

//     calculate principal component analysis scores

** subjective decision: whether to impute component variables or whole scales

pca Cint* [aweight = f2f1wt] if Benroll==1, mineigen(1)
rotate, varimax normalize blanks(.3)
predict intacad intsoc if Benroll==1, score
label var intacad "F2 academic integration"
label var intsoc "F2 social integration"
notes intacad: principal component analysis scores based on Cint* \ `tag'
notes intsoc: principal component analysis scores based on Cint* \ `tag'

//     drop cases and variables that are not used for analysis

* drop race data that is too small
drop if Crace==8                 // American Indian (recoded to 8)
drop if Crace==6                 // multiracial

* drop vars that were used to recode vars
drop Cpspattern                  // only keep Bpersistps, Benroll
drop Bdegdone                    // only keep Bpersistps
drop Cofrtypes                   // only keep Benroll
drop Bdatagender                 // only keep Bfemale
drop Cdatarace                   // only keep Crace
drop Cpspipeline                 // only keep Bpspipe
drop Cofraid                     // only keep Baidofferd
```

```
* drop variables with NCES imputed data and imputation flags
drop Bnces* Cnces*               // for gender raceall
drop Cim*                        // NCES imputation flags

* drop vars used in principal components analysis
drop Cint*                       // only keep intacad, intsoc

//  check the variables

codebook, compact

isid id_stu                      // check the id variable
codebook id_stu, compact         // compare to after mi

//  closeup and save data

quietly compress
label data "example \ ELS:2002-06 dataset, prepared for mi \ `date'"      // 80 chars
note: example-data01.dta \ ELS data prepared for multiple imputation \ `tag'
datasignature set, reset
save example-data01, replace

* check the dataset
use example-data01, clear
datasignature confirm
note _dta

log close
exit
```

Appendix C. Stata 12 code illustrating use of MI – Step 2, setting up and conducting MI

```
capture log close
log using example-data02-mi, replace text

//  program:    example-data02-mi.do
//  task:       multiple imputation example - impute
//  project:    multiple imputation in higher education, NERA 2012 Conference
//  author:     cathy manly and ryan wells \ 2012-12-15

//     program setup, date and tag

version 12
set linesize 80
clear all
macro drop _all
set mem 500m
set more off

local date "2012-12-15"
local tag "example-data01-prep.do cam `date'."

//     load data

use example-data01, clear
datasignature confirm
notes _dta

//     define locals

* variables with no missing data (to be registered as regular variables)
local regularlist   "f2f1wt f2qstat g12cohrt id_stu psu strat_id"

** recommendation: report variables used in imputation
* variables with missing data (to be registered as imputed variables)
local imputelist    "Baidofferd Benroll Bfemale Bofrloan Bpersistps Bpspipe Crace
intacad intsoc sesnces"

** recommendation: report the number of imputations
* number of imputations (m=nummi)
       * start with nummi=2 to determine model setup and debug Stata code
       * try nummi=4 to test speed when doubling m
       * nummi=5 was standard but more may be better
       * consider at least nummi=20
       * guideline: set nummi slightly larger than the largest % of missing data
local nummi   30

* base imputation command
* note: will be used several places, so keep consistent using a local macro
* note: imputation may be too slow with >50-70 variables
local micommand "mi impute chained (regress) sesnces intacad intsoc (logit) Bfemale
Bpspipe Benroll Baidofferd (logit, conditional(if Benroll==1) omit(i.Benroll))
Bpersistps (logit, conditional(if Baidofferd==1) omit(i.Baidofferd)) Bofrloan (mlogit)
Crace = i.psu [pweight = f2f1wt] , add(`nummi') rseed(394857235) augment dots
chaindots report"

//     verify vars (in local regularlist) that have no missing data

misstable summarize `regularlist'       // note: will be blank if nothing missing

//     check the missing data to be imputed (in local imputelist)

describe `imputelist'
summarize `imputelist'
```

```
misstable summarize `imputelist'        // shows variables with missing data

* examine patterns of missing data (long output) - need to 'mi set flong' to do
*mi set flong
*mi misstable patterns, bypatterns

