Received: 2 November 2021 | Revised: 17 January 2023 | Accepted: 21 January 2023

DOI: 10.1002/icd.2407

METHODOLOGICAL ARTICLE

WILEY

Best practices for addressing missing data through multiple imputation

Adrienne D. Woods[1] | Daria Gerasimova[2] | Ben Van Dusen[3] | Jayson Nissen[4] | Sierra Bainter[5] | Alex Uzdavines[6,7] | Pamela E. Davis-Kean[8] | Max Halvorson[9] | Kevin M. King[9] | Jessica A. R. Logan[10] | Menglin Xu[11] | Martin R. Vasilev[12] | James M. Clay[13] | David Moreau[14,15] | Keven Joyal-Desmarais[16,17] | Rick A. Cruz[18] | Denver M. Y. Brown[19] | Kathleen Schmidt[20] | Mahmoud M. Elsherif[21]

Correspondence
Adrienne D. Woods, SRI International, 1100 Wilson Blvd, Suite 2800, Arlington, VA 22209, USA.
Email: adrienne.woods@sri.com

Abstract

A common challenge in developmental research is the amount of incomplete and missing data that occurs from respondents failing to complete tasks or questionnaires, as well as from disengaging from the study (i.e., attrition). This missingness can lead to biases in parameter estimates and, hence, in the interpretation of findings. These biases can be addressed through statistical techniques that adjust for missing data, such as multiple imputation. Although multiple imputation is highly effective, it has not been widely adopted by developmental scientists given barriers such as lack of training or misconceptions about imputation

The foundation for this paper was created during a 'hackathon' session occurring on 23 June 2021, at the annual virtual meeting of the Society for Improving Psychological Science. We invited anyone interested in the topic to attend, welcoming both experts and those with little experience addressing missing data in their research, specifically welcoming participation from those who were not sure how to address the missing data they experienced. Decisional guidelines for analyzing the type and extent of missing data were then crowdsourced and curated during this hackathon, resulting in a missing data and multiple imputation decision tree (Woods et al., 2021, available at https://doi.org/10.31234/osf.io/mdw5r) and a companion infographic (Woods & Schmidt, 2021, available at https://miro.com/app/board/o9_J18JGJQk=/). We also created multiple imputation coding templates for several prominent software languages (Stata, Mplus, R, SPSS, SAS and Blimp). All hackathon materials and coding templates are available at https://osf.io/j3f8m/.

For affiliations refer to page 30

methods. Utilizing default methods within statistical software programs like listwise deletion is common but may introduce additional bias. This manuscript is intended to provide practical guidelines for developmental researchers to follow when examining their data for missingness, making decisions about how to handle that missingness and reporting the extent of missing data biases and specific multiple imputation procedures in publications.

KEYWORDS
development, missing data, missingness mechanisms, multiple imputation, open scholarship

1 | INTRODUCTION

Adequately addressing missing data is a common challenge in the developmental sciences. Multiple imputation is a feasible, credible and powerful approach to handling missing data that helps reduce bias in several scenarios (Enders, 2017). Multiple imputation attempts to minimize the impact of attrition or non-response bias on the analysis by using available information about individuals to adjust the parameter estimates. Using multiple imputation thus approximates what results would look like with complete observations while allowing for representation of uncertainty in the results and maximizing the data set's statistical power (see Box 1 for an overview) (Cheema, 2014; Dong & Peng, 2013).

Table of Acronyms and Definitions

Name	Acronym	Definition
Auxiliary variables		Variables that researchers include in the imputation model (but not the analytic model) because they are either correlates of missingness or correlates of an incomplete variable. This helps to account for the missingness of variables directly related to the research question(s) (Collins et al., 2001; Enders, 2010).
Burn in iterations		Discarding the first *N* samples, with *N* being chosen to be large enough that the chain has reached its stationary regime by this time. The default in the mice package is 5000 (van Buuren, 2018).
Complete case analysis		A procedure that removes participants with any missing information from the analysis. Also known as 'listwise deletion' (van Buuren, 2018).
Convergence		Occurs for a test statistic when the multiple imputations of that test statistic overlap (e.g., they do not diverge or run in parallel) around a consistent value (e.g., they do not tend to increase or decrease; see van Buuren, 2018 for examples). Researchers can diagnose convergence for a test statistic as occurring when the variance between different imputations is no larger than the variance within each individual imputation (van Buuren, 2018).
ECLS-K 2011		Early Childhood Longitudinal Study, Kindergarten Cohort of 2010–2011 (Tourangeau et al., 2015).

Name	Acronym	Definition
Fully conditional specification	FCS	Another term for 'multiple imputation by chained equations' (MICE; van Buuren, 2018).
Full information maximum likelihood	FIML	An approach to handling missing data that computes a casewise likelihood function using only those variables that are observed for each case (Enders, 2010). FIML is embedded into the estimation process and can be described as 'implicit imputation', as the technique creates temporary imputations during the estimation process (Widaman, 2006).
Intraclass/intracluster correlation coefficient	ICC	A statistic that describes the degree of dependence between the observations taken on a specific unit/patient/participant within the same group/cluster. The values range between 0 (i.e., weak within-cluster correlation) and 1 (strong within-cluster correlation) (Katzmarzyk et al., 2022).
Listwise deletion		A procedure that removes participants with any missing information from the analysis. Also known as 'complete case analysis' (van Buuren, 2018).
m		The number of imputed data sets generated in a multiple imputation procedure.
Missing at random	MAR	Situations when missing data are generated in a systematic manner that can be fully accounted for using information contained within a data set (Bhaskaran & Smeeth, 2014).
Maximum iterations	Maxit	The number of iterations beyond the burn in iterations used for each imputation in MICE. The plots of the iterations inform if the imputation achieved convergence (Oberman et al., 2021).
Model-based imputation	MBI	In MBI, one first specifies their intended analytic model. The MBI procedure then creates *m* multiply imputed data sets that are tailored to this model. One can analyze the imputed data sets using the specified model or a model that is nested within the specified model (Keller & Enders, 2021).
Missing completely at random	MCAR	The likelihood of any given data point being missing is the same across all data points and unrelated to any other measured or unmeasured variables (Bhaskaran & Smeeth, 2014).
Multiple Imputation	MI	Existing data is used to generate multiple (*m*) data sets of plausible values for missing data that each incorporate random components to reflect the uncertainty of these values. Each data set is then analyzed individually according to a common statistical model, and parameter estimates are pooled into one set of estimates, variances, and confidence intervals (van Buuren, 2018). See Box 1.
Multiple imputation by chained equations	MICE	A multi-step process to create each imputed data set. The steps and an example are laid out in Azur et al. (2011); see also van Buuren (2018).
Missing not at random	MNAR	Situations when missing data occur in a way that we cannot fully account for through measured data (Bhaskaran & Smeeth, 2014).
Planned missing design		A data collection design in which the researcher randomly assigns certain participants to be missing observation occasions or measurements to minimize research costs and participant burden. Because the missing values are MCAR given the random assignment, they can be imputed without bias or auxiliary variables during analysis (Graham et al., 2006; Rhemtulla & Hancock, 2016; Rhemtulla & Little, 2012; Wu & Jia, 2021)
Pairwise deletion		To only use a participant's information when they offer complete data for a given analysis. This approach is less restrictive than listwise deletion (van Buuren, 2018).

(Continues)

Name	Acronym	Definition
Pooling		See Multiple Imputation (van Buuren, 2018).
Predictive Mean Matching	PMM	PMM uses regression models (linear, logistic, or multinomial, depending on the variable) to find the user-specified number of nearest observed cases that most closely resemble the predicted values of the respondents with missing data rather than imputing random values from the conditional distribution. This results in imputed values that are actually observed in the data set and that are more robust to violations of normality than other approaches (i.e., *regress*, *logit* and *mlogit*) (van Ginkel et al., 2020).
Seed value		An integer that offsets the random number generator in model estimation. Setting a seed value in generating multiple imputations will make the multiple imputation analysis reproducible, assuming the data and other parameters (e.g., iterations, *m*, auxiliary variables) are the same.

Yet, despite its benefits, developmental scientists have been slow to adopt multiple imputation. Many scientists perceive barriers to both understanding and implementing multiple imputation including uncertainties about when it is appropriate to use multiple imputation, and concerns that multiple imputation is 'making data up' (Nguyen et al., 2021; Rombach et al., 2018; White et al., 2010). In addition, researchers often find that lower-quality methods for handling missing data are both easy to use and still readily accepted by many developmental scientists. Developmental scientists might be more willing to overcome these barriers if they had good examples of multiple imputation that they could apply to their own work. Unfortunately, few practical examples have been offered using the complex data and analyses commonly encountered in developmental research, such as a multilevel data structure and analysis using multilevel or growth curve models.

Our aim is to provide a set of decision points to address this gap. We are basing these decision points on prior work detailing best practices for addressing missing data through multiple imputation. Similar to work on best practices in preregistration (van den Akker et al., 2021) and open science (Adelson et al., 2019), we hope this paper demystifies the process of understanding and applying multiple imputation. We provide a practical guide for authors, reviewers and editors, and include recommendations for the information that should be included in peer-reviewed manuscripts and their supplements.

We begin with a brief overview of why developmental scientists should adjust for missingness in quantitative analyses, including discussions of common barriers to adopting best practices for handling missing data, misconceptions of employing multiple imputation, and the implications of failing to adjust for missing data in developmental science. Next, we review the mechanisms that lead to missingness and the multiple imputation model. We conclude with a worked example of missing data analysis and multiple imputation using complex data and analyses to match the kind of work done by developmental scientists. This example uses publicly available data from the Early Childhood Longitudinal Study, Kindergarten Cohort of 2010–2011 (ECLS-K: 2011; Tourangeau et al., 2015). Though this worked example will be particularly helpful for developmental scientists, we hope to persuade all quantitative researchers to consider the implications of missing data and more appropriately adjust for missing data in their research.

2 | WHY IS APPROPRIATELY ADDRESSING MISSING DATA IMPORTANT?

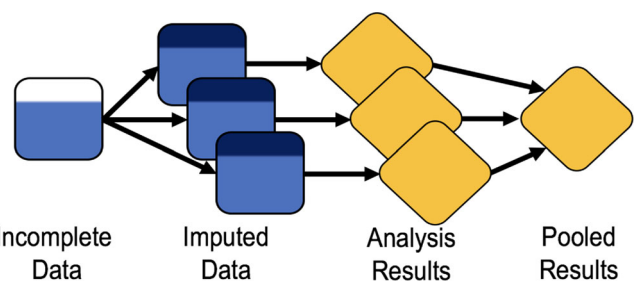
Missing data have been described as the norm rather than an exception in quantitative research (Dong & Peng, 2013). Missingness can occur when a participant disengages with a task before completing enough items or trials for a reliable answer (e.g., skips parts of a questionnaire or stops responding during a task measure), misses measurement occasions

BOX 1 An Overview of Multiple Imputation

Following data collection, several strategies may be used to handle missing data. The correct choice depends on the context of the analysis (see Supporting Information: Table A1 for a summary of these strategies). However, before building any missing data models, you should think about the missingness mechanism(s) in your data set. Why are your data missing? You should visualize and summarize the missing data patterns to develop your hypotheses. Are specific measures, items, time points missing, or does missingness vary by cluster or site, or reporter? Does this fit with MCAR, MAR or MNAR? Do some variables have missing data due to an MCAR mechanism while others have missing data due to an MNAR mechanism? What, if any, auxiliary variables can you include to best account for these mechanisms? The imputation model you construct will vary, maybe drastically, depending on your reasoning behind the mechanisms of missing data in your data set.

Once you are ready to impute, you will follow a series of steps summarized in the figure below. First, you will specify an imputation model to generate *m* complete data sets ($m = 3$ to simplify Figure 1, but *m* is often much larger in practice). Specifying an imputation model is by itself a multi-step process, which we describe in more detail in our worked example, below. The *m* complete data sets that are generated from this imputation model contain estimated plausible values for each missing data point based on the observed data. Each imputed data point incorporates the participant's available data, a regression model predicting that data point based on the associations observed for other participants in the data set, and random noise to reflect the uncertainty of these values. Each imputed data set is then analyzed individually according to a common statistical model (e.g., ordinary least squares, logistic or multinomial regression). The results of analyses on each *m* data set will differ, as random components will have led to different values being generated within each data set. Finally, parameter estimates are pooled into a single set of estimates, variances and confidence intervals (Baraldi & Enders, 2010; Enders, 2016; Schafer & Graham, 2002; van Buuren, 2018).

