Center for Demography and Ecology

University of Wisconsin-Madison

Population-Representative Analysis in High School & Beyond

Michael J. Culbertson

CDE Working Paper No. 2025-01

Population-Representative Analysis in High School & Beyond

Michael J. Culbertson

University of Wisconsin–Madison

* Address correspondence by e-mail to mculbertson@wisc.edu. This research was carried out using the facilities of the Center for Demography and Ecology at the University of Wisconsin–Madison (P2CHD047873) and was supported by the National Institute on Aging (U01AG058719) and the Alzheimer's Association (SG-20717567). I thank Eric Grodsky for helpful comments on an earlier draft.

Abstract

National surveys are designed to yield population estimates, but obtaining valid estimates requires respecting complex survey design. Inverse probability weighting and multiple imputation are both approaches to handle differences between analytic samples and the populations they are intended to represent. This guide describes the application of these techniques to the analysis of data from the High School & Beyond study of 1980.

EdSHARe

Population-Representative Analysis in High School & Beyond

Michael J. Culbertson

May 29, 2025

Revised October 31, 2025

Conceptual Background

One of the advantages of national surveys, such as the High School & Beyond study of 1980 (HS&B:80), is that they are specifically designed to allow inferences about an entire national population—in the case of HS&B:80, the population of individuals who were U.S. high school sophomores or seniors in 1980, individuals who are now around 60 years old. But, analyses of survey data yield valid population inferences only if they are conducted in a way compatible with the survey design. This guide discusses how to analyze HS&B:80 data to yield valid population inferences.

What is population representativeness?

Going back to the basics, in all of statistics, we use samples to represent populations. The population includes all the individuals we are interested in; but, we don't have data about everyone (that would be far too expensive). So instead, we collect data only for a selection of individuals, which we call the sample. We want to use the information we collect from the sample to make an inference about the population. In that sense, the sample members "stand in" for or are "representatives" for the members of the population who weren't included in the sample. For our inferences to be correct, the sample members have to be good representatives for the rest of the population.

The quality of our sample members as representatives for the rest of the population depends entirely on **how** the sample members were selected. For valid population inferences, getting into the sample has to be **independent** of the thing you're analyzing. For example, if you're analyzing income, the likelihood of someone getting into your sample can't be related (explicitly or implicitly) to their income.

The simplest way to achieve this is to take a list of all the members of the population and pull names out of a well-shaken hat—this is a **simple random sample**. In this kind of sample, getting into the sample has nothing to do with income (it was only based on the mixed-up locations of names in the hat), so this sample would yield valid inferences about average income in the population (or just about anything else you wanted to measure). Analysis of a simple random sample is easy, because theorems from mathematical statistics tell us that whatever you calculate for the sample (e.g., the sample mean income) is a good estimate for what's happening in the population—no additional complications are needed in the analysis phase, because the sample members are representative of the population directly.

But, there are plenty of reasons why a simple random sample may not be ideal or practical. For example, you may not have a list of the entire population (there is no list of all high school sophomores); or, you may want to increase the statistical power for the analysis of small subgroups of special interest (e.g., examining Black–White health gaps).

In HS&B:80, they didn't have a list of all high school students, so they started with a list of all high schools. They selected high schools randomly, with a higher selection probability for certain types of high schools (e.g., Catholic schools, elite private schools) and high schools that were likely to have students from small subgroups of interest (e.g., Hispanic or Black). Then, they obtained a list of all the high school sophomores and seniors at the selected schools, and chose a simple random sample of up to 18 students in each of those grades within selected schools. This is a **two-stage stratified random sample** with clustering at the high school level.

Because selection for HS&B:80 wasn't related to income (etc.), the HS&B:80 sample can be used to obtain valid inferences about the population; but, because it wasn't a simple random sample, we can't just use the sample statistics as population estimates: Due to the intentional sampling design, some subgroups of students were **overrepresented** in the sample, so the sample is a slightly distorted model of the population. Furthermore, because of the two-stage design, students from the same high school may be **more similar to one another** than they would be in a simple random sample. To get valid population inferences from the HS&B:80 sample, we have to **account for the sampling design** in our analysis.

Accounting for Sampling Design

We account for the unequal probabilities of selection into the sample by **reproportioning** the sample members to reflect the population proportions of the groups they represent: Instead of treating each sample member equally when calculating the mean (or any other statistic), we give a **higher** weight to individuals who were **less likely** to be included (from under-represented groups), and we give a **lower** weight to individuals who were **more likely** to be included (from over-represented groups). The weights counteract the unbalanced sampling probabilities, reshaping the distorted model that the sample provides to resemble the proportions of the population. Notice that for a simple random sample, every individual had the same probability of inclusion, so no weighting is necessary (their weights are all the same).

Weighting based on sample inclusion probabilities is great, if you know the probability of being included in the sample! But, despite survey designers' best intentions, people don't always respond to surveys. So, the inclusion probability doesn't depend only on factors in the study design (intentional oversampling), but also on known and potentially unknown factors that motivate someone to **respond** to the survey invitation.

Because we don't know ahead of time all these factors that affect survey response, we can **model** after the fact individuals' **propensity to respond** and use those estimated propensities as part of the weights used to correct over- and under-representation in the sample. As long as the response propensity model includes enough of the relevant variables that are related to both propensity to respond and your outcome variable, analysis weighted using these estimated response propensities can yield valid population inferences. (Refer to the section on Assumptions at the end of this guide for important discussion related to this "as long as" bit.)