* check whether the missing data pattern is monotone or arbitrary
misstable nested `imputelist'
* result: data not monotone (so do not use 'mi impute monotone' here)
* note: if you use conditional imputation, you need nested variables

** recommendation: report overall percentage of missing data
* check overall percentage of missing data
quietly sum id_stu
local totaln = r(N)                     // capture total N
quietly logit `imputelist'
local totalld = e(N)                    // capture N under listwise deletion
display _newline "total N: `totaln'" _newline ///
"N if all cases with missing data dropped (listwise deletion): `totalld'" _newline
"percent of cases with missing data: "1-`totalld'/`totaln'

** recommendation: report range of missing data rates across variables
* check percentage of missing data for each var
misstable sum, gen(miss_)               // generate missingness indicator variables
label def Lismiss 0 0_valid 1 1_missing
foreach varname in `imputelist' {
       label var miss_`varname' "`varname' is missing?"
       label val miss_`varname' Lismiss
}
tab1 miss_*                             // best if 1_missing<10%, look out for >50%

//     determine conditional imputation relationships-placed in local macro micommand

tab Bpersistps Benroll, miss
/* logical statement: in order to persist at your first institution, you had to attend
a first institution
conditional to use: (logit, conditional(if Benroll==1) omit(i.Benroll)) Bpersistps
*/

tab Bofrloan Baidofferd, miss
/* logical statement: if you were offered a loan by your first institution, then you
must have been offered aid from your first institution
conditional to use: (logit, conditional(if Baidofferd==1) omit(i.Baidofferd)) Bofrloan
*/

//     visual check of values for continuous vars

dotplot sesnces int*       // compare to after imputation to verify valid imputations
graph export example-data02-premi-continuousvars.png, replace

//     set and register the mi data

* check if data are already mi set
mi query                                        // expect data not mi set yet

mi set flong

* register vars with missing data as imputation vars
mi register imputed `imputelist'

* register other vars as regular (not for imputation)
mi register regular `regularlist'

* register any passive variables (e.g. var transformations) with mi register passive
```

```
//     update and verify the mi data

mi update                                       // do this after all changes to mi data
mi query                                        // expect mi data are set
mi describe                                     // gives # of vars to be imputed

//     use dryrun option for mi impute to refine prediction model specification

display ". `micommand' dryrun"
`micommand' dryrun

* debugging note: try out models from dryrun output individually to make sure they run
* debugging note: use the noisily option to see what the mi impute command does

//     investigate convergence (subjective decision)

* trace plots of means and standard deviations of imputed values in 1 chain
save x-temp, replace
display ". `micommand' chainonly burnin(50) savetrace(x-impstats1, replace)"
`micommand' chainonly burnin(50) savetrace(x-impstats1, replace)

use x-impstats1, clear
sum *_mean *_sd                                 // to identify means for drawing lines
tsset iter
tsline sesnces_mean, name(gr1, replace) nodraw yline(-.26)
tsline sesnces_sd, name(gr2, replace) nodraw yline(.69)
tsline Benroll_mean, name(gr3, replace) nodraw yline(.60)
tsline Benroll_sd, name(gr4, replace) nodraw yline(.49)
tsline Bofrloan_mean, name(gr5, replace) nodraw yline(.61)
tsline Bofrloan_sd, name(gr6, replace) nodraw yline(.49)
tsline intacad_mean, name(gr7, replace) nodraw yline(-.76)
tsline intacad_sd, name(gr8, replace) nodraw yline(1.42)
graph combine gr1 gr2 gr3 gr4 gr5 gr6 gr7 gr8, title(Trace plots of summaries of
imputed values) rows(4)
graph export example-data02-mi-diagnostics-chainvalues.png, replace

//     impute data

timer clear 1
timer on 1                  // set timer to find out how long the imputation takes

** recommendation: report software, version, and command used in order to communicate
the algorithm/procedure chosen, including key non-default model options (e.g. burn-in
and between-imputation iterations)
* issue the 'mi impute chained' command from local macro defined above
use x-temp, clear
display ". `micommand' burnin(30) savetrace(x-impstats2, replace)"
`micommand' burnin(30) savetrace(x-impstats2, replace)