The main steps in multiple imputation:



Note: although single imputation (e.g., mean imputation) methods exist, we do not recommend their use under most circumstances due to resulting bias and reduced generalizability of results.

(e.g., is not present during a specific assessment session) or withdraws from the study completely (i.e., attrition). These scenarios almost always occur in developmental research, especially in longitudinal studies.

Missing data can negatively affect our ability to draw valid conclusions because it both reduces statistical power and introduces bias to parameter estimates. Yet rather than adjust for this bias, many developmental researchers opt to simply remove participants with missing information from the data set (i.e., listwise deletion, also referred to as complete case analyses) or to only use their information when they offer complete data for a given analysis

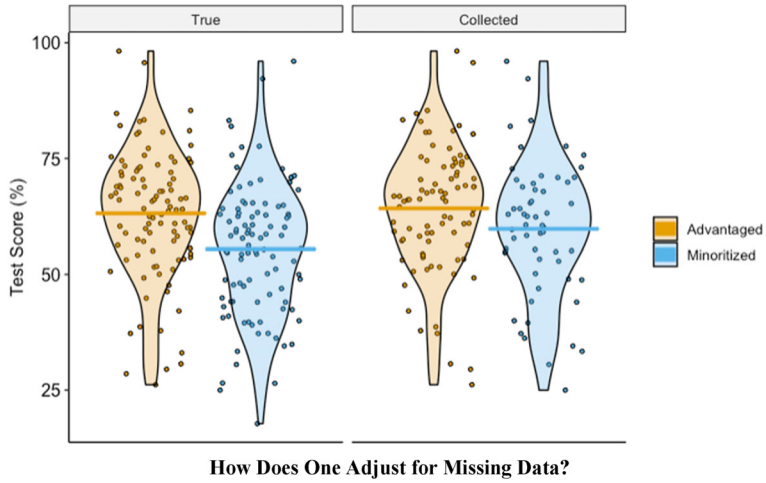


FIGURE 1 Simulated student test data showing how missingness skews results based on performance and failure rates from university physics courses in Nissen et al. (2018) and Van Dusen and Nissen (2020). The true data represent scores for every student in a course. The collected data is missing data from students who failed the course. Minoritized students fail courses at higher rates. The figure shows how this inequality leads to the collected data underreporting the actual inequalities in test scores: $d = 0.75$ versus $d = 0.43$.

(i.e., pairwise deletion). These 'easy' deletion methods are often the default setting in common software programs. However, these options often increase bias and create inefficient estimation of parameters, confidence intervals, and significance tests (Baraldi & Enders, 2010). When conclusions are drawn from biased statistics, the work that comes after is likely to be biased or fail to replicate previous findings (Lee et al., 2021), and the line of research can ultimately lead to intervention or policy recommendations that are grounded in biased results. Underpowered studies are also often cited as a significant contributing factor to the 'Replication Crisis' in psychology (Button et al., 2013; Nosek et al., 2015). These negative consequences ultimately reduce the validity and reliability of inferences to the population.

Below, we discuss why and how barriers have slowed the widespread adoption of missing data practices. We also outline how failing to adjust for missing data has ethical implications that are especially relevant for developmental researchers invested in open science, diversity, equity, inclusion and/or accessibility initiatives.

3 | WHY IS APPROPRIATELY ADDRESSING MISSING DATA IMPORTANT?

3.1 | Barriers to widespread adoption of missing data analyses and adjustments

Systemic and individual barriers slow the adoption of evidence-based practices, whether in psychology (Nosek et al., 2015), economics (Delios et al., 2022; Tierney et al., 2020, 2021), medicine (Grol & Wensing, 2004) or other disciplines (see Proctor et al., 2009 for an overview). These barriers are no different for quantitative methods (King et al., 2019). Many 'best practices' in statistics are slowly (if ever) adopted. This lag may be accounted for by both individual factors (i.e., lack of access to statistical training or technology) and systematic barriers (i.e., field-wide norms about what data analysis methods are considered acceptable). The practice of transparently and appropriately addressing missing data has achieved widespread methodological support (Appelbaum et al., 2018; Manly & Wells, 2015; Nicholson et al., 2017; Sterne et al., 2009; Sterner, 2011; Vandenbroucke et al., 2007). Yet, repeatedly,

reviews have found that progress has been slow in increasing its implementation (Bodner, 2006; Burton & Altman, 2004; Karahalios et al., 2012; Lang & Little, 2018).

Beyond the systemic barriers that are common across research fields, developmental scientists have other reasons for being slow to adopt modern missing data practices like multiple imputation. It is difficult to adopt any practice when guidelines and practical demonstrations of that practice have not been tailored to the research of developmental scientists. This gap around addressing missing data likely disproportionately affects early career researchers, especially those from backgrounds that are traditionally underrepresented in science. A well-established or more privileged researcher is more likely to (1) have the resources needed to enroll in a course on handling missing data; (2) hire a statistician to do the work; and/or (3) seek formal or informal mentorship about designing studies to minimize missingness including planned-missing designs, as well as about appropriately addressing missing data once it occurs. Early career researchers and/or those from underrepresented backgrounds are less likely to have access to these resources and would probably benefit the most from a well-tailored guide to handling missing data.

Developmental psychologists may also be hindered in adopting multiple imputation by several individual-level barriers. These may include (1) a lack of familiarity with or confidence using statistical software; (2) pressure from colleagues and advisors to submit and publish manuscripts as quickly as possible by using software defaults that match established norms for handling missing data (i.e., listwise deletion methods); (3) worries over whether the decision-making process required by multiple imputation is 'correct'; and (4) added complexity in the data analysis process. On top of these barriers, a number of common misconceptions about multiple imputation further limit its adoption (e.g., multiple imputation is 'making up' data, should not be used for dependent variables, and/or is only appropriate when data are missing at random, which is defined in further detail below; see Table 1 for a full list of common misconceptions). Many of these misconceptions are based on the general idea that using multiple imputation to manage missing data is ethically questionable. We argue the opposite—the ethical risks from failing to properly adjust for missing data far outweigh those raised by multiple imputation.

3.2 | Ethical implications of failure to adjust for missing data

Properly adjusting for missing data is vital for investigations of diversity, equity, inclusion and accessibility. These investigations aim to counteract the historical and continued oppression of minoritized groups in scientific research (Zuberi, 2001; Zuberi & Bonilla-Silva, 2008) and are crucial to creating a more open science. There are similar implications for clinical trials, interventions and meta-analyses (see review by Rioux & Little, 2021). For example, missing participants might experience more favourable outcomes in the treatment group and poorer outcomes in the control group (or vice versa), which would bias conclusions toward (or away) from the true efficacy of the intervention. It is important to consider and adjust for missing data because this can invalidate the conclusions we draw and, in turn, waste resources and lead to poor policies (Mavridis et al., 2014; Rioux & Little, 2021).

Adjusting for missing data through appropriate and replicable methods is also an important step in promoting open science initiatives. Many developmental scientists advocating for open scholarship work to improve openness, integrity, social justice, diversity, equity, inclusivity and accessibility in all areas of their scholarly activities. By extension, they hope to improve both their academic field and the societies they live in (Ledgerwood et al., 2022; Pownall et al., 2021). Streamlining procedures to address missing data and increasing the transparency of those procedures through consensus on reporting standards will advance these goals (Randall et al., 2021). Several outlets, including *Infant and Child Development*, have called for researchers to prioritize similar 'rigorous, transparent, credible and robust' methods in the work they submit for publication (Syed, 2021).

To maximize the contribution of our participants' data, we must plan for handling missing data during the early phases of research design — for example, by designing data collection procedures to minimize missing data. Practically, this means researchers need to collect information on additional (auxiliary) variables that may be related to missing data. This is because structural barriers to participation in research can lead to participants from minoritized

groups disproportionately dropping out of longitudinal studies or not completing measures (Randall et al., 2021). Thus, data from minoritized students may be most likely to be excluded from longitudinal studies investigating academic achievement using pairwise or listwise deletion methods. This selection effect can bias model estimates and confidence intervals, obscuring the inequities in student outcomes and possibly leading to unsubstantiated claims about achieving equity (Rhodes, 2015). Collecting demographic data that is often associated with attrition (e.g., income, education level and occupation) during recruitment or early in the study can help researchers better understand missingness in their data set, even if a participant is lost to follow-up or fails to complete the full trial.

We must also identify ways to address missingness when it occurs. Our participants donate valuable time to us when they participate in our studies. When participants provide valid, albeit partial, data, we should maximize their contributions whenever possible by leveraging the incomplete data. When scientists drop records because of partial missing data by using deletion methods, they nullify the donation of time from their participants.

These ethical considerations are not an exhaustive list. Additional considerations may need to be weighed when choosing strategies to adjust for missing data (e.g., cultural considerations, protection of data or participants, etc.). In a subsequent section, we address specific ethical considerations in the process of conducting multiple imputation analysis.

3.3 | Deletion methods increase bias and decrease representation

Deletion methods for handling missing data are the default option in most software analysis platforms. We argue that there are increasingly limited situations in which deletion methods may be used. Deletion methods exacerbate bias in parameter estimates when some participants are more likely to have missing data than others (e.g., Curran, Bacchi, et al., 1998; Curran, Molenberghs, et al., 1998; Fairclough et al., 1998; Widaman, 2006). The two most common deletion methods are *pairwise deletion* and *listwise deletion*. Pairwise deletion is a common practice that excludes missing data on an analysis-by-analysis basis; only complete cases for relevant variables are included (Myers, 2011). Entirely excluding participants who have any missing data on at least one of the variables included in the analysis is known as listwise deletion (Myers, 2011). This approach further exacerbates bias as it ignores *all* information from participants who have *any* missing data (Altmann & Bland, 2007; Howell, 2007; Kang, 2013). Deletion methods are simple to implement and time efficient, particularly when the loss of statistical power is inconsequential (Kang, 2013; Schafer, 1999). But these deletion methods may be misaligned with the researcher's intentions to make their work as inclusive as possible.

Deletion methods are appropriate only in certain limited circumstances because they generally assume that the data are Missing Completely at Random (MCAR; discussed in more detail, below). With MCAR data, and *only* with MCAR data, deletion methods will not bias inferences. This is because the complete records in an MCAR data set are a random sample drawn from the larger sample of participants. This larger sample includes records with missing data and is, in turn, drawn from the population (Kang, 2013). When researchers conduct analyses using this 'random sample' of complete records, the analyses will not lead to biased parameter estimates, although tests of statistical significance will have decreased power due to the loss of observations. In practice, MCAR data are very rare. This is why we *do not recommend deletion methods*[1]—because of the resulting loss of statistical power, constraints on the generalizability of the results, and the likelihood that the MCAR assumption is not met. The conclusions researchers draw when they use deletion methods are generalizable *only* to a population similar to participants with complete data (e.g., those participants who fully complete surveys). The use of deletion methods with data Missing at Random or Missing Not at Random (MAR and MNAR; explained in detail, below) will *always* introduce selection bias into inferences. This bias undermines the validity of researchers' conclusions by greatly decreasing the probability that researchers will statistically detect true inequalities across groups (Hernán et al., 2004).

Here, we offer an example of how listwise deletion may bias estimates, impede replicability, and disproportionately impact minoritized individuals from Nissen et al. (2018) and Van Dusen and Nissen (2020). Developmental

TABLE 1 Debunking misconceptions about multiple imputation.