Complex survey designs can also have implications for the **standard errors** of population parameter estimates. Without adjustment, most modeling approaches assume residuals are **independently and identically distributed** (IID) across the sample, which would be the case in a simple random sample. In the two-stage design of HS&B:80, however, residuals may be correlated for those attending the same high school. Because knowing something about one correlated observation tells you something about another correlated observation, having two correlated observations gives you less information than two independent observations would, and a two-stage sample contains less information than an equivalently sized IID sample. Less information means greater uncertainty in population parameter estimates, which needs to be reflected in the standard errors of estimates. One common technique for accounting for correlated residuals is to use the **clustered sandwich operator** with the Stata estimation option vce(`cluster SCHID`).

Two Methods: Explicit and Implicit

There are two main methods for reproportioning the observed sample to represent the population: inverse probability weighting and multiple imputation. **Inverse probability weighting** involves using sampling probabilities (or estimated response propensities) **explicitly** as weights in calculation of population estimates. **Multiple imputation**, on the other hand, copies patterns in the observed data to cases that are missing data, **implicitly** reproportioning the responding sample to match the correct proportions in the full sample.

Both methods can effectively reduce bias due to non-representative sampling,[1] and both methods rely on largely the same set of assumptions, differing somewhat in the type of parametric assumptions they make. Refer to the sections “Advantages and Limitations” and “Assumptions,” below, for more information on the similarities and differences in these two techniques.

Additionally, inverse probability weighting and multiple imputation can be **combined** to draw on their relative strengths. In fact, our recommended default procedure for analyzing HS&B:80 data involves using multiple imputation to adjust for differences between respondents and the full HS&B:80 panel and using inverse probability weighting to adjust for differences between the HS&B:80 panel and the population. The recommended setup is described in detail below.

What is your population?

A crucial consideration before beginning your population-representative analysis is to determine the population to which you would like to make inferences. If you are using HS&B:80 data, this population should be **similar to the study’s original population**: U.S. high school sophomores and seniors in 1980 (potentially adding the qualifications “surviving, non-institutionalized” for analysis of later-life outcomes). If your population of interest is more **general** that the original study population (e.g., “U.S. adults in their late 50s in 2021”), you will have to consider carefully whether the differences between your population of interest and the analytical population permitted by the HS&B:80 study design are minimal enough for your analysis to meaningfully answer your research question—such differences (which are not determinable from the HS&B:80 data alone) may represent substantial potential limitations to your analysis.

On the other hand, your population of interest may be more **specific** than the HS&B:80 study population. For example, you may be interested only in surviving U.S. 1980 high school sophomores and seniors who have been diagnosed with periodontal disease by 2021. Such restrictions are much more easily handled, as discussed in greater detail below.

Importantly, just as being selected for a sample does not alter the population being sampled, the presence or absence of **missing data** is generally **not** a consideration in determining your population of inference. Rather, you will use a reweighting method to

[1] Little, R. J., Carpenter, J. R., & Lee, K. J. (2024). A comparison of three popular methods for handling missing data: Complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological Methods & Research,* **53**, 1105–1135. https://doi.org/10.1177/00491241221113873

adjust for differences between your analytical sample and your intended population. The only cases in which absence of data could be an indicator of membership in your population of interest are those for which data are mechanically missing for individuals who do not have a focal characteristic, such as membership in a focal program (e.g., Medicaid, SNAP, etc.).

For example, individuals who do not have a credit card are necessarily missing the credit card balance in the consumer credit dataset (this is a **structurally** missing value, rather than a value missing due to match failure or non-response). If your population of interest is surviving U.S. 1980 high school sophomores and seniors who had a credit card in 2021, you would need to carefully distinguish sample members whose credit balance is missing due to not having a credit card from sample members whose credit balance is missing due to failure to match to the consumer credit administrative dataset. The former you exclude from your analysis, while the latter you might adjust for using a reweighting technique.

An important special case of this issue is **mortality**. Whether to exclude deceased sample members from your analysis depends on whether your intended population of interest is restricted to **surviving** U.S. 1980 sophomore and seniors. This is an analytic decision that must be grounded in your research questions and relevant theory. Because mortality is associated with many variables of general sociological interest, mortality may be a **confounding** factor in your analysis. The extent to which you can adjust for the confounding of selective mortality will depend on whether you can achieve conditional independence of mortality and your outcome, given a set of auxiliary variables associated with both.

Inverse Probability Weighting

As the name implies, inverse probability weighting involves using the reciprocal of the probability of sample inclusion as a weight in the calculation of weighted statistics as population-representative estimates of population parameters. In Stata, this usually amounts to specifying the weight variable as a **`pweight`** in your estimation command.[2]

[2] Do not use `aweight` for inverse probability weighting. The `aweight` is reserved for handling differential measurement error such as arises when your unit of analysis (i.e., rows in your dataset) consists of aggregates (e.g., county mean income) of different sizes, and the `aweight` procedure does not appropriately modify standard errors due to violations of IID assumptions.

In principle, the inverse probability weights you use should correspond to exactly the set of individuals included in your analysis, incorporating all sources of selection and non-response. In practice, inverse probability weights are usually constructed to handle a study's intentional sampling design and general unit-level non-response. If exclusions from your analytic sample due to population restrictions and item non-response are relatively small or are independent of your outcome variable, this strategy will probably yield results that are close to correct. On the other hand, if your analytic sample differs substantially from the set of weights you are using, results from inverse probability weighting could be biased, and you should either construct your own inverse probability weights or use the multiple imputation strategy described below.

Weights Available in HS&B:80/21

The HS&B:80/21 data release includes a suite of weights that can be used to analyze frequently used subsets of data from HS&B:80/21. The data release includes no less than 12 weights—each weight is tailored to analyze a different HS&B:80 subsample.