* show time for imputation (3 equivalent ways: seconds, minutes, and hours)
timer off 1
timer list 1                // imputation time in seconds
local tsec = r(t1)
local tmin = r(t1)/60
local thr  = r(t1)/60/60
display "timer results for m=`nummi': `tsec' sec, or `tmin' min, or `thr' hrs"

//     verify mi data

mi update
mi query
mi describe
mi varying                  // identify variables that vary over imputations
```

```
** recommendation: describe any relevant special considerations for your dataset
* (e.g. special handling of scales, multilevel data)

//     create a variable to identify the intended sample subpopulation

mi passive: generate subsample = Benroll==1 & Baidofferd==1 & g12cohrt!=0
label var subsample "analysis subsample"
label val subsample Lyesno
notes: subsample: binary based on Benroll==1 & Baidofferd==1 & g12cohrt!=0 \ `tag'
mi update

//     recode variables post-mi

* intacad and intsoc only have valid values for individuals enrolled in postsec ed
foreach varname in intacad intsoc {
       * make a copy of acad/soc intetration variables
       display ". mi passive: generate X`varname' = `varname'"
       mi passive: generate X`varname' = `varname'          // generate copy
       note X`varname': copy of `varname' retaining all data \ `tag'

       * decode acad/soc integration to sysmiss if no postsecondary
       display ". replace `varname'=99 if Benroll==0"
       replace `varname'=99 if Benroll==0                   // make int*=99 if no ps
       mvdecode `varname', mv(99)                           // decode cases to sysmiss
       note `varname': coded as sysmiss if no initial ps enrollment \ `tag'
}
capture drop Xint*                      // only keep intacad intsoc
mi update

//     set for complex survey design

mi svyset psu [pweight=f2f1wt], strata(strat_id) singleunit(centered)
mi update

//     verify values for all vars make sense

unab varlist : _all                     // get a list of all vars
* create a random variable
set seed 1951
generate xselect = int( (runiform()*_N)+ 1 )
label var xselect "Random numbers from 1 to _N"
summarize xselect                       // verify range

* look at a random selection of observations of each var
* note: should include only the missing data from the original dataset (m=0)
foreach varname in `varlist' {
       codebook `varname', compact
       sort `varname'
       list `varname' if xselect<20, clean
}
drop xselect                            // get rid of xselect once done using it

dotplot sesnces int*                    // check values (still) in right range
graph export example-data02-postmi-continuousvars.png, replace

//     check for problems, id variable check/comparison

codebook, problems
codebook id_stu, compact                // # unique id values should = pre-mi
```

```
//  save data and check

quietly compress
label data "example \ ELS:2002-06 dataset, trimmed, mi, svyset \ `date'"  // 80 chars
note: example-data02.dta \ dataset ready to use for analysis (mi) \ `tag'
datasignature set, reset
save example-data02, replace

* check the dataset
use example-data02, clear
datasignature confirm
note _dta

// trace plots of means and std devs of imputed values from multiple chains

use x-impstats2, clear
reshape wide *mean *sd, i(iter) j(m)
tsset iter
tsline sesnces_mean*, name(gr100, replace) nodraw legend(off) ytitle(Mean of ses)
yline(-.26)
tsline sesnces_sd*, name(gr200, replace) nodraw legend(off) ytitle(Std Dev of ses)
yline(.69)
tsline Benroll_mean*, name(gr300, replace) nodraw legend(off) ytitle(Mean of
enrollment) yline(.60)
tsline Benroll_sd*, name(gr400, replace) nodraw legend(off) ytitle(Std Dev of
enrollment) yline(.49)
tsline Bofrloan_mean*, name(gr500, replace) nodraw legend(off) ytitle(Mean of loan
offered) yline(.61)
tsline Bofrloan_sd*, name(gr600, replace) nodraw legend(off) ytitle(Std Dev of loan
offered) yline(.49)
tsline intacad_mean*, name(gr700, replace) nodraw legend(off) ytitle(Mean of academic
integration) yline(-.76)
tsline intacad_sd*, name(gr800, replace) nodraw legend(off) ytitle(Std Dev of academic
integration) yline(1.42)
graph combine gr100 gr200 gr300 gr400 gr500 gr600 gr700 gr800, title(Trace plots of
summaries of imputed values from `nummi' chains) rows(4)
graph export example-data02-mi-diagnostics-imputations.png, replace