Misconceptions	Reality
Multiple imputation should only be used when the missingness is MAR	MAR is the least restrictive assumption for multiple imputation. Therefore, multiple imputation is also appropriate (and better than listwise deletion due to increased statistical power) under the more restrictive MCAR assumption. Even under MNAR, multiple imputation (used with sufficient auxiliary variables) can offer advantages over other approaches (e.g., deletion-based methods).
Multiple imputation should only be used when too few cases are left after listwise deletion	Multiple imputation has advantages even when the amount of missing data is low (i.e., because multiple imputation will eliminate bias under MAR and can partially eliminate bias under MNAR).
If results from statistical analyses obtained from multiple imputation differ from those of listwise deletion, the results of multiple imputations must be wrong	Results of multiple imputation have been shown to be more accurate and reduce bias in parameter estimates compared to deletion techniques when the multiple imputation model is correctly specified.
Certain variables must not be imputed (outcomes/predictors)	With the exception of special instances, most variables can be multiply imputed with benefits. Caution in using multiple imputations is, however, warranted for missing social identity data for ethical concerns (Randall et al., 2021).
Multiple imputation must not be used because it can produce several different outcomes in statistical analyses	Following the computation of multiply imputed data, point estimates from the analysis of each data set are pooled to provide one overall estimate. Generally, this is done using Rubin's (1987) rules. However, sometimes a pooling method is not available for certain commands in your software package of choice. In these instances, we recommend switching to another package. If this is not possible, transparently reporting an ad hoc solution is key.
Multiple imputation is making data up	Algorithms for imputing missing data use the available data to optimize the accuracy of missing values that are replaced. Sufficient multiple imputations allow researchers to estimate the most likely values for the variable and case while incorporating uncertainty.
Doing anything other than listwise or pairwise deletion is hard enough that it is not worth doing	With some training, researchers can develop skills to implement best practices for handling missingness such as multiple imputations, which can be completed in a reasonable amount of time and will ultimately provide knowledge producers and consumers with a more accurate understanding of the relations that are being examined. Researchers may also utilize the skills of a methodological consultant to help incorporate best practices for missing data analysis in their design and analysis.
The computational demands of multiple imputation are too intensive and/or will take too long to complete	Thanks to advances in computing power, only very complex analyses or 'big data' such as neuroimaging and genomics data sets are likely to have computational constraints. For most studies, multiple imputation can be performed in a reasonable amount of time with modern hardware. Multi-core processors are common, and modern software can create multiple imputed data sets concurrently. Moreover, refusing to adjust for missing data given time constraints is not a valid reason to avoid multiple imputation. Good science is not always fast science.

Note: Adapted from van Ginkel et al. (2020).

psychologists often administer assessments before and after an intervention, for example, to measure growth in students' knowledge (Singer & Smith, 2013). While most students participate in the pretest, research suggests that students who earn lower grades in a course are less likely to participate in the posttest (Kost et al., 2009; Kost-Smith

et al., 2010; Nissen et al., 2018; Nissen & Shemwell, 2016). This means students with lower grades are *more* likely to have missing data. If the researcher uses listwise deletion, students in their sample with lower grades are most likely to be removed from analyses. Because minoritized students experience structural barriers to success that increase their likelihood of having higher rates of failing grades (Benford & Gess-Newsome, 2006; Van Dusen & Nissen, 2020), using listwise deletion may shrink sample sizes for minoritized students. This will artificially inflate group mean grades, making inequalities in outcomes between majority and minoritized groups appear smaller and ultimately biasing estimates and interpretations based on post-intervention assessment scores (Dynan & Rouse, 1997; Hutchins et al., 1999; Kanim & Cid, 2020; National Academy of Sciences, 2011)[2].

Figure 1 illustrates this by showing how analyzing complete data can skew findings about inequities across student groups using simulated data based on performance and failure rates from university physics courses. In these data, the true mean score for non-Hispanic White students (65%) is similar to the collected data (68%), while it is meaningfully lower for minoritized students (53%) than the collected data (63%). This bias in data collection reduced the effect sizes between groups from $d = 0.75$ to $d = 0.43$ and misrepresented the impacts of systemic barriers to minoritized student success. In contrast, using multiple imputation will retain students across the grade distribution, and more accurately estimate the true group means for students from all groups.

4 | HOW DOES ONE ADJUST FOR MISSING DATA?

4.1 | Missing data mechanisms

Researchers first should try to understand *why* data may be missing before making any adjustments or conducting analyses (see Box 1). Data can be missing for many different reasons, including item non-response, attrition during longitudinal studies (Jeličić et al., 2009), participants' inability to complete tasks, or not passing quality controls. When researchers discuss missing data, they usually make a distinction between three main reasons why data may be missing, referred to as *missing data mechanisms*. These mechanisms are *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR; Heitjan & Basu, 1996; Little & Rubin, 2002). MCAR, MAR and MNAR each lead to distinct assumptions about the generalizability and validity of the inferences drawn from a data set. Although these distinctions are useful for researchers thinking about why data might be missing from a given data set, these are *theoretical* distinctions. In practice, with a few important exceptions (e.g., planned missing designs), knowing or uncovering the true mechanism causing missing data is not possible. Absolutely distinguishing between these mechanisms would require observing values that are unobserved in the data set.

4.1.1 | Missing completely at random

MCAR refers to situations when missing data are the result of a truly random process. Formally, MCAR means that the likelihood of any given data point being missing for a participant is unrelated to the rest of the participant's data. The most unambiguous cases of MCAR come from missingness generated at random *by design*. Researchers can implement planned missing designs when collecting data (Graham et al., 2006; Rhemtulla & Hancock, 2016; Rhemtulla & Little, 2012; Wu & Jia, 2021). For example, each participant may only be given a random subset of the assessments to complete (e.g., the design of the National Assessment of Educational Progress). Another example is when a random subset of participants are given resource-intensive measures (i.e., direct observations of classroom behaviour) in addition to similar but less intensive measures that may be more biased (i.e., teacher-reported classroom behaviour). In a third variation of planned missing designs, different participants are given random subsets of scale items to collect data on more variables overall while minimizing participant burden.

MCAR is considered safe to 'ignore' because most missing data approaches (including listwise deletion) provide unbiased parameter estimates under MCAR. However, the precision of parameter estimates is still reduced (Pedersen et al., 2017) (see Supporting Information: Table A1 for a list of missing data approaches). In theory, researchers can test if data are *not* MCAR by examining distributional differences between cases with fully observed data and cases with missing data (Raykov, 2011). However, the absence of evidence that data are MAR or MNAR does not constitute evidence that data are MCAR. In practice, determining that data are MCAR is impossible unless the researcher used a planned missing design and there are no other sources of missing data.

4.1.2 | Missing at random

MAR refers to situations when missing data are generated in a systematic manner that can be fully accounted for using information contained within a data set. Following our previous example on listwise deletion, students who do not complete post-tests are more likely to have lower scores on pre-tests of educational knowledge (Kost et al., 2009; Kost-Smith et al., 2010; Nissen et al., 2018; Nissen & Shemwell, 2016). Since lower pre-test scores predict students' missingness on post-tests, the missing post-test data are MAR. Researchers can use modern missing data methods (e.g., multiple imputation, full information maximum likelihood estimation or FIML) to incorporate variables that account for MAR missingness in their data. These methods allow researchers to estimate parameters with less bias.

Variables that explain the mechanism behind missing data are called *auxiliary variables* if they are included in the missing data model, but not included in the analytic model (Collins et al., 2001). In the example about test scores, above, pre-test scores should be included in the imputation model to help adjust for the missing post-test scores. However, if pre-test scores were included in the imputation model but *not* included in the final analytic model, they would be considered an auxiliary variable. As another example, suppose a researcher did not want to control for socioeconomic status (SES) in their analytic model but SES predicted patterns of missingness in other variables. If the researcher included SES in the missing data model, SES would be an auxiliary variable. Notably, non-auxiliary variables used in the analysis could also account for MAR. For example, achievement may be both a predictor in analysis and could predict patterns of missing data. Including this variable in the imputation model could also help account for MAR.

Researchers would ideally design their studies to collect auxiliary variables to help account for missingness and aid in building missing data models. Including many auxiliary variables in a model can increase the plausibility that missing data are MAR (Collins et al., 2001). However, including lots of auxiliary variables may not be feasible in large secondary data sets, in part because increasing the number of variables in a model can lead to computational problems like non-convergence due to multicollinearity (van Buuren & Groothuis-Oudshoorn, 2011). When building the missing data model, van Buuren and Groothuis-Oudshoorn (2011) and van Buuren and Oudshoorn (2000) recommend including all variables the researcher plans to use in the analytic model as well as all auxiliary variables for which the distributions between the response and nonresponse groups differ by a certain reasonable magnitude (e.g., based on an expected minimum correlation with the target variables or that explain a predetermined amount of variance). Some software programs include functions that help select these variables automatically, such as the *quickpred* function in the *mice* package in *R* (further discussed in the worked example section, below).

4.1.3 | Missing not at random

MNAR refers to missingness that cannot be fully accounted for with other variables in the data set. With MNAR, we can only guess what the missing data mechanism may be. This is because MNAR can occur if missingness depends on either the unobserved data or on the missing values themselves (Fielding et al., 2008). This is true regardless of

whether missingness also depends on observed data. For example, parents of children experiencing more behavioural concerns might be less likely to return a questionnaire on the impact of certain parenting practices on their child's behaviour. In this example, the missing value for behavioural problems depends on the missing parent questionnaires. Ignoring this missingness would lead to significant biases in estimating the relation between parenting practices and behaviour due to selection effects.

Unfortunately, there is no simple or straightforward way to combat MNAR missingness. Methods for handling MNAR data attempt to model the reason, or mechanism, for missingness. These methods include selection and pattern mixture models (Heckman, 1979; Little, 1993). Modern efforts have focused on applying selection or pattern mixture models towards longitudinal data (e.g., Enders, 2010, 2011). The quality of the correction depends on the quality of the model for the missing data mechanism. In principle, if the mechanism is modelled correctly, bias due to MNAR would be negated; however, in practice and by definition, knowing the exact nature of the missingness mechanism is impossible. Consequently, some researchers avoid using most missing data tools like multiple imputation under MNAR due to concerns about inadequately addressing bias in their models.

On the other hand, the use of modern missing data adjustments is likely a better solution than simply ignoring missingness (i.e., defaulting to methods like listwise deletion), even under MNAR. For instance, van Ginkel et al. (2020) argued that using a sufficient number of auxiliary variables for multiple imputation can still produce less biased estimates than listwise deletion under MNAR. In addition, using multiple imputation with auxiliary variables can restore statistical power lost due to missingness (Collins et al., 2001; Graham, 2009). Other recommendations suggest conducting sensitivity analyses. For example, a researcher would fit multiple types of missing data models (e.g., selection or pattern mixture models) to the same data set to check the impact of different MNAR assumptions on parameter estimates (Demirtas & Schafer, 2003). Overall, we believe the use of multiple imputation under MNAR is justified and provides important advantages over more common deletion techniques. That said, because bias cannot be fully eliminated, keeping this limitation in mind when reporting findings is important. Ideally, researchers will prevent MNAR from the outset by designing robust studies.

4.2 | Multiple imputation considerations

4.2.1 | Time

The amount of time that multiple imputation takes will vary according to the complexity of the multiple imputation model and size of your data set as well as by the software used. For example, our worked example below uses a highly complex data set with thousands of participants. The complexity of the multiple imputation models we created to handle missingness is reflected in the many potential auxiliary variables we could have used and the nested structure of the models (measurements within students, within schools). van Buuren (2018) recommended including no more than 15–25 auxiliary variables (van Buuren, 2018), but there are minimal downsides to including a large number of auxiliary variables (Enders, 2010). However, including more auxiliary variables does increase model complexity. This increased complexity could lead to substantial increases in computation time and risk of non-convergence. Researchers need to balance the information added from additional auxiliary variables with computation time and potential non-convergence when choosing the final set of variables. Multiple imputation models for less complex data sets may not contain nearly as many auxiliary variables; yet, as discussed earlier, auxiliary variables that could lead to a reasonable assumption of MAR should be considered during the design of the study. Categorical variables are also associated with added computational time. Including information associated with attrition or non-response, though requiring more time, can reduce the likelihood of encountering MNAR missingness, especially in prospective longitudinal designs. Some software, such as *mice* in R (van Buuren & Groothuis-Oudshoorn, 2011), include parallel processing functions that greatly reduce the amount of time needed for the imputations to run.

BOX 2 Basic Reporting Standards: A Checklist for Reviewers and Editors

Information that must be reported in academic articles:

☐ Did the authors include a 'Missing Data' section? Failure to discuss missing data could be grounds for rejection or major revision. Within this section, the authors should also include:

- ☐ The proportion of missing data by variable and by case, including the sample size available under listwise deletion
- ☐ A brief justification for why and how the authors are addressing missing data (e.g., it is plausible that the data are MAR?)
- ☐ A brief justification as to whether auxiliary variables were included and how data on key variables are MAR with inclusion of these variables
- ☐ The number of auxiliary variables and what these variables represent

If the authors specifically used multiple imputation to adjust for missing data, this section should also include:

- ☐ The algorithm used to impute missing data (e.g., MICE)
- ☐ The number of data sets imputed and a justification for this decision
- ☐ The number of iterations (if using chained equations or MBI) and the rationale for this decision
- ☐ Whether and why it is believed that model convergence was achieved
- ☐ Whether any alterations were needed to achieve model convergence
- ☐ Results from model checks and sensitivity analysis (can go in the supplement):
  - ☐ Tests for inclusion of auxiliary variables, if applicable
  - ☐ Convergence plots
  - ☐ Descriptive statistics before and after multiple imputation (preferred in the main document)
  - ☐ Results obtained under listwise deletion relative to multiple imputation

Example methods paragraphs can be examined in Enders (2010) and Manly and Wells (2015).