Weight Variable	**Analytic Sample**
RAWWT	Entire HS&B:80 panel
HS&B:80/21 respondents	
CROSS_SECTIONAL_WEIGHT	HS&B:80/21 respondents
PANEL_WEIGHT	All-wave respondents
PANEL_WEIGHT_MINUS_2014_2015	All-wave (other than 2014/15) respondents
COG_WEB_WEIGHT	Those with web cognitive data in 2021
COG_CATI_WEIGHT	Those with phone cognitive data in 2021
HS&B:80/21 biomarker participants	
HHV_WEIGHT	2021 home health visit participants
F6WTBLOOD	2021 blood samples
F6WTSALIVA	2021 saliva samples
F6WTGENETIC	2021 blood or saliva samples
F6WTPHARM	Those consenting to link to pharmacy records
Niche	
GRADE12_WEIGHT	HS&B:80/21 respondents who attended their 12th-grade year

The **RAWWT** accounts for the intentional HS&B:80 sampling design (without accounting for any non-response) and is applicable when using the entire panel (26,820 sample members) to estimate parameters of the population of all U.S. 1980 high school sophomores and seniors.

The remaining weights listed in the table above (except for GRADE12_WEIGHT, discussed below) are applicable when using the **subsample indicated** in the table to estimate parameters of the population of all **surviving, non-institutionalized** U.S. 1980 high school sophomores and seniors. For example, if you are analyzing all HS&B:80/21 respondents, you would use the CROSS_SECTIONAL_WEIGHT. If you are analyzing all HS&B:80/21 respondents who have web cognitive data, you would use the COG_WEB_WEIGHT. If you are analyzing individuals who responded in every wave other than 2014/15, you would use the PANEL_WEIGHT_MINUS_2014_2015.

To verify that you have selected the correct weight, create a variable that indicates whether each panel member will be included in your analysis (do this for the full panel, without removing any cases). Then, create a variable that indicates whether your selected weight is positive (i.e., greater than 0). Cross-tabulate these two indicators—if almost all cases are on the diagonal (i.e., the two indicators are equal), you have selected the correct weight. Note that anyone with a value of **zero** for a given weight will be **dropped** from your analysis if you use that weight, even if they have no missing data.

The **GRADE12_WEIGHT** differs from all other HS&B:80/21 weights in that it is used to generalize to a different population than the original HS&B:80 study population. The GRADE12_WEIGHT generalizes the sample of HS&B:80/21 respondents to the population of surviving, non-institutionalized U.S. 1980 high school sophomores and seniors **who attended their 12th-grade year**. By definition, all members of the senior cohort attended their 12th-grade year and are included in the population for this weight. The difference between the primary HS&B:80 study population and the population targeted by the GRADE12_WEIGHT is in handling of 1980 sophomores: Sophomores are included only if they were still enrolled two years later (in 1982), when they were expected to be in 12th grade. This population is not of general interest, and is primarily intended for cross-study comparisons with other NCES surveys of high school seniors (e.g., the National Longitudinal Study of 1972, National Education Longitudinal Study of 1988, etc.).

Constructing Custom Weights

A detailed discussion of considerations when constructing your own inverse probability weights is beyond the scope of this guide; however, the process generally involves the following steps for each stage of selection (i.e., intentional sampling or opportunities for non-response):

1. Calculate (from the sampling design) or estimate using a propensity score model the probability of inclusion for the current step of the process.
2. Optionally, smooth estimated propensity scores in this way:
    a. Divide the sample into quantiles based on estimated propensity scores.
    b. For each quantile group (“weighting cell”), calculate the sum of weights from the previous selection stage (or 1, if this is the first selection stage) for everyone in this cell.
    c. Calculate the sum of previous-stage weights for everyone in this cell who is included in this stage.
    d. Replace the estimated propensity scores for everyone in a given cell with the ratio of the included sum to the total sum from parts (c) and (b), respectively.
3. For individuals who are excluded in this selection stage, set the weight for this stage to 0. For individuals who are included in this selection stage, set the weight for this stage to the product of the weight from the previous stage (or 1, if this is the first selection stage) and the reciprocal of the (smoothed) selection probability.

If eligibility is unknown for a given selection stage (e.g., it is unknown whether someone who could not be located for surveying is still living, or it is unknown whether someone who was not matched to administrative records qualifies for the program of interest), treat “known eligibility” as an additional selection stage inserted before the iteration that requires the given eligibility information (e.g., use “matched to administrative records” as the inclusion characteristic in a stage before the stage corresponding to “participated in the focal program” whose value depends on data from the matched administrative records).

In this procedure, propensity score models should include **all observed variables that are associated with the selection process**. For a given selection stage, this can only be accomplished using variables **known from a previous selection stage**. For example, selection into the HS&B:80/21 follow-up (i.e., non-response) can be modeled only using information available from prior survey rounds, not data from the HS&B:80/21 follow-up itself.

Multilevel Models

Using inverse probability weights in multilevel models is more complicated than for simple statistics and ordinary regression because weights must be specified **separately for each level** of the analysis. In Stata `mixed`, the first-level (observation) weights are specified in the usual manner with the `pweight` keyword in square brackets after the first-level explanatory variables, while the second-level (group) weights are specified using the `pweight` option after the comma in the second-level section of the command (i.e., after the double bar, ||).

If you are analyzing **students within schools**, students are the first level and schools are the second level. The school design weights can be found in the variable WEIGHT in the SOPHOMORE/HSCHOOL data file in the HS&B:80 revised restricted data release (see also ICPSR 07896/DS001). The usual weights you would use in non-multilevel analyses (e.g., RAWWT, CROSS_SECTIONAL_WEIGHT) combine selection of schools and selection of students within schools. To obtain a student-level weight to use as the level-one weight in a multilevel model, divide your weight of interest by the school weight.

If you are analyzing **repeated measures** (e.g., consumer credit outcomes across multiple years) within sample members, the repeated measures are the first level and sample members are the second level. In this case, the usual weights you would use for sample members in non-multilevel analyses (e.g., RAWWT, CROSS_SECTIONAL_WEIGHT) are the second-level weight. None of the provided weights are applicable to repeated measures, so you will either need to treat the repeated measures as a simple random sample (equally weighted), potentially ignoring any non-response across repeated measures, or you will need to construct your own weights for the repeated measures within sample member. (Alternatively, use the multiple imputation-based process described below, instead.)