//     verify replication ability (need 'mi impute chained' rseed() option)

local nummi 2                                    // set number of imputations for speed
display "nummi: `nummi'"

* impute the first time
use x-temp, clear
display ". `micommand'"
`micommand'
save x-temp-rep1, replace

* impute the second time (should be the same)
use x-temp, clear
display ". `micommand'"
`micommand'
save x-temp-rep2, replace

* verification method 1: cf - compare dataset in memory to this one to verify match
capture noisily cf _all using x-temp-rep1      // blank if match, error if problems

* verification method 2: dta_equal - compare data in 2 datasets to verify match
dta_equal x-temp-rep1 x-temp-rep2              // error listing mismatches if problems

log close
exit
```

Appendix D. Stata 12 code illustrating use of MI – Step 3, analyzing the imputed dataset

```
capture log close
log using example-stat-analysis, replace text

//  program:    example-stat-analysis.do
//  task:       multiple imputation example - statistical analysis
//  project:    multiple imputation in higher education, NERA 2012 Conference
//  author:     cathy manly and ryan wells \ 2012-12-15

//  program setup

version 12
set linesize 80
clear all
macro drop _all
set mem 500m
set more off

//     load data

use example-data02, clear

//     setup local macros for descriptive statistics

mi query
local M = r(M)                                  // use all imputations
display "M: `M'"

local lhs "Bpersistps"                          // dependent (left hand side) variable

* analysis block 1 - independent (right hand side) variable of interest
local rhs1    "Bofrloan"

* analysis block 2 - include controls
local rhs     "Bofrloan Bfemale Bpspipe i.Crace sesnces intacad intsoc"

//     check correlation matrix

xi: corr `rhs'

//     descriptive stats - means

xi: mi estimate, nimputations(`M') post: svy, subpop(subsample): mean `lhs' `rhs'
estimates store alldata
outreg2 using example-stat-desc, replace ///
       title("Estimated (weighted) means and standard errors of the estimates") ///
       ctitle("Overall Mean") sideway noaster dec(3)

//     logistic blocks - impact on persistence of whether loans were offered for aid

mi estimate, or nimputations(`M') post ///
       cformat(%9.3fc) pformat(%5.3fc) sformat(%8.3fc) : ///
       svy, subpop(subsample) : logistic `lhs' `rhs1'
estimates store block1
outreg2 using example-stat-logit, replace ///
       title("Logistic blocks for persistence, odds ratios reported") ///
       ctitle("block1") eform alpha(0.001, 0.01, 0.05) symbol(**, *, +) dec(3)

mi estimate, or nimputations(`M') post ///
       cformat(%9.3fc) pformat(%5.3fc) sformat(%8.3fc) : ///
       svy, subpop(subsample) : logistic `lhs' `rhs'
estimates store block2
outreg2 using example-stat-logit, ///
       ctitle("block2") eform alpha(0.001, 0.01, 0.05) 10pct dec(3)
```

```
estimates table block1 block2, b(%9.3f) star eform   // show results in log

* check the fraction of missing information (gamma)
matrix list e(fmi_mi)            // be wary over .5 or 50%, and try more imputations

//     compare results of listwise deletion (to results from block2 above)

** recommendation: discuss any discrepancies in MI/listwise deletion results
mi xeq 0: svy, subpop(subsample) : logit `lhs' `rhs', or
estimates store block3-ld
outreg2 using example-stat-logit, ///
       ctitle("block3-ld") eform alpha(0.001, 0.01, 0.05) 10pct dec(3)
estimates table block3-ld, b(%9.3f) star eform

log close
exit
```