Optional but recommended: Preregistration of missing data decisions

Preregistration is where a researcher publishes their planned study procedure as an immutable document in a time-stamped database (e.g., Open Science Framework, As Predicted; Baum et al., 2022; Parsons et al., 2022). Typically, preregistration includes specifying research questions/hypotheses, the research design, and data analysis plan before conducting analyses (e.g. Mertens & Krypotos, 2019; Nosek et al., 2018; Pownall et al., 2021, 2022; Tierney et al., 2020, 2021; Topor et al., 2020). This process helps to avoid too many 'researcher degrees of freedom' leading to potentially spurious findings (Azevedo et al., 2019, 2022; Wicherts et al., 2016). Therefore, preregistering missing data decisions is ideal. We provide more information, including links to templates, in Woods et al., (2021) at https://doi.org/10.31234/osf.io/mdw5r. A registered/exploratory report is more ideal to implement than pre-registration, as the rationale, methods and analysis can be reported *a priori* and reviewed by peer reviewers. Once in-principle acceptance is received, authors can begin data collection and analysis; however, they cannot change the rationale or methodology. Thus, in a registered report, the focus is less on the findings and more on the research question, methodology and analysis (see review by Chambers & Tzavella, 2022). This process helps reduce publication bias, allowing the literature to be less distorted (Findley et al., 2016). If time is a factor that affects project completion, pre-registration is an adequate approach to reduce researcher degrees of freedom.

4.2.2 | Revisiting earlier decisions

In practice, researchers making decisions about missing data analysis and multiple imputation may need to revisit choices made earlier in the process. It is very common to uncover information down-stream that leads to rethinking an up-stream choice. The goal is not to approach multiple imputation perfectly, or even linearly, but instead to think carefully about decisions and report these decisions with transparency. Although circling back to an earlier step or redefining assumptions during the process is normal, these revisions and the rationale behind them should be clearly documented (see Box 2).

4.2.3 | Ethical considerations when imputing[3] social identifiers

While multiple imputation will nearly always provide less biased findings than listwise deletion, imputing some variables raises ethical considerations (Brown et al., 2021). There is a long history of minoritized individuals having their social identifiers assigned in ways that do not align with their identities (Ford, 2001; Puthillam et al., 2022; Shih & Sanchez, 2009). This context has made some researchers wary of any practices that ascribe social identifiers to research participants beyond what they have self-selected (Brown et al., 2021). However, some diversity, equity, inclusion, and accessibility research would not be possible without multiple imputation (Rhodes, 2015). While this practice of assigning social identifiers can be problematic in many settings, the nature of multiple imputation limits its potential harms in two important ways. First, multiple imputation does not simply 'assign' a social identifier to an individual with missing data. It creates a probability distribution of multiple social identifiers based on the rest of the information known about an individual. Second, multiply imputed data do not create findings about any specific individual who has had their data imputed; instead, conclusions are drawn about an aggregated population from the models containing individual data. Researchers have found that imputing social identifiers can limit bias from missing data and meaningfully improve the accuracy of model predictions (Rhodes, 2015).

While multiple imputation of social identifiers can be an important step to preparing data, it is worth considering the specifics of a data set before imputation. For example, if a question about gender identity is asked as a binary (only offering man or woman), a blank answer might be a participant's way of communicating that they do not identify as either gender. In this case, imputing gender as man or woman would be misinterpreting the participant's response. Brown et al. (2021) have provided a set of recommendations on when and how researchers should impute social identifier data for investigations of racial equity. In sum, there is no singular answer to whether missing social identifier data should be imputed. Researchers should weigh the ethical considerations for and against imputing social identifier data within the context of their research and the communities that their research impacts.

4.2.4 | Data structure

Multiple imputation models should be based on and match their corresponding analytic model to satisfy the congeniality assumption (Meng, 1994; see also a discussion in van Buuren, 2018). The data set we use in our worked example, below, contains nested data (occasions within students, students within schools). Thus, a question arises about how to best accommodate clustered data in multiple imputation models.

Researchers planning to analyze both single- and multilevel models could develop separate imputation models for each planned analytic model, or impute the most complex model and use these imputed data sets to analyze similar but less complex models (Graham, 2012). For example, when the researcher plans to analyze both single- and multilevel models, they could impute into a multilevel model and use these imputed data for their single-level analyses. However, researchers generally should not use single-level multiple imputation when they intend to analyze

multilevel models. Lüdtke et al. (2017) showed that when a single-level multiple imputation model was used and data were then analyzed via a multilevel model, both the resulting within- and between-group coefficients and their standard errors were biased, especially when the intraclass correlation coefficients (ICCs) were larger. The only scenario in which single-level multiple imputation produced results similar to multilevel multiple imputation occurred when the missing data rate was low and ICCs were small (Lüdtke et al., 2017). In contrast, when a multilevel multiple imputation model was used, no substantial bias in between- or within-group coefficients or standard errors was observed.

Multilevel multiple imputation is not yet available in some statistical software programs that are commonly used in developmental psychology. Software capabilities are being rapidly developed or expanded, and it is reasonable to expect that these best options may soon be available across all platforms. In the meantime, researchers may not be willing or able to learn a new software language to conduct multiple imputation, and hence, may need to divert to less optimal solutions that would still be an improvement over listwise deletion.

There are several alternative approaches one could use to ensure their multiple imputation model is as congenial as possible to their planned multilevel analytic models. Two of these approaches are to include cluster variables as dummy indicators in the multiple imputation model, or to multiply impute data separately within each individual cluster (Graham, 2012). However, these options do not work equally well, nor are they as efficient as multilevel multiple imputation. In both approaches, imputing higher-level variables (e.g., school characteristics) is not straightforward. The dummy-indicator approach leads to overestimated ICCs and underestimated between-group coefficients and standard errors, especially for smaller cluster sizes, although within-group coefficients and standard errors may not be substantially biased (Lüdtke et al., 2017). The Impute-Within-Clusters strategy preserves means, variances, and covariances within each cluster, but nonetheless still leads to the problem of overestimated ICCs. This approach also needs large cluster sizes, which are not always available (Graham, 2012). For example, our worked example data set contains a large number of schools ($N = 893$) and fairly few students per school (average n per school $= 8.4$), which produced convergence problems with the Impute-Within-Clusters approach.

Another alternative ad hoc solution is to conduct both a single-level multiple imputation model at level 1 that includes dummy indicators for the clusters as predictors (e.g., school ID variables) as well as a separate level 2 multiple imputation model using aggregated level 1 data. These two data sets can then be merged into one multilevel data set for analysis (Grund et al., 2018; van Buuren, 2011). We took this approach to demonstrating multiple imputation in our worked example for one software program that does not yet allow multiple imputation (Stata). We caution that there is little evidence evaluating the effectiveness of this approach. However, our results are similar to those obtained from other software that implemented a multilevel imputation model (see Tables 3–6). We also believe that adjusting for missing data in this ad hoc fashion represents an improvement over both listwise deletion methods and single-level imputation models, because it is more congenial with our planned multilevel analyses. But simulation studies are needed to further evaluate this approach.

4.2.5 | Handling derived variables

Developmental researchers may often include derived variables in their analytic models. Examples of derived variables are multi-item scales, in which individual items are averaged or added together, or interaction effects, which may include cross-level interactions. Multiple ways to handle derived variables have been proposed. For derived categorical variables (e.g., interaction effects with categorical predictors), imputation may be conducted separately for each category (van Buuren, 2018). For multi-item scales, scale-level imputation could be conducted, in which only the derived variable (the composite) is used in the imputation and the individual items are not included. However, this method disregards information from participants who answered some but not all items, leading to loss of power (Gottschall et al., 2012).

An alternative option is to *impute, then transform* (von Hippel, 2009). In this approach, variables used to create the derived variables are imputed individually, and the derived variables are computed after the data are imputed.

TABLE 2 Main variables and complete list of all potential auxiliary variables evaluated for use in our multiple imputation worked example.

Variables	Fall K	Spring K	1st Grade	2nd Grade	3rd Grade	4th Grade	5th Grade
Main							
WM		X					
Math		X[a]					
Income		X					
Age		X					
Disability		X[a]					
Parent Educ.		X[b]					
Par. Employ	X						
Cog. Stim.	X[c]						
Male		X					
Race		X					
% Lunch		X					
Auxiliary							
Age	X		X	X	X	X	X
WM direct	X		X[d]	X[d]	X[d]	X[d]	X[d]
WM parent					X	X	
Math	X		X	X	X	X	X
Income			X	X	X	X	X
Disability							
DCCS	X						
Tch behavior	X[e]	X	X	X	X	X	X
Par Behavior:							
Approaches		X[b]					
Control		X[b]					
Impulsive		X[b]					
Bilingual		X					
Single Par.		X	X	X	X	X	X
Burnout		X					
Public Sch.		X	X[e]	X[e]	X[e]	X[e]	X[e]
% Non-White		X	X[e]	X[e]	X[e]	X[e]	X[e]
Title I funds		X	X	X	X	X	X
% Lunch			X	X	X	X	X
Disadvantage		X	X[e]	X[e]	X[e]	X[e]	X[e]

[a]Disability*Math interaction term could be computed from these two items before or after imputation; see Section 4.2.5.
[b]One variable constructed with information averaged across the fall and spring of kindergarten.
[c]Scale created from nine items. Either the final scale could be created and imputed, or the nine items could be used separately in the imputation model; see Section 4.2.5.
[d]Variables were auxiliary for RQ1–3 imputation models and main for RQ4 imputation models.
[e]Variables were discovered to be collinear with the spring K value and were dropped from all imputation models.
Abbreviations: Cog. Stim., cognitive stimulation; DCCS, Dimensional Change Card Sort, an executive function task; Par, parent; Sch, school; Tch, teacher; WM, working memory.

TABLE 3 Sample descriptives before and after multiple imputations ($N = 7509$)

			Under listwise deletion		Under multiple imputation			
	% Miss	Range	Mean	SD	Stata	R	Blimp RQ1–3	Blimp RQ4
Key variables								
WM K	4.29	393–563	451.82	30.15	451.62	451.62	451.62	451.92
WM 1st	4.91	393–596	470.35	25.34	470.21	470.21	–	470.01
WM 2nd	5.95	403–581	481.90	22.41	481.78	481.76	–	481.47
WM 3rd	6.83	403–603	490.79	21.53	490.63	490.65	–	490.4
WM 4th	7.96	403–588	498.22	20.71	498.01	498.00	–	497.96
WM 5th	8.62	403–588	504.32	21.47	504.08	504.08	–	504.44
Math K	4.33	11.78–112.54	51.39	13.38	51.29	51.29	51.27	51.29
Male	0.23	0–1	0.50	0.50	0.50	0.50	0.50	0.50
Race	0.13	1–5						
White %			0.50		0.50	0.50	0.50	0.50
Black %			0.09		0.09	0.09	0.09	0.09
Hisp. %			0.26		0.26	0.26	0.26	0.26
Asian %			0.11		0.11	0.11	0.11	0.11
Others %			0.05		0.05	0.05	0.05	0.05
Age at K	3.98	52.31–97.41	73.54	4.39	73.53	73.52	73.52	73.53
Education	10.73	1–3						
HS or less %			0.35		0.36	0.34	0.36	0.36
Some College %			0.30		0.30	0.31	0.30	0.30
Bachelor's %			0.35		0.34	0.34	0.34	0.34
Employed	28.03	1–3						
Full Time %			0.44		0.43	0.42	0.43	0.43
Part Time %			0.22		0.22	0.27	0.22	0.22
Unemploy %			0.34		0.35	0.31	0.35	0.35
Income K	21.75	1–18	11.10	5.45	10.75	10.89	10.80	10.79
Disability K	24.28	0–1	0.18	0.39	0.18	0.26	0.18	0.18
Cog. Stim.	26.85	1.11–4	2.90	0.47	2.89	2.90	2.89	2.89
Lunch K	20.36	1–4	2.45	1.16	2.54	2.54	2.55	2.54

Note: 'Lunch K' in Stata was obtained from a level 2 imputation model.
Abbreviations: K, kindergarten; WM, working memory.