Stata will not automatically normalize the scale of the second-level weights, and large weights can lead to numerical issues in the maximum likelihood optimization algorithm. It is generally recommended to **rescale** your **second-level weights** manually before running the `mixed` command, by dividing the second-level weights by the mean of the second-level weights (so that the rescaled weights have mean 1).

Stata `mixed` also provides an option to **rescale** the **first-level weights** (`pwscale`). Rescaling the first-level weights may not be strictly necessary if you have separated the weight into level-one and level-two components, but using this option may resolve some convergence issues if the level-one weights are numerically large.

Refer to the Stata `mixed` manual for more information on using inverse probability weights in multilevel models, particularly the section on “Survey Data” under “Remarks and examples.”

Multiple Imputation

Both inverse probability weighting and multiple imputation are common strategies for handling missing data. Inverse probability weighting is often used to handle unit non-response, while multiple imputation is more common for item non-response. Statistically, the objective of reweighting samples to yield population-representative can be framed as a missing-data problem. As such, both inverse probability weighting and multiple imputation are effective strategies for achieving population-representative estimates. This section describes the multiple imputation-based strategy that we recommend for analyzing HS&B:80 data. The section on Advantages and Limitations at the end of the guide describes the relative merits of inverse probability weighting and multiple imputation and explains why the strategy described here is our general recommendation.

This strategy actually blends inverse probability weight and multiple imputation to draw on their relative strengths. Inverse probability weighting excels at adjusting for known differences in inclusion probability, such as for intentional oversampling in a complex survey sampling design, while multiple imputation provides substantial flexibility in tailoring implicit weights to handle non-response, customized to each analysis. Accordingly, our strategy uses multiple imputation to adjust for differences between your analytic sample and the (possibly restricted) HS&B:80 panel due to non-response, combined with inverse probability weighting by the full panel RAWWT to adjust for intentional differences between the HS&B:80 sampling design and the population.

The general procedure has the following steps:

1. Determine the **exclusions** from the original HS&B:80 study population necessary to achieve your intended inferential population (refer to the section on "What is your population," above).
2. Assemble the **variables** to include in your imputation model.
3. **Generate** multiple imputation datasets.
4. **Analyze** the multiply imputed datasets using standard multiple imputation techniques, weighted with **RAWWT**.[3]
5. Check that the **number of imputations** is adequate.

[3] We also recommend weighting the imputation models with RAWWT. Since the main stratification variables used in the sampling design are included in our recommended minimal set of imputation variables, weighting the imputation models may not be strictly necessary and may not make much difference in practice. But, weighting the imputation models may be a prudent precaution.

These steps are described in greater detail below, with recommendations specific to analysis of HS&B:80; however, this guide is not a comprehensive introduction to multiple imputation. Refer to the documentation for Stata `mi` and Stef van Buuren's *Flexible Imputation of Missing Data* for more-general introductions to multiple imputation.

Am I Making Up Data?

Mechanically, the multiple imputation procedure involves plugging holes in a dataset with concrete numbers. Sometimes, newcomers to multiple imputation wonder if this means that they are "making up" data. This is **not at all** the case.

First, the "multiple" part of multiple imputation is critical. You are not making up a single number to stand in for a missing value, as if you were pretending that the number you make up actually represents the missing value. Rather, you are assuming that the unknown value comes from a range of possible values, governed by a probability distribution conditional on all the other information you have at your disposal. When you invoke multiple imputation, you are actually using a standard numerical technique called **Monte Carlo integration** to approximate the **expectation** of your analytic estimates over the entire **joint distribution** of the missing values in your dataset. You have not made anything up because you do not assert any single value for missing data, but rather allow for a range of possibilities for what you don't know, constrained by what you do know.

Second, multiple imputation does not force a specific distribution on the unknown values (aside from a broad functional form, as described in the Assumptions section, below), as if you were nudging the data in the direction you wanted. Rather, multiple imputation takes the **patterns** of association found in the **observed data** and **replicates** them across the unobserved data. The cases with missing data do not alter these conditional patterns—the imputed data only reflect whatever relationships are present in the observed data. The purpose of including the cases with missing data is to ensure that the groups they represent in the population are included in the correct proportions after dropping (marginalizing) the auxiliary variables used in the imputation.

Third, you are not "cheating" by artificially inflating your sample size with made up data. Multiple imputation is carefully designed to **track all sources of uncertainty**—this includes the uncertainty of the imputation process itself, given its assumptions. While adding more cases (with known values) decreases uncertainty, if additional cases have missing data, the uncertainty of the missing values will prevent the added cases from artificially deflating the standard errors of your estimates, because the multiple imputation calculations will counter balance the larger sample size (i.e., apparently greater certainty) with the larger imputation uncertainty. Refer to the section on "Reporting Sample Sizes," below, for more information about how to communicate your sample size accurately when using multiple imputation.

Cases to Include

As mentioned in the section "What is your population," above, you should first determine how your intended population of inference differs from the original HS&B:80 study population (U.S. 1980 high school sophomores and seniors). In particular, if you are analyzing midlife outcomes, you will need to consider whether you are making inferences to the **surviving** population (as well as implications for possible violations of assumptions if you are attempting to make inferences about the full original population). You may have other population **exclusions** related to your research question (e.g., you may be making inferences specifically about Bachelor-degree-holders, those who have heart disease, those who do not have chronic kidney disease, those who have had a credit card, etc.).

In general, you will always analyze the **full HS&B:80 panel** minus your intentional population exclusions. To reiterate, these exclusions are based **only** on your desired population of inference, not on non-response or missing-data considerations.