This option is problematic because it may bias parameter estimates involving the derived variable toward zero (van Buuren, 2018). Another option is to treat derived variables as *'just another variable'* (*JAV*; White et al., 2011), also referred to as *transform, then impute* (von Hippel, 2009). With this option, both the variables used to create the derived variables and the derived variables themselves are imputed. The problem here is that the imputed derived score might differ from the score computed from the imputed variables (van Buuren, 2018). As an alternative, *passive imputation* (van Buuren & Oudshoorn, 2000) occurs when computation of the derived variable is conducted as part of multiple imputation 'on-the-fly' (see section 6.4 in van Buuren, 2018). This method aims to address the problems with *impute, then transform* and *JAV*. Other, newer options accommodate substantive models, such as the *smcfcs* method (Bartlett et al., 2015) and fully Bayesian model-based imputation (Enders et al., 2020). Although a variety of

TABLE 4 Results for RQ1–2 under multiple imputation in R, Stata and Blimp relative to listwise deletion.

	RQ1				RQ2			
		Multiple imputation				Multiple imputation		
	Listwise Deletion	R	Stata	Blimp	Listwise Deletion	R	Stata	Blimp
Constant	456.35***	454.52***	454.72***	455.11***	456.33***	454.49***	454.71***	455.10***
Level 1								
Math	1.23***	1.25***	1.24***	1.25***	1.20***	1.21***	1.21***	1.22***
Disability	−3.76***	−2.96***	−3.65***	−3.77***	−3.46***	−2.68***	−3.34***	−3.45***
Math*Disab	–	–	–	–	0.14*	0.15***	0.13**	0.14**
Income	**0.13**	**0.18***	**0.19***	**0.17***	**0.13**	**0.18***	**0.19***	**0.17***
Age	−0.16	−0.08	−0.05	−0.06	−0.15	−0.07	−0.05	−0.05
Some college	3.28**	2.41**	2.30**	2.06*	3.32***	2.49**	2.34**	2.10*
Bachelor's	3.08**	2.47**	2.66**	2.32*	3.15**	2.62**	2.70**	2.37*
Part time	−1.27	−0.60	−0.86	−1.36	−1.24	−0.58	−0.84	−1.34
Unemployed	**−0.94**	**−0.74**	**−1.62***	**−1.70***	**−0.96**	**−0.77**	**−1.92***	**−1.71***
Cog. Stimul.	**1.57***	**1.26+**	**1.56***	**1.32+**	**1.60***	**1.27+**	**1.56***	**1.32+**
Male	−2.77***	−2.70***	−2.69***	−2.62***	−2.72***	−2.68***	−2.68***	−2.61***
Black	−3.66*	−3.43**	−3.50**	−3.68**	−3.72*	−3.47**	−3.51**	−3.71**
Hispanic	−6.37***	−5.77***	−5.48***	−5.72***	−6.44***	−5.80***	−5.53***	−5.77***
Asian	**−2.62+**	**−1.78+**	**−1.96***	**−1.88+**	−2.51+	−1.77+	−1.89+	−1.81+
Other race	−0.17	0.40	0.57	0.56	−0.13	0.46	0.62	0.63
Residual Var.	520.90***	564.09***	582.69***	561.83***	520.33***	563.30***	562.17***	561.22***

Note: Unstandardized coefficients presented. All data analyzed in Mplus. Boldface is used to indicate there is a significant difference in results between the types of missing data analyses used.

$^{+}p < 0.10$.

$^{*}p < 0.05.^{**}p < 0.01.^{***}p < 0.001$.

TABLE 5 Results for RQ3 under multiple imputation in R, Stata and Blimp relative to listwise deletion.

	Listwise Deletion	Multiple imputation		
		R	Stata	Blimp
Constant	456.32***	454.25***	454.61***	455.00***
Level 1				
Math	1.23***	1.26***	1.25***	1.25***
Disability	−3.81***	−2.83***	−3.62***	−3.77***
Income	0.06	0.12	0.13	0.10
Age	**−0.20***	**−0.09**	**−0.06**	**−0.07**
Some college	3.02**	2.15*	1.99*	1.71*
Bachelor's	2.86*	2.07*	2.21*	1.84*
Part time	**−1.96+**	**−0.57**	**−0.91**	**−1.44**
Unemployed	**−0.86**	**−0.74**	**−1.68***	**−1.76***
Cog. Stimul.	**1.33**	**1.20+**	**1.51***	**1.27+**
Male	−2.91***	2.70***	−2.69***	−2.62***
Black	−4.05*	−2.59*	−2.80*	−2.96*
Hispanic	−5.86***	−4.79***	−4.64***	−4.82***
Asian	**−2.01**	**−1.54**	**−1.78+**	**−1.69+**
Other race	0.30	0.69	0.82	0.86
Level 2				
% Free lunch	**−0.54**	**−1.00****	**−0.91****	**−0.98****
Variance				
School	14.35***	14.11***	14.20***	15.28***
Residual	501.85***	549.18***	547.86***	545.88***

Note: Unstandardized coefficients presented. All data analyzed in Mplus. Boldface is used to indicate there is a significant difference in results between the types of missing data analyses used.
$^{+}p < 0.10$.
$^{*}p < 0.05.^{**}p < 0.01.^{***}p < 0.001$.

methods exist, methodological work to determine which method is best in which situation is ongoing. Researchers should therefore be clear about which method of handling derived variables they used in their imputation procedures.

5 | A WORKED EXAMPLE: DEVELOPMENTAL PREDICTORS OF WORKING MEMORY

The remainder of this manuscript presents a worked example in which we highlight how to adopt multiple imputation techniques using a nested series of research questions that build in complexity and are often encountered in quantitative developmental research. Although we recommend that researchers perform any appropriate adjustments for missing data rather than defaulting to deletion methods, addressing full information maximum likelihood estimation (FIML) in addition to multiple imputation is beyond the scope of this paper. Multiple imputation generally produces similar results as FIML (Lee & Shi, 2021). We have described common differences between FIML and multiple imputation elsewhere (see Woods et al., 2021). Cham et al. (2017), Dong and Peng (2013), and Lee and Shi (2021) provide additional information about and examples of FIML.

TABLE 6 Results under multiple imputation in R, Stata and Blimp relative to listwise deletion for RQ4.

	Listwise Deletion	Multiple imputation		
		R	Stata	Blimp
Constant	462.52***	460.73***	461.13***	461.21***
Intercept				
Level 1				
Math	1.05***	1.09***	1.08***	1.08***
Disability	−3.83***	−3.02***	−3.97***	−4.36***
Income	0.09	0.09	0.09	0.08
Age	−0.21**	−0.13*	−0.11*	−0.11*
Some college	**1.50**	**1.16**+	**0.99**	**0.83**
Bachelor's	0.75	0.64	0.59	0.48
Part time	**−1.61***	**−0.49**	**−0.82**	**−0.91**
Unemployed	**−1.53***	**−1.16**+	**−2.00****	**−1.91****
Cognitive Stim	**0.92**	**0.94**+	**1.41****	**1.20***
Male	−2.95***	−2.66***	−2.63***	−2.59***
Black	−3.54**	−3.49***	−3.59***	−3.89***
Hispanic	−4.21***	−3.74***	−3.62***	−3.70***
Asian	**−1.06**	**−1.04**	**−1.30**+	**−1.15**
Other race	0.86	0.30	0.37	0.34
Level 2				
% Free lunch	−0.76*	−1.04***	−1.03***	−1.02***
Slope				
Level 1				
Time	9.29***	9.58***	9.49***	9.49***
Time*Math	−0.09***	−0.10***	−0.10***	−0.10***
Time*Disability	0.32	0.19	0.29	0.36
Time*Income	−0.01	0.00	−0.00	−0.00
Time*Age	−0.04*	−0.05***	−0.06***	−0.06***
Time*Some college	**−0.40**+	**−0.37***	**−0.34**+	**−0.30**+
Time*Bachelor's	−0.22	−0.04	−0.03	0.03
Time*Part time	0.24	−0.10	0.00	0.00
Time*Unemployed	0.05	0.05	0.18	0.14
Time *Cog. Stimul.	−0.36*	−0.33*	−0.46**	−0.40**
Time*Male	0.48**	0.39***	0.41***	0.40***
Time*Black	0.38	0.24	0.33	0.31
Time*Hispanic	1.27***	1.16***	1.12***	1.13***
Time*Asian	1.17***	1.00***	1.02***	0.99***
Time*Other	0.03	0.08	0.09	0.09
Level 2				
Time*Free lunch	**0.04**	**0.15**+	**0.12**	**0.15**+
Residual variances				
Time (L1)	254.81***	270.85***	270.92***	270.16***

TABLE 6 (Continued)

	Listwise Deletion	Multiple imputation		
		R	Stata	Blimp
Students (L2)	163.16***	179.95***	178.69***	179.10***
Students*Time (L2)	7.69***	7.95***	7.90***	7.87***
Schools (L3)	9.01**	10.16***	10.21***	11.12***
Schools*Time (L3)	**0.41**+	**0.65*****	**0.66*****	**0.73*****
Covariances				
Time with students	−16.99***	−19.65***	−19.49***	−19.53***
Students with schools	**−0.80**	**−1.63*****	**−1.68*****	**−1.77*****

Note: Unstandardized coefficients presented. All continuous variables are grand mean centred. L1 = Level 1. L2 = Level 2. L3 = Level 3. Boldface is used to indicate there is a significant difference in results between the types of missing data analyses used.

$^{+}p < 0.10$.

$^{*}p < 0.05.^{**}p < 0.01.^{***}p < 0.001$.

5.1 | Motivation and data set

We used data from the Early Childhood Longitudinal Study, Kindergarten Cohort of 2010-2011 (ECLS-K: 2011; Tourangeau et al., 2015) to demonstrate how multiple imputation procedures can be implemented, including how performing multiple imputation may differ depending on software, research questions, and planned substantive analyses. The ECLS-K: 2011 is a nationally representative study of 18,174 US students who began kindergarten during the 2010–2011 school year and were followed longitudinally through the spring of expected fifth grade in 2016. Data were collected on a variety of factors thought to influence children's development across elementary school. Specifically, the researchers collected data about home, neighbourhood, cognitive, behavioural, academic, and school factors. The ECLS-K: 2011 data are publicly available and maintained by the National Center for Education Statistics.

Our cleaned data sets and all code for Stata, R and Blimp are available at https://osf.io/j3f8m. Our worked example is based on Ahmed et al. (2022)[4]. This worked example follows the steps outlined in our multiple imputation decision tree (Woods et al., 2021), available at https://doi.org/10.31234/osf.io/mdw5r. Readers may find the decision tree useful as a step-by-step procedure for handling missing data. This procedure covers decision points researchers will encounter based on considerations for their data and the missingness mechanisms in their data. We mapped the choices we made during this worked example including within each software package onto each step of the decision tree, which is available in Supporting Information: Table A5.

5.2 | Research questions

We addressed four research questions (RQs) about child- and school-level predictors and longitudinal development of working memory, the complex cognitive ability to maintain and manipulate information in immediately accessible memory systems (Cowan, 2008). Our questions were designed to emulate common developmental research questions. To demonstrate several common considerations and approaches to multiple imputation, we included different types of variables (e.g., binary, ordinal, nominal, continuous and scales) at different levels of analysis (e.g., child-level and school-level). We also looked at interaction effects between variables.

(RQ1) What predicts kindergarten working memory? Expanding on Ahmed et al. (2022), we used a linear regression model to evaluate whether any of several variables predict working memory in the spring of kindergarten.

These variables were: math achievement, age at assessment, disability status, cognitive stimulation, sex, race or ethnicity, household income, parent education, and parent employment. Variables are discussed in detail below.

(RQ2) Does disability status moderate the relation between working memory and math achievement in the spring of kindergarten? We included this research question to demonstrate how to include an interaction term in multiple imputation models. To answer RQ2, we expanded the RQ1 model by including an interaction term evaluating whether disability status moderated the relation between math achievement and working memory.

(RQ3) Do students who attend more economically advantaged schools have higher working memory in the spring of kindergarten? To answer this question, we evaluated a two-level random-intercept model in the spring of kindergarten. In this model, students are nested within schools. We also added an additional school-level predictor, the proportion of students receiving free- or reduced-price school lunch, that serves as a proxy for schoolwide economic advantage.

(RQ4) What kindergarten factors predict growth in working memory from kindergarten to fifth grade? For our final research question, we evaluated a three-level growth curve model. Level 1 modelled the influence of time. Level 2 modelled the longitudinal influence of kindergarten child-level characteristics. Level 3 modelled the longitudinal influence of the proportion of students receiving free- or reduced-price school lunch in the school a given child attended for kindergarten.

5.3 | Sample and variables

Researchers evaluating a question like RQ4 using nested, longitudinal data may need to decide whether they want to allow group membership to be dynamic over time. For our worked example, we chose to restrict the analytic sample to only students who did not change schools[5] between kindergarten and fifth grade. We also chose to remove students who were homeschooled or who did not have a proper school ID (i.e., the data collectors could not locate a child, or a child had moved into a non-sampled county during data collection). With these listwise deletions, our model accounts for time-invariant school-level features in a sample of 7509 students (43% of the original sample). Researchers faced with a similar scenario should choose whichever model best fits the research question, the complexity of the data set, and the ability for their data to meet the congeniality assumption.

As is appropriate to do in circumstances when listwise deletion cannot be avoided, we evaluated how these excluded students differed from our included participants. The students we removed from our analytic sample appeared to have more socioeconomic markers of disadvantage, lower executive function, and lower achievement than students we retained in analyses (Table 3). This means that our results likely only generalize to populations of relatively more advantaged students who remain in the same school from kindergarten to fifth grade. Descriptive sample information for key variables before and after multiple imputation is available in Table 3. The same information for auxiliary variables is available in Supporting Information: Table A2.

5.3.1 | Outcome

Researchers collecting data for the ECLS-K: 2011 measured working memory at each wave using the Numbers Reversed subtest of the Woodcock-Johnson Tests of Achievement (WJ-NR; Mather et al., 2001). At each wave, students were asked to orally repeat increasing number sequences in reverse order, beginning with two-number sequences up to a maximum of eight numbers. Performance was converted into *W*-scores, a standardized scale of equal intervals that is normed to a mean of 500 and a standard deviation of 100 among children aged 10 years 0 months. Because *W*-scores are sensitive to longitudinal change, they can be considered a growth scale and used across multiple age ranges. Younger children will typically display scores below the mean. The working memory variables included in the analytic model were measured by ECLS-K: 2011 data collectors in the spring term of each academic year.

5.3.2 | Key predictors

We included the same key predictors of working memory Ahmed et al. (2022) included in their longitudinal analysis of working memory development: math achievement, male sex, racial or ethnic identity, age at assessment (in months), disability status, household income, parent education level and parent employment status. These are the predictors for RQ1 and RQ4. For RQ2, we include the same predictors in addition to a key interaction term for *disability* by *math*. For RQ3, we expand on RQ1 by adding the schoolwide proportion of students receiving free and reduced-price lunch as a key predictor in a multilevel framework. The models we developed for RQ1–3 model working memory cross-sectionally at kindergarten. Our RQ4 model includes working memory across all timepoints in a longitudinal framework.

Math achievement was directly measured in the spring of kindergarten using Item Response Theory (IRT) procedures in a two-stage assessment. This assessment was designed to capture conceptual knowledge, procedural knowledge, and problem-solving skills ($\alpha = 0.94$). We also included age at assessment in months. Our decision was based on ECLS-K: 2011 recommendations for using direct assessment data (Tourangeau et al., 2015). Racial or ethnic identity was a variable constructed by ECLS-K: 2011 staff. We recoded the variable so that 1 = *White*, 2 = *Black*, 3 = *Hispanic*, 4 = *Asian* and 5 = *Others*. Disability status was included as a binary variable in which parents reported at the spring of kindergarten whether their child was professionally diagnosed with or had received therapy for an emotional, psychological, learning, communicative or developmental difference.[6] Household income was measured in the spring of kindergarten. We treated it as a continuous variable because it contained 18 nearly equal-interval categories ranging from 1 = *$5,000 or less* to 18 = *$200,001 or more*. Parent education level was constructed by ECLS-K: 2011 staff from both fall and spring parent surveys. We treated these data as ordinal and recoded the original values to 1 = *High school diploma or less*, 2 = *Some college*, 3 = *College degree or higher*. Employment was measured in the fall of kindergarten and treated as a nominal variable coded as 1 = *Employed full-time*, 2 = *Employed part-time*, 3 = *Not employed or looking for work*.

In addition to using the same variables as Ahmed et al. (2022), we created a cognitive stimulation scale to demonstrate how multi-item scale variables can be imputed. We averaged together nine items measured at the fall of kindergarten assessing how often any member of the family cognitively engaged with the child. These engagement items included: telling stories; singing songs; helping with arts and crafts; involving the child in household chores; playing games or doing puzzles; talking about nature or science projects; building something or playing with construction toys; playing a sport or exercising together; or practising reading, writing or working with numbers (where 1 = *not at all* and 4 = *every day*).

Finally, we included a proxy marker of student socioeconomic disadvantage in the focal child's school to demonstrate the influence of a school-level predictor in a multilevel model. At each wave, the school administrator reported what percentage of students attending the school received free- or reduced-price lunch (recoded so that 1 = *0%–25%*, 2 = *26%–50%*, 3 = *51%–75%* and 4 = *76%–100%*).

Data were normally distributed on our variables of interest.[7] We did not transform any variables before imputation. Researchers who do encounter non-normality will need to investigate the impact this may have on their analyses and take necessary steps. Non-normality is discussed in section 3.3 of van Buuren (2018) with several references. Readers may also take guidance from Lüdtke et al. (2020), Lun and Khattree (2022) and Lee and Carlin (2010).

5.3.3 | Auxiliary variables

Before conducting any evaluations with the data, we thought about why data may be missing on these key variables (as outlined by Woods et al., 2021; see Supporting Information: Table A5). We hypothesized that data could be MNAR if the child's working memory was too low to complete the direct assessment. In these cases, the child would probably also be missing math achievement and other direct assessment data. To adjust for this possible bias, we

chose a set of auxiliary variables that could approximate low working memory (i.e., variables that could account for or be related to missing observations). These variables included parent and teacher observations of child's behaviour and other aspects of executive functioning. We also included working memory and math achievement from the fall of kindergarten as auxiliary variables, along with a host of other child, home and school-level variables that could influence both patterns of missingness as well as working memory and math. The final number of auxiliary variables differed by software to maximize convergence (see Supporting Information: Table A5).

In addition to our 11 main analytic variables, we tested 18 auxiliary variables to evaluate the MAR missingness mechanism. Eleven of these auxiliary variables contained repeated observations (see Table 2 for a complete list of variables available across timepoints). Repeated observations of auxiliary variables occurred at the fall of kindergarten for both *working memory* and *math* as well as from first to fifth grade for *working memory* (note these are considered main analytic variables for RQ4), *math*, *income*, *disability* and *lunch*. There were four school-level auxiliary variables each with six repeated observations that we hypothesized may predict missingness, particularly on *lunch*; these variables measured neighbourhood disadvantage, Title I funding, proportion of non-white students, and whether the school was public or private. Five kindergarten variables captured parent-rated behaviour, language status, and parenting stress, which could influence longitudinal study attrition. Six repeated observations of parent marital status were reported from kindergarten through fifth grade, and parents rated students' working memory capabilities at third and fourth grade, which could help predict missingness on the direct assessment of working memory. Finally, there were seven repeated observations each of directly assessed executive function and teacher ratings of behaviour.

We created several auxiliary variable composites from this information before imputation. We had many potential auxiliary variables and hoped to minimize convergence issues. We hypothesized that parent- and teacher-reported behaviour and executive function would influence missingness on the direct assessments, including for working memory. However, there were five teacher-reported auxiliary variables at the spring of each wave and three parent-reported variables at the fall and spring of kindergarten. To manage this suite of potential auxiliary variables when using data in the wide format, we conducted additional data exploration of the relations between these variables (e.g., correlation matrices, evaluation of missing patterns among these variables) and distilled them into two auxiliary variables: one parent-reported composite averaging scores from the fall and spring of kindergarten, and one teacher-reported composite at the spring of kindergarten. Each of these composites was created using all available data (e.g., through pairwise deletion) and had high reliability ($\alpha > 0.80$). In testing our auxiliary variables, we discovered that repeated observations of *public*, *non-white* and *neighbourhood disadvantage* variables were collinear with their spring kindergarten values (i.e., strongly enough correlated to cause problems in estimation; $r > 0.80$; Berry & Feldman, 1985), so we dropped these repeated observations to minimize convergence issues.

To evaluate which of these potential auxiliary variables should be included in our imputation models to adjust for MAR data, we created dummy variables for our key predictors and outcome where 1 = *missing*, 0 = *non-missing*. We then conducted *t* tests between these dummy variables and the auxiliary variables as well as examined correlations between these missing dummies and key variables (Supporting Information: Table A3). All tests were conducted in Stata. The results file is available at https://osf.io/j3f8m.

We retained those auxiliary variables that showed significant mean differences between missing and non-missing values on key variables at $p < 0.05$ and with correlations between missing dummies and key variables $r > 0.10$ (i.e., at least a small effect size; Funder & Ozer, 2019). All of our hypothesized auxiliary variables were significantly and meaningfully related to missingness on at least some of our key predictors and outcomes, including longitudinally. For example, there was more study attrition among students attending school in more disadvantaged neighbourhoods relative to students attending school in less disadvantaged neighbourhoods (i.e., students attending kindergarten schools receiving Title I funding were no more or less likely to have missing working memory scores from kindergarten to third grade, but they were more likely to be missing these scores in fourth and fifth grade). Further, there were significant differences between missing and non-missing values on direct cognitive assessments (i.e., math achievement and working memory, as well as lagged predictors of achievement and executive function

measured in the fall of kindergarten) and on parent and teacher ratings of problem behaviours and executive functioning including self-regulation. Consistent with our expectation, this indicates that students with lower executive functioning were less likely to complete the direct assessments, resulting in missing data on these items. Failing to include these auxiliary variables could bias our analyses consistent with MNAR. Incorporating these variables into our imputation model results in a reasonable assumption of MAR.

For the purposes of demonstrating multiple imputation procedures, we chose to multiply impute race/ethnicity and sex. We were only missing 0.23% of cases for race/ethnicity and 0.13% of cases for sex, so this decision was unlikely to substantively impact our results. Other researchers using the ECLS-K: 2011 who have different research questions or different predictors may choose not to impute these variables. Either way, this decision should always be transparent and well-justified or, alternatively, in the absence of a good justification, both ways could be conducted and compared.

5.4 | Sampling weights

Many researchers using large secondary data sets like the ECLS-K: 2011 are interested in making claims about whether their results are nationally representative. This can be accomplished through weighting. For congeniality, if weights are to be used in the analytic model, they should also be used in the imputation model. However, the addition of weights demonstrates an important congeniality issue for our present example. If a researcher were only estimating our RQ1-2, the multiple imputation models would be identically estimated to RQ3 *but* for the weight variable. This is because the ECLS-K user's guide (Tourangeau et al., 2015) instructs researchers to use a child-level weight (e.g., *W1C0*) with single-level kindergarten data and the school-level weight (*W2SCH0*) with nested or multilevel kindergarten data. Because RQ1-2 does not ask about the influence of schools, we need only account for the nested structure of the data by clustering standard errors by school ID in these analyses. We would use a child-level weight if we were to weight these analyses, per the ECLS-K: 2011 user's guide. It would be inappropriate to use the school-level weight in lieu of a child-level weight. Yet, in contrast, we would need to use the school-level weight for RQ3 since we specifically analyze a multilevel model. Without weights, we can estimate one imputation model for RQ1–3 since our RQs and analytic models are all nested. However, our imputation and analytic models would not be congenial if we used one weight in the imputation and a different weight in the analysis. Therefore, researchers who need to weight their data to make nationally representative claims may find that their imputation models differ from those who do not need to weight their data.

Moreover, there is only one school-level weight in the ECLS-K: 2011 because students (not schools) were followed longitudinally. This means that after the kindergarten wave, there is no way to weight the data to obtain a nationally representative sample of US first- to fifth-grade schools. For RQ4, we could opt to use a child-level attrition weight (i.e., one that accounts for both selection into the sample and longitudinal non-response bias). Davis-Kean et al. (2015) recommend against the use of attrition weights because they can diminish sample size and power. Instead, researchers can account for the same factors that influence attrition and retention rates in their multiple imputation models as auxiliary variables and continue to use the base-year selection weights. Thus, for a weighted RQ4, a researcher might include additional variables that can explain this longitudinal attrition alongside the school-level base-year weight for initial selection into the study (*W2SCH0*).