If you **know the exclusion status** for everyone in the panel, you may **delete** the excluded individuals from your dataset before generating your multiple imputations.

If you have **missing values** in the exclusion criteria, you will need to impute inclusion status along with your other variables. In this case, do **not** delete cases you know you will later exclude, since you will first need them to impute the unknown eligibility status for other cases. Refer to the section "Unknown Eligibility," below, for more information about how to handle this scenario.

If you plan on making inferences to **multiple sub-populations**, do **not** create multiple datasets for each sub-population. Instead, keep everyone who will appear in at least one of your analyses in the dataset you impute, and use `if` statements at the analysis stage to subset your single imputed dataset to the sub-population of interest for each model.

Variables to Include

Create a **single, unified dataset** per project/paper with all the variables you will need for multiple imputation and your analysis, even if you are analyzing multiple outcomes or multiple sub-populations. If you are analyzing **time series** data, reshape the data into **wide format** for imputation (one row for each sample member, one variable for each measurement); you may reshape the data into long format **after** imputation, if long format is more convenient for your analysis.

Your imputation model should include:

- All outcomes you plan on analyzing
- All exposures you plan on analyzing
- Anything that is plausibly related to non-response/missingness
- Anything that is plausibly associated with your outcomes
- Anything that is plausibly associated with your exposures
- Any control variables you plan on including in your analytic models
- Any interaction terms you plan on including in your analytic models, including interaction terms for any stratified analyses (i.e., models estimated by subgroup)
- Any measurement of the included variables at other points in time (either earlier **or** later than your focal measurement)

The final list will include variables that you do not plan to analyze—you should still include these variables in your imputation model, even if you do not include them in your analysis. Variables included in an imputation model but not in analytic models are called **auxiliary variables**. Auxiliary variables are very important for reducing uncertainty (increasing statistical efficiency/precision) and making the conditional independence assumption of multiple imputation more plausible (refer to the section on Assumptions, below).

Based on the considerations above, we have created the following **minimal list of recommended variables** to be included in all HS&B:80 imputations (in addition to the variables specific to your analysis). This list includes the main stratification variables in the original HS&B:80 sampling design, as well as variables included in the propensity score model used in development of the inverse probability weights in the HS&B:80/21 data release. Including these variables should ensure that the non-response bias reduction you achieve via multiple imputation will be at least as good as the bias reduction you would achieve using the provided weights.

- Cohort, race, ethnicity, sex, high school disability
- High school region, urbanicity, type
- High school achievement scores (at least math in grade 12), GPA, math coursetaking
- Parental education, high school family income, high school family owned home

- Educational attainment, employment status, marital status
- 2021 general cognition, self-reported health
- 2021 Kessler score, science knowledge score

As mentioned above, when these variables are measured at multiple points in time (e.g., employment status), you should include all measurements. If you are not analyzing midlife outcomes, you could probably leave out the 2021 outcomes from your imputation model, unless they are related to your outcome of interest (e.g., educational attainment in early adulthood).

Impute the Outcome?

The guidelines above indicate that you should use imputed outcome values in your analysis—in some scenarios, you may even have more imputed outcome values than observed outcome values (refer to the section "Am I Making Up Data," above, for reassurance that this is not evil). This may feel uncomfortable as it runs counter to common advice either not to impute cases with missing outcomes or to impute and then discard such cases. However, the older advice, while not necessarily misguided, is **insufficiently nuanced** to meet the needs of population-representative estimation in a complex survey, such as HS&B:80.

Arguments against imputing the outcome variable (by deleting cases before imputation) point out that cases with no outcome value provide no information about the relationship between the outcome and your exposure of interest. This is, strictly speaking, true: Cases missing the outcome variable indeed do not provide usable information for estimating patterns of association with the outcome. However, their role in the strategy presented here is not to provide additional information about the relationship between outcome and exposure; rather, their role is to ensure that subgroups are **correctly proportioned** so that the sample reflects the population.

Moreover, when first proposing the impute-then-delete strategy,[4] von Hippel argued that including imputed outcome values simply adds noise to estimation of the analytic model, and thus these cases yield more trouble than value. However, von Hippel was primarily considering scenarios with a very small number of imputations; with a larger number of

[4] Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology,* **37**, 83–117. https://doi.org/10.1111/j.1467-9531.2007.00180.x

imputations, **noise becomes negligible**, and including multiply imputed outcomes does no harm.

There are, in fact, scenarios in which failing to impute and include cases with missing outcomes will lead to **biased estimates**: namely, when the imputation model includes **auxiliary variables** that are correlated with the outcome.[5] Given the recommended minimal list of variables to include in multiple imputation for HS&B:80, you will almost always have such auxiliary variables in your imputation model, so removing cases with missing outcome variables runs the risk of biasing your estimates.

Unknown Eligibility

The only panel members that should be excluded from your analysis are individuals who are not members of your population of interest, when your population of interest is more restricted than the original HS&B:80 study population. However, this requires that you know whether panel members have the necessary characteristics, which may depend on data with missing values. For example, you may want to exclude individuals with chronic kidney disease, but this condition is known only for respondents to the HS&B:80/21 follow-up survey.

In such scenarios, you will need to **impute a flag for membership** in your restricted population. To do so, create a binary variable that is 1 for all panel members known to be in your restricted population, 0 for all panel members known **not** to be in your restricted population, and missing for all panel members whose membership status is not fully known. Then, using the full panel, impute this variable along with all other variables in your imputation model, as you would any other binary variable.

If you are concerned that relationships in your imputation model may differ for your restricted population, compared with those outside your restricted population, or if relationships make no sense for individuals outside your restricted population (e.g., maternity variables when working with the female sub-population), you will want to estimate the imputation model only for members of your restricted population, rather than the entire panel. This means that your imputation of some variables depends on the imputation of the membership flag. This internal dependency requires special handling called **conditional imputation**.

---

[5] Sullivan, T. R., Salter, A. B., Ryan, P., & Lee, K. J. (2015). Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data. *American Journal of Epidemiology*, **182**, 528–534. https://doi.org/10.1093/aje/kwv100

To set up conditional imputation with Stata `mi`, you should first **fill in** missing values for conditionally imputed variables with 0 for just those cases who are **known to be excluded** from your restricted population. Leave the missing values for cases that have unknown eligibility or that are known to be included. Then, use the `conditional` option to specify an `if` statement based on your imputed membership flag for the portions of the `mi impute` command that require conditional imputation, for example:

```
mi impute chained (logit) mypop (reg) `uncond_vars'     ///
    (reg, conditional(if mypop == 1)) `cond_vars'       ///
    = `aux_vars' [pw=RAWWT], add(20)
```

where `mypop` is the membership flag, the macro `uncond_vars` lists variables that do not need conditional imputation and will be estimated using the full panel, and the macro `cond_vars` lists variables that should be imputed only for members of the restricted population.

When using conditional imputation with an imputed membership flag, it is very important to ensure that the membership flag is **imputed before** the variables that are imputed conditional on membership. When Stata `mi impute chained` lists the models in the chain, verify that the model for the membership flag outcome is listed before models for the conditionally imputed variables. If it is not, stop the imputation, and specify the order of variables for the imputation manually with the option `orderasis`.

Finally, after imputation, use an `if` statement to **restrict estimation** of your analytic models to cases where the (imputed) membership flag is 1. Cases with unknown eligibility will be included for some imputed datasets, but not for others, which results in a variable analytic sample size across imputed datasets (refer to the section on "Reporting Sample Sizes," below, for more information). Stata `mi est` will complain about this unless you supply the **option `esampvaryok`** to reassure `mi` that this is intentional.

Number of Imputations

Early work in multiple imputation suggested that only a few imputations were necessary for valid results; however, the number of imputations required depends on the fraction of missing information.[6] In many HS&B:80 analytic scenarios, you may have a large number of cases with missing data due to non-response in later data collection periods, which may

[6] 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. https://doi.org/10.1002/sim.4067

require a much **larger number of imputations** than is often seen in other studies (e.g., 50-100 imputations, or possibly more). We recommend using estimates of the **Monte Carlo error** to determine whether you have used a sufficient number of imputations.

Monte Carlo error refers to the variability in your estimates due not to survey sampling or variability in the population, but rather due to the imputation of missing values. Multiple imputation uses an approximation of the multidimensional integral needed to compute the expectation of your estimates over the joint distribution of the missing data. The more imputations you use, the more accurate the approximation becomes. The Monte Carlo error is an indication of how much variability there is in your model parameter estimates across different runs of the multiple imputation procedure for a given number of imputations. The goal is to increase the number of imputations until the imprecision of the multiple imputation approximation is smaller than the level of precision that you care about (e.g., in decimal places that you won't report).

There is a separate Monte Carlo error estimate for each statistic you calculate (each model parameter, each test statistic, each p-value, etc.). You do not necessarily need to worry about the Monte Carlo error for everything that Stata might report—focus on the Monte Carlo error of the quantities that you care about and will interpret (e.g., regression coefficients of interest, p-values for statistical tests that you report, etc.). We suggest two rules of thumb for checking Monte Carlo error levels:[7]

- For regression **coefficients**, the Monte Carlo error for the coefficient should be **less than 10 percent** of the coefficient's standard error.
- The distance from **p-values** to your significance level (α) should be **more than twice** the p-value's Monte Carlo error.

If you find that the Monte Carlo error is too high for a quantity of interest, simply generate additional imputations (you can re-use the original imputed data to save computation time), re-estimate with the larger number imputations, and re-check the updated Monte Carlo error. The Monte Carlo error should decrease roughly with the square root of the number of imputations, which can be helpful for determining how many additional imputations to generate. Note also that the recommendations above are general guidelines, rather than strict rules.

To obtain estimates of the Monte Carlo error of analytic results in Stata, add the **`mcerror` option** to the `mi est` command. This will generate two rows for every row in the usual

[7] *Ibid.*

output of your estimation command. The first row of the pair will be the usual output; the second row will be the Monte Carlo error corresponding to each statistic on the preceding row. To find the Monte Carlo error for a regression coefficient, locate the regression coefficient estimate in the output and look directly below the coefficient estimate (you will compare this Monte Carlo error with the coefficient's standard error, which is on the first row—same as the coefficient estimate—but in the next column over). To find the Monte Carlo error for a p-value, locate the p-value in the output and look directly below the p-value. You may wish to compare the output with and without the `mcerror` option to verify that you are reading the output correctly.

It is not generally necessary to report the Monte Carlo error for every statistic. Rather, it may be sufficient simply to report "**We used XX imputations to reduce Monte Carlo error to negligible levels**," optionally giving the criteria listed above if editors/reviewers request specificity.

Reducing the Monte Carlo error to acceptable levels may take a large number of imputations, but using a large number of imputations can be slow and cumbersome. We recommend using a **modest number of imputations (e.g., 10-20) during model development**. Once you have finalized your analytic code, increase the number of imputations for a final run before checking the Monte Carlo error and pasting the results into your manuscript. Note that the apparent significance of marginal statistical tests may change when you increase the number of imputations, if the Monte Carlo error with the smaller number of imputations was large enough to push the p-value to the wrong side of your significance level—treat marginal p-values very cautiously until you have finalized the number of imputations. Marginal p-values will require more imputations than statistical tests that have strong evidence against the null hypothesis.