5.5 | Software considerations

We conducted multiple imputation by chained equations (MICE), also known as *fully conditional specification* (FCS), in Stata and R (using the *mice* package). We also conducted multiple imputation via fully Bayesian model-based imputation (MBI) in Blimp. Whereas recommendations have been made with respect to ideal or best practices in multiple imputation, software capabilities differ. Thus, each software produced similar imputation results with some

exceptions. The exact number of both main analytic variables and auxiliary variables included in the imputation model varied slightly by software program. We decided each given computation and convergence concerns as well as how the software handled derived variables (i.e., whether we created the *disability* by *math* interaction term for RQ2 before, during, or after multiple imputation, as well as whether we chose to create the cognitive stimulation scale from its 9 items before or after multiple imputation).

In Blimp, we developed two multiple imputation models: one two-level model for RQ1–3 and one three-level model for RQ4. We conducted multilevel multiple imputation in R. We implemented the dummy-indicator approach with the previously described ad hoc solution for imputing school-level variables in Stata. For RQ4, the longitudinal aspect of the data was handled by working with data in the wide format in both R and Stata, whereas it was handled by working with data in the long format in Blimp. More detailed discussion of specific models and current considerations for each software program are available in the Appendix. Syntax files, results from convergence checks, and all imputed data sets can be found at https://osf.io/j3f8m/. For additional step-by-step information regarding best practices for setting up imputation models and checking for appropriate convergence and results, please see our decision tree at https://doi.org/10.31234/osf.io/mdw5r (Woods et al., 2021). We also overlap this decision tree with the specific decisions made in each software package in Supporting Information: Table A4.

5.5.1 | Stata

To closely approximate multilevel congeniality (Grund et al., 2018; van Buuren, 2011), we created two data sets in Stata v.15.1: one at the individual level (level 1), and one in which the individual level variables were aggregated to a within-school (level 2) mean. We then ran separate imputation models for each level, including the ID variables *childid* (child identification number) and *s_id* (kindergarten school identification number) as predictors in the level 1 imputation model following Stata's recommendations for clustered data. The level 2 imputation model was a single-level model incorporating the same main and auxiliary variables and model specifications (*m*, burn-ins, etc.) as those included at level 1. Following imputation, we recombined these two imputed data sets into one using the *mi merge* command. The level 1 imputation model was congenial with the analysis models for RQ1–2, and the level 2 model combined with the level 1 model was congenial with the analysis model for RQ3–4.

We included all auxiliary variables noted in Table 2 except for repeated observations of age at assessment and fall teacher-reported behaviour given convergence problems. Data for all variables were imputed using predictive mean matching (PMM) with 10 nearest neighbours. The passive imputation approach (*mi passive: generate*) was used to create the cognitive stimulation scale and the *disability* X *math* interaction term following multiple imputation. Consistent with recommendations by White et al. (2011) to set $m > 100$ times the highest fraction of missing information (FMI), we imputed $m = 40$ data sets. Convergence appeared adequate based on visual inspection of convergence plots (for an illustration of sufficient vs. non-sufficient convergence, see Nassiri et al., 2020; van Buuren & Groothuis-Oudshoorn, 2011) and plots of imputed versus observed values (e.g., box plots, scatterplots and density distributions) were similar between imputed and observed values. The distributions of imputed values can slightly differ from observed values given the reduction in bias due to missing data, but these differences should not be unreasonable. Anomalies evident in a few imputations but not others would indicate problems with the imputation model; White et al. (2010).

5.5.2 | R (*mice*)

The *mice* package v3.13.0 (van Buuren & Groothuis-Oudshoorn, 2011) in R uses FCS to multiply impute data. *mice* provides imputation methods for different types of variables (nominal, ordinal and continuous) and different levels of categorical variables. Each variable has its own model. The *mice* imputation model experienced issues with

convergence given sparse cell sizes for some auxiliary variables; convergence was achieved by removing repeated observations of variables capturing school Title I funding status from first to fifth grade, school neighbourhood disadvantage, parent-reported working memory at third and fourth grade, and all repeated observations of single parent status. To specify the model, we developed a *predictor matrix* using the *quickpred* function in *mice*. The predictor matrix is a matrix that defines the equations that will be used to impute each variable. The *quickpred* function analyzes the correlations between variables to define which variables should be included in the imputation model for each other variable. We were unable to achieve convergence with a three-level model in long format (i.e., perfectly congenial to RQ4). To create two-level imputation models in wide format, we defined the school ID as the nesting variable in the predictor matrix. We then identified the imputation algorithms as *pmm* for our level 1 variables and *2lonly.norm* for our level 2 variables. We ran $m = 30$ imputations, *maxit* = 30 iterations, and used the default burn-in of 5000. Visual inspection of the model plots indicates that the models reached sufficient convergence.

5.5.3 | Blimp

Blimp (Keller & Enders, 2021) has two algorithms: FCS and MBI (Enders et al., 2020). With FCS, the process of specifying the multiple imputation model is similar to *mice* in R, but cannot accommodate nonlinear terms (e.g., interaction effects, random slopes, polynomial terms, etc.). In contrast, MBI allows one to specify complex models up to three levels with non-linear terms. For congeniality to our RQs, we elected to use MBI.

We specified one random intercept imputation model in long format in Blimp, Version 2 for RQ1–3. We specified the school ID as a cluster variable and working memory as an outcome. We included our main predictors (coded as nominal or ordinal where appropriate), 11 auxiliary variables (9 at the student level and 2 at the school level, all measured during the fall and/or spring of kindergarten), and an interaction effect between disability and mathematics achievement for congeniality with RQ2. We imputed $m = 30$ data sets. We set a very large number of burn-in iterations (50,000) to solve convergence problems, which led to an acceptable potential scale reduction (psr) factor of 1.074 (Gelman & Rubin, 1992; Keller & Enders, 2021).

We modified the syntax for RQ1–3 to produce a multiple imputation model congenial with RQ4 by adding student ID as a cluster variable, adding a time variable to model linear growth, and specifying a random slope for time along with interaction effects of analytic predictors of growth and time. As in the model for RQ1–3, we imputed $m = 30$ data sets and set the number of burn-in iterations to 50. This model was extremely computationally intensive but eventually reached convergence. The psr factor was 1.061, which we deemed acceptable.

6 | RESULTS

Descriptive statistics for auxiliary variables and derived variables for the sample after multiple imputation in each software program is compared to the complete case sample in Table 3. Based on these descriptive statistics, we find that complete case analysis would restrict the sample to include more white students, students with higher working memory scores, and students from households with higher parental education and income. The imputation model for *mice* was the only model to not include the kindergarten measure of school neighborhood disadvantage given convergence problems. Potentially highlighting the importance of auxiliary variables, results produced by *mice* display some differences in disability, employment, and education values (Table 3). This might have led to slight variations in analysis results (Tables 4-6).

The average per cent of missing observations across analytic variables was 16.8%, ranging from a low of 0.1% for race and a high of 28.0% on employment status. We would have retained only 59% of our sample under listwise deletion methods for RQ1–2. Twenty-five per cent of cases were missing parent survey items (i.e., presumably from attrition non-response in failing to return the entire survey rather than item non-response on individual questions;

14% of these cases were missing responses from the fall of kindergarten and 9% were missing both fall and spring waves). Nine per cent of cases were only missing kindergarten income and disability status, and an additional 2% were missing only kindergarten disability status. Finally, 1% of cases were missing all items from the spring of kindergarten (direct assessment and parent survey responses). The remaining 6% of cases had no discernable pattern in non-response (e.g., could have been due to selectively or inadvertently skipping a question, coding errors, etc.). For RQ3, we would have retained 49% of our sample under listwise deletion methods. Seventeen per cent of cases were missing parent survey items (10% from the fall of kindergarten, and 7% from the spring of kindergarten). Sixteen per cent of cases were missing school administrator data (10% missing only the administrator survey, and an additional 6% missing fall and/or spring parent surveys). Ten per cent of cases were missing disability status either alone (1%) or in combination with income (7%) and administrator data (2%). The remaining 8% of cases had no discernable pattern in non-response. For RQ4, which added repeated observations of our dependent variable working memory, we would have retained only 46% of our measurement occasions under listwise deletion methods. Twenty-two per cent of these observations were missing kindergarten parent survey data (9% were missing information from the fall of kindergarten, 7% were missing information from the spring of kindergarten, and 6% were missing all parent survey information). Eighteen per cent of these observations were missing school administrator data (10% of cases missing only school administrator data, and an additional 8% missing school administrator and fall and/or spring parent survey data). One per cent of these observations were missing only disability status. The remaining 14% had no discernible missing pattern.

6.1 | Developmental predictors of working memory

For each research question, there were few dissimilarities across results from different software packages but marked differences in the pattern of results compared to complete case analysis. For RQ1–3, estimated gaps in working memory for racially minoritized students were overestimated relative to white students using complete case analysis. The magnitude of these differences sometimes affected statistical significance. In the models for RQ1 and RQ2 (Table 4), family income was a significant predictor of working memory after (but not before) multiple imputation, and the effects of parent education were overestimated in complete case analysis. In RQ3 (Table 5), the effect of the school-level predictor of economic disadvantage (per cent of students receiving free or reduced-price lunch) significantly predicted kindergarten working memory scores after multiple imputation but not in complete case analysis. Interestingly, in the model for RQ3, complete case analysis would lead researchers to conclude that age of assessment was a significant predictor of working memory. After accounting for missing data using multiple imputation, we observe marked differences in effects for the sociodemographic variables of parent employment, parent education, child sex and cognitive stimulation. Examining predictors of trajectories of working memory in RQ4 (Table 6) via a growth curve modelling approach again reveals minimal differences in multiple imputation results across software programs, but larger differences in results between multiple imputation and listwise deletion. As shown in Table 6, the pattern is not entirely consistent, but listwise deletion appeared to underestimate the effect of parent unemployment, cognitive stimulation and free lunch, and overestimate the effects of age, part-time employment and parent education.

For each research question, listwise deletion might lead researchers to overestimate the working memory gap between students from different sociodemographic and socioeconomic backgrounds. For example, researchers using complete case analysis might overestimate the working memory gap between white and minoritized students, particularly Black or Hispanic students, despite the fact that less than 1% of these observations were missing. Similarly, complete case analysis would lead researchers to overestimate the gap by parent education level as well as underestimate the effect of parent unemployment as well as schoolwide socioeconomic disadvantage. Thus, even when using population data like the ECLS-K: 2011, failure to adjust for missing data can introduce bias into results and analysis, particularly on important sociodemographic predictors. Although we found results to be mostly similar across imputation models in different software, other studies with different analytic models and data may have produced discrepant results. Future

methodological research on fairly complex analytic models is needed to evaluate the effects of different imputation models implemented in different software packages on the bias in parameter estimates and standard errors. Recommendations from simulation studies will be helpful for applied researchers when choosing between different imputation models.

7 | CONCLUSIONS

Missing data are ubiquitous in developmental research. The choice of how to address missing data is as crucial to the validity of results as the choice of analysis. The goal of this paper was to elucidate the importance of addressing missing data, to outline recommended multiple imputation reporting standards (e.g., Box 2), and to provide worked software examples across multiple approaches to handling missing data. Our recommendations are applicable to all social scientists but are critical for developmental scientists who often use complex analytic models.

Even when researchers do not explicitly adjust for missing data, there is often still an adjustment for missingness made in analyses that can impact results (e.g., software programs usually default to deletion methods when nothing else is specified). We argue that this process should be conscious and well-informed given the ethical, practical and moral implications of ignoring missing data. Because the choice of how to handle missing data can have important effects on the accuracy and precision of one's inferences, researchers should not only carefully consider *why* they are implementing a chosen method, but also how such decisions will affect their final study outcomes. We recommend that decisions be clearly communicated, driven by theory including a thorough conceptual understanding of one's data, and delineated at the level of the proposed analysis rather than specified for a data set as a whole. Because there are many potential decisions, researchers should conduct sensitivity analyses such as applying different decision-making rules to test the robustness of results (e.g., threshold of significance or meaningful effect sizes for the inclusion of auxiliary variables) or examining samples and model results before and after multiple imputation. We also recommend that researchers incorporate open science practices into multiple imputation so that others may replicate their work (e.g., pre-registering imputation decisions or conducting registered reports, openly sharing data and code). Overall, despite the many potential decisions that can be made in the process of multiply imputing data, choosing to *not* consider robust ways of addressing missingness is a decision that is likely to have more serious consequences than using one type of approach (e.g., multiple imputation, FIML) or algorithm (e.g., MICE) over another.