Diagnostics

Multiple imputation via chained equations depends on the specification of a model for each imputed variable—for valid imputations, these models must fit the data well. Best practice is to check the fit of the imputation models, though few researchers seem to report doing so. To check **model fit**, after imputation, simply re-estimate the imputation model (Stata `mi impute` lists these at the start of its output) using the imputed data, and check model fit as you would for any model (e.g., examine a plot of residuals against predicted values). If time is limited, you could focus on the most important variables (e.g., your outcomes and exposures of interest).

In imputation via chained equations, each imputed dataset is the end of a sequence of estimations of the models for each imputed variable, taken in turn, forming a Markov chain. In principle, this Markov chain must reach **stationarity (convergence)** for valid imputations. If the Markov chain has not reached stationarity, the **length of the chain** should be increased (`mi impute chained` **option `burnin`**). Few researchers using multiple imputation seem to check this, though researchers taking a fully Bayesian approach are much more used to examining the dynamics of their Markov chains. The `chainonly` and **`savetrace` options** can be used to store summary statistics from the Markov chain for inspection. Line plots of the summary statistics should converge to approximately flat (though noisy) lines, with no upward or downward trend; an apparent trend in the first few iterations is not problematic, but if the latter iterations do not clearly show the absence of a trend, increase the number of iterations and re-inspect the traces.

Reporting Sample Sizes

In the early days, sample sizes could be used directly to estimate the uncertainty in simple statistics, such as the sample mean or sample proportion. However, with complex statistical models and complex sampling designs, specially computed standard errors for each estimand are preferred for communicating levels of uncertainty, rather than the sample size itself. Nevertheless, the sample size remains a common rough indicator of the amount of information used in the analysis.

The multiple imputation-based strategy presented in this guide makes use of the full HS&B:80 panel (minus population exclusions) for every analysis, but this does not mean that you have 26,820 pieces of information available for every model, because cases that are missing the outcome provide no information about conditional relationships of interest. In the spirit of using the sample size to index the amount of information available, report instead the **number of observed outcome values** used in the analysis. If your exposure of interest has substantially more missing data than your outcome, report the number of observed exposure values, instead.

In the methods section, report the size of the full HS&B:80 panel minus population exclusions where you describe your multiple imputation procedure. If you report sample sizes separately for various models (e.g., with different outcomes or for different subgroups), clearly indicate in the methods section that the sample size reported for each model is the number of observed outcome values used for that model.

Illustrative Stata Code

```
** Categorical auxiliary variables: Code missing category
foreach v in marital86 marital92 marital14 employ14 health14 distress15 {
    replace `v' = 0 if missing(`v')
}

* For conditional imputation of income82, fill in 0 for seniors
replace income82 = 0 if cohort == 1  // Seniors


** Post-secondary education variables

* Auxiliary variables
local pse_aux  psetranso psetransr pse82sr edattain92so ///
               pse14so pse15sr pse84sr pse86sr pse84so pse86so

* Code missing category for auxiliary PSE variables
foreach v of local pse_aux {
    replace `v' = 0 if missing(`v')
}


** Variables for imputation (customize for your analysis)

* Binary variables: logistic regression models
local imp_var_logit    home physdisab learndisab

* Multi-category variables: multinomial logistic regression models
local imp_var_mlogit   pared income80 highmath employ21 marital21 health21

* Linear regression for variables with Normally distributed residuals
local imp_var_reg      gencog21 kessler21 sciknow21

* Predictive mean matching for variables with troublesome distributions
local imp_var_pmm      gpa12

* Variables that require special settings during imputation
local imp_var_special  edattain21 income82

* Auxiliary variables that do not require imputation
local aux_var_cat      cohort sex race marital86 marital92 marital14  ///
                       employ14 health14 distress15
local aux_var_cont     math12 vocab12 read12
```

```
* Auxiliary post-secondary education variables
local pse_vars
foreach v of local pse_aux {
    * Accumulate i. operator for each auxiliary pse* variable
    local pse_vars `pse_vars' i.`v'
}

* Combine all imputed variables into a single convenience macro
local imp_var_all  `imp_var_logit' `imp_var_mlogit' `imp_var_reg'  ///
                   `imp_var_pmm' `imp_var_special'

* Generate missing variable indicators
misstable summarize `imp_var_all', generate(miss_)


** Configure imputation
mi set wide
mi register imputed  `imp_var_all'


** Impute
//    Note: Set a random number generator seed where indicated below
mi impute chained  ///
    (logit) `imp_var_logit' ///
    (mlogit) `imp_var_mlogit'  ///
    (reg) `imp_var_reg'  ///
    (pmm, knn(10)) `imp_var_pmm'  ///
    /* Variables with special imputation settings: */  ///
    (ologit, include(`pse_vars')) edattain21  ///
    (mlogit, conditional(if cohort == 0)) income82  ///
    = i.(`aux_var_cat') i.marital86#i.cohort `aux_var_cont'  ///
    [pw=RAWWT], add(20) dots chaindots augment rseed(/*SEED*/)

* Example of regression estimation
mi estimate, dots mcerror post:  ///
    reg gencog21 i.race [pw=RAWWT], vce(cluster SCHID)
```

Assumptions

Both inverse probability weighting and multiple imputation share the crucial assumption that **selection into the sample** (via both sampling and non-response/missing data processes) is **conditionally independent of the outcome variable**, given a set of observed variables—namely, the variables used either in the propensity score model for inverse probability weighting or in the imputation model for the outcome variable in multiple imputation. Since researchers many times use pre-computed weights constructed by others and since the propensity score model is not directly related to the outcome variable, this assumption may appear more hidden with inverse probability weighting, compared with multiple imputation, in which the conditional relationship between the outcome variable and the auxiliary variables is more explicit. But, the assumption applies equally to both methods.

You will not have observed every variable related to the selection mechanism. But, you may have observed variables that are correlated with important unobserved variables. By including an expansive set of auxiliary variables, you have likely included some observed variable that can act as a proxy for components of the unobserved selection mechanism, increasing the plausibility of the conditional independence assumption. Your population estimates may be biased to the extent that your included variables are **not** correlated with some unobserved variable that is associated with both selection and your outcome

In addition to the conditional independence assumption, both inverse probability weighting and multiple imputation share their own version of a **functional form** assumption. Inverse probability weighting assumes that the propensity score model is a good representation of the actual selection mechanism. If there is no non-response, the sampling process is known exactly, and this assumption is met automatically.

Multiple imputation assumes that the imputation model is a good representation of the conditional distribution of each variable—this includes the functional form of the conditional mean given the other variables in the imputation model, as well as the shape (family) of the residual distribution. The non-parametric technique **predictive mean matching** can reduce assumptions of the distribution of residuals.[8] Because multiple imputation involves specifying conditional distributions for many different variables, the functional form assumption could be considered more substantial for multiple imputation than for inverse probability weighting. Using chained equations for multiple imputation has the further elaboration of the functional form assumption that the univariate conditional distributions must be **compatible** with one another to represent a valid multivariate distribution, though the results of analysis of multiply imputed datasets may be robust to

[8] Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of Educational and Behavioral Statistics,* **42**, 371–404. https://doi.org/10.3102/1076998616687084

some forms of incompatible univariate models.[9] Using a **sequential modeling approach**[10] in a fully Bayesian framework maintains the modeling flexibility of chained equations while removing concerns over conditional model incompatibility (see, e.g., Blimp).

Posthumous imputation?

As the cohort ages, mortality increasingly becomes an important consideration in population inference. Risk of death is not evenly distributed across the population, so changes over time in group differences in an outcome may be confounded with differential mortality across groups, rather than reflecting within-individual changes.

One might be tempted to "factor out" mortality to obtain parameter estimates for the population as it appeared before the deaths occurred. In a multiple imputation context, this involves imputing outcomes for those who have died; in an inverse probability weighting context, this involves treating mortality as a selection stage. There is nothing strictly wrong with this approach; however, it still requires the same conditional independence assumption: The conditional distribution of the outcome variable must be the same for both those who died and those who survived, given the available variables—those who died and those who survived must in some sense be the "same."

However, decedents died for some reason, and (bizarre accidents aside) it seems very likely that whatever those reasons may be (things such as poor health, morbidity risk, restricted access to health care, etc.), they differentiate those who died from those who survived. Certainly, it should be possible to account for some of the differences in correlates of mortality risk with available data; but, it remains a significant and open question whether the available data cover enough of unobserved mortality risk to render the conditional independence assumption plausible. To reiterate, this concern applies no matter whether you are using multiple imputation or inverse probability weighting.

---

[9] 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**, 1049–1064. https://doi.org/10.1080/10629360600810434

[10] Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods,* **25**, 157–181. https://doi.org/10.1037/met0000233

Advantages and Limitations

Both inverse probability weighting and multiple imputation can reduce bias in estimates of population parameters when survey samples suffer from non-response. Inverse probability weighting with pre-computed weights is straightforward and computationally efficient. However, the pre-computed weights only cover unit non-response, not item non-response. Moreover, they are applicable primarily only to one sub-population of interest (surviving, non-institutionalized U.S. 1980 high school sophomores and seniors). If your analytic sample differs from the weighted sample (either due to item non-response or due to additional restrictions to the inferential population), your estimates may still be biased if the additional analytic sample restrictions are correlated with your outcome. Pre-computed weights also include only a limited set of variables in the response propensity model, which may not include key variables relevant for your outcome.

You may already be using multiple imputation to handle missing data in your covariates; if so, very little additional work is needed to leverage multiple imputation for population representativeness as well. By shifting from inverse probability weighting to multiple imputation, you gain the flexibility of including all the observed variables you have available to render the conditional independence assumption as plausible as possible. This means that your strategy for achieving population representative estimates can be easily and flexibly customized for each analysis, rather than relying on the imperfect match of the design of general-purpose pre-computed weights. Moreover, in some cases, estimates via multiple imputation can be more statistically efficient (i.e., smaller standard errors) than estimates via inverse probability weighting,[11] namely when auxiliary variables provide explanatory power for the outcome. Inverse probability weighting, on the other hand, can sometimes degrade statistical efficiency (i.e., inflate standard errors) when the propensity to respond is low, such as when there are many selection stages leading to widely varying weights.

Multiple imputation is, of course, much more computationally intensive than inverse probability weighting, and the distributional functional form assumptions are more extensive for multiple imputation than for inverse probability weighting. But, if you are already using multiple imputation to handle missing data in your covariates, you have already accepted these costs, to some extent. We suggest that the additional costs

[11] Little, Carpenter, & Lee (2024), *op. cit.,* Table 1.

associated with using multiple imputation to handle population-representative estimation are relatively minor compared with its flexibility and potential gains in statistical efficiency.

To reiterate, both inverse probability weighting and multiple imputation require the crucial assumption that non-response is conditionally independent of the outcome, given the available observed variables. While multiple imputation may make this assumption feel more pressing by explicitly stating a model for the outcome conditional on auxiliary variables, the assumption applies equally well to inverse probability weighting: Neither technique can remove bias due to unobserved factors associated with both the outcome and the response process.

How to choose?

Given its flexibility, we recommend using the multiple imputation strategy presented here for most "production" analyses of HS&B:80 data. For a "quick peak" or initial exploratory work, using the pre-computed weights can save time and effort while approximately removing some non-response bias. For the "development" work of writing error-free analytic code and refining model specifications, we recommend using multiple imputation with a modest number of imputations (e.g., 20). Then, increase the number of imputations in one final (long) run before pasting the results into your manuscript, after checking the Monte Carlo error to ensure the adequacy of the number of imputations.

Center for Demography and Ecology
University of Wisconsin–Madison
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400