In sum, addressing missing data appropriately takes additional time, effort and thought, and involves additional analysis steps than what is done automatically in most programs. Such barriers may prevent adoption of multiple imputation for many researchers, but any well-designed analysis or study design includes consideration of missing data. We hope our guidance will inspire researchers to question their default practices, describe and justify their approach to missing data when reporting results, and implement multiple imputation in future analyses. Appropriately addressing missing data is key to transparent analyses and to engaging in the most robust, most unbiased science possible.

AUTHOR CONTRIBUTIONS

Adrienne D. Woods: Conceptualization; data curation; formal analysis; methodology; project administration; software; supervision; writing – original draft; writing – review and editing. **Daria Gerasimova:** Formal analysis; methodology; software; writing – original draft; writing – review and editing. **Ben Van Dusen:** Formal analysis; software; writing – original draft; writing – review and editing. **Jayson Nissen:** Formal analysis; software; writing – original draft; writing – review and editing. **Sierra Bainter:** Formal analysis; software; writing – original draft; writing – review and editing. **Alex Uzdavines:** Writing – original draft; writing – review and editing. **Pamela Davis-Kean:** Conceptualization; writing – original draft; writing – review and editing. **Max Halvorson:** Conceptualization; writing – review and editing. **Kevin King:** Conceptualization; writing – original draft; writing – review and editing. **Jessica Logan:** Conceptualization; writing – original draft; writing – review and editing. **Menglin Xu:** Conceptualization; writing – original draft; writing – review and

editing. **Martin R. Vasilev:** Writing – original draft; writing – review and editing. **James M. Clay:** Writing – original draft; writing – review and editing. **David Moreau:** Writing – original draft; writing – review and editing. **Keven Joyal-Desmarais:** Writing – original draft; writing – review and editing. **Rick A. Cruz:** Writing – original draft; writing – review and editing. **Denver M. Y. Brown:** Writing – original draft; writing – review and editing. **Kathleen Schmidt:** Writing – original draft; writing – review and editing. **Mahmoud M. Elsherif:** Writing – original draft; writing – review and editing.

AFFILIATIONS

[1]Center for Learning and Development, Education, SRI International, Arlington, Virginia, USA
[2]Kansas University Center on Developmental Disabilities, University of Kansas, Lawrence, Kansas, USA
[3]School of Education, Iowa State University, Ames, Iowa, USA
[4]Nissen Education Research and Design, Corvallis, Oregon, USA
[5]Department of Psychology, University of Miami, Coral Gables, Florida, USA
[6]South Central Mental Illness Research, Education, and Clinical Center, Michael E. DeBakey VA Medical Center, Houston, Texas, USA
[7]Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, Texas, USA
[8]Department of Psychology, University of Michigan, Ann Arbor, Michigan, USA
[9]Department of Psychology, University of Washington, Seattle, Washington, USA
[10]Department of Educational Studies, The Ohio State University, Columbus, Ohio, USA
[11]Department of Internal Medicine, The Ohio State University, Columbus, Ohio, USA
[12]Department of Psychology, Bournemouth University, Bournemouth, UK
[13]Department of Psychology, University of Portsmouth, Portsmouth, UK
[14]School of Psychology, University of Auckland, Auckland, New Zealand
[15]Centre for Brain Research, University of Auckland, Auckland, New Zealand
[16]Department of Health, Kinesiology, and Applied Physiology, Concordia University, Montreal, Quebec, Canada
[17]Montreal Behavioral Medicine Centre, Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, Montreal, Quebec, Canada
[18]Department of Psychology, Arizona State University, Tempe, Arizona, USA
[19]Department of Psychology, University of Texas at San Antonio, San Antonio, Texas, USA
[20]School of Psychological and Behavioral Sciences, Southern Illinois University, Carbondale, Illinois, USA
[21]Department of Psychology, University of Birmingham, Birmingham, UK

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/icd.2407.

DATA AVAILABILITY STATEMENT

Data are publicly available and are maintained by the National Center for Education Statistics. Our cleaned data sets and all code for Stata, R and Blimp are available at our osf.io page (https://osf.io/j3f8m/).

ORCID

Adrienne D. Woods https://orcid.org/0000-0003-1101-6975
Daria Gerasimova https://orcid.org/0000-0002-9669-1648
Ben Van Dusen https://orcid.org/0000-0003-1264-0550
Jayson Nissen https://orcid.org/0000-0003-3507-4993
Sierra Bainter https://orcid.org/0000-0001-7461-0803
Alex Uzdavines https://orcid.org/0000-0001-5829-9648
Pamela E. Davis-Kean https://orcid.org/0000-0001-8389-6268
Max Halvorson https://orcid.org/0000-0002-3113-2458

Kevin M. King https://orcid.org/0000-0001-8358-9946
Jessica A. R. Logan https://orcid.org/0000-0003-3113-4346
Martin R. Vasilev https://orcid.org/0000-0003-1944-8828
James M. Clay https://orcid.org/0000-0002-3392-5099
David Moreau https://orcid.org/0000-0002-1957-1941
Keven Joyal-Desmarais https://orcid.org/0000-0003-0657-8367
Rick A. Cruz https://orcid.org/0000-0001-9139-8170
Denver M. Y. Brown https://orcid.org/0000-0003-4078-8253
Kathleen Schmidt https://orcid.org/0000-0002-9946-5953
Mahmoud M. Elsherif https://orcid.org/0000-0002-0540-3998

ENDNOTES

[1] But see Jakobsen et al. (2017) for a differing opinion.

[2] This is also exacerbated when papers fail to report reasonable descriptive statistics, making it impossible to determine whether and how listwise deletion is further limiting statistical power for minoritized groups.

[3] This discussion pertains to imputation of social identifiers rather than using social identifiers as auxiliary variables. Regardless of whether the auxiliary variables are social identifiers, using variables that predict missing observations on a given variable should result in more precise imputed values.

[4] Ahmed et al. (2022) capitalized on the planned missing design of the ECLS-K: 2011, where data were only collected on a random subsample of students in the fall of first grade (wave 3) and the fall of second grade (wave 5). In a departure from Ahmed's analyses and for simplicity in demonstration, we do not use data from the planned missing waves 3 and 5, but instead use data from waves 1 and 2 (fall and spring of kindergarten), 4, 6, 7, 8 and 9 (spring of first, second, third, fifth and fifth grade, respectively). Readers interested in an example of imputation with planned missing data should consult Ahmed et al. (2022).

[5] In this worked example, allowing group membership to vary over time would have produced a more complex model for RQ4 (e.g., a cross-classified random effects model where group membership within schools may change over time instead of a linear growth model; see Cafri et al., 2015). Retaining these participants who changed schools by running a more complex model is a commendable goal for a substantive study but is beyond the scope of our worked methodological example.

[6] The term *disorder* was used in the parent questionnaire. To enhance inclusivity, we follow the neurodiverse movement and the social model of disability by using the word *difference* instead of disorder (Elsherif et al., 2022).

[7] Distributional assumptions are usually placed only on the residuals of the dependent variable Y in typical models. In imputation models, the independent variables (*X*) take turns serving as the 'outcome' to be imputed, so distributional assumptions also apply to the residuals of any *X* variable that is imputed.

REFERENCES

Adelson, J. L., Barton, E., Bradshaw, C., Bryant, B., Bryant, D., Cook, B. G., Coyne, M., DeBettencourt, L., DeHaven, A. C., Dymond, S., Esposito, J., Farmer, T., Flake, J. K., Gage, N. A., Kennedy, M., Kern, L., Lane, K. L., Lee, D., Lembke, E., … Troia, G. A. (2019). A roadmap for transparent research in special education and related disciplines [Preprint]. *EdArXiv*. https://doi.org/10.35542/osf.io/sqfy3

Ahmed, S. F., Ellis, A., Ward, K. P., Chaku, N., & Davis-Kean, P. E. (2022). Working memory development from early childhood to adolescence using two nationally representative samples. *Developmental Psychology*, *58*(10), 1962–1973. https://doi.org/10.1037/dev0001396

Altmann, D. G., & Bland, J. M. (2007). Missing data. *British Medical Journal*, *334*(7590), 424. https://doi.org/10.1136/bmj.38977.682025.2C

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*, 3–25. https://doi.org/10.1037/amp0000191

Azevedo, F., Liu, M., Pennington, C. R., Pownall, M., Evans, T. R., Parsons, S., Elsherif, M., Micheli, L., Westwood, S. J., & FORRT. (2022). Towards a culture of open scholarship: The role of pedagogical communities. *BMC Research Notes*, *15*, 75. https://doi.org/10.1186/s13104-022-05944-1

Azevedo, F., Parsons, S., Micheli, L., Strand, J., Rinke, E., Guay, S., Elsherif, M., Quinn, K., Wagge, J. R., Steltenpohl, C., Kalandadze, T., Vasilev, M., de Oliveira, C. F., Aczel, B., Miranda, J., Galang, C. M., Baker, B. J., Pennington, C. R., Marques, T., … FORRT. (2019). *Introducing a Framework for Open and Reproducible Research Training* (*FORRT*). https://doi.org/10.31219/osf.io/bnh7p

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. https://doi.org/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. https://doi.org/10.1016/j.jsp.2009.10.001

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348

Baum, M., Hart, A., Elsherif, M., Ilchovska, Z., Moreau, D., Dokovova, M., LaPlume, A. A., Krautter, K., & Staal, J. (2022). Research without borders: How to identify and overcome potential pitfalls in international large-team online research projects. In *SAGE Research Methods Cases*. SAGE Publications, Ltd.

Benford, R., & Gess-Newsome, J. (2006). *Factors affecting student academic success in gateway courses at Northern Arizona University* (*ERIC Document Reproduction Service No. ED495693*).

Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice*. SAGE Publications.

Bhaskaran, K., & Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, *43*(4), 1336–1339. https://doi.org/10.1093/ije/dyu080

Bodner, T. E. (2006). Missing data: Prevalence and reporting practices. *Psychological Reports*, *99*(3), 675–680. https://doi.org/10.2466/PR0.99.3.675-680

Brown, K. S., Su, Y., Jagganath, J., Rayfield, J., & Randall, M. (2021). *Ethics and empathy in using imputation to disaggregate data for racial equity*. Urban Institute. https://www.urban.org/sites/default/files/publication/104678/ethics-and-empathy-in-using-imputation-to-disaggregate-data-for-racial-equity.pdf

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. https://doi.org/10.1038/sj.bjc.6601907

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cafri, G., Hedeker, D., & Aarons, G. A. (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological Methods*, *20*(4), 407–421. https://doi.org/10.1037/met0000043

Cham, H., Reshetnyak, E., Rosenfeld, B., & Breitbart, W. (2017). Full information maximum likelihood estimation for latent variable interactions with incomplete indicators. *Multivariate Behavioral Research*, *52*, 12–30. https://doi.org/10.1080/00273171.2016.1245600

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487–508. https://doi.org/10.3102/0034654314532697

Collins, L., Schafer, J., & Kam, C.-M. (2001). A comparison of restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9

Curran, D., Bacchi, M., Schmitz, S. F., Molenberghs, G., & Sylvester, R. J. (1998). Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine*, *17*(5–7), 739–756. https://doi.org/10.1002/(sici)1097-0258(19980315/15)17:5/7<739::aid-sim818>3.0.co;2-m

Curran, D., Molenberghs, G., Fayers, P. M., & Machin, D. (1998). Incomplete quality of life data in randomized trials: Missing forms. *Statistics in Medicine*, *17*(5–7), 697–709. https://doi.org/10.1002/(sici)1097-0258(19980315/15)17:5/7<697::aid-sim815>3.0.co;2-y

Davis-Kean, P. E., Jager, J., & Maslowsky, J. (2015). Answering developmental questions using secondary data. *Child Development Perspectives*, *9*, 256–261. https://doi.org/10.1111/cdep.12151

Delios, A., Clemente, E. G., Wu, T., Tan, H., Wang, Y., Gordon, M., Viganola, D., Chen, Z., Dreber Johannesson, M., Pfeiffer, T., Generalizability Tests Forecasting Collaboration, & Uhlmann, E. L. (2022). Examining the generalizability of research findings from archival data. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(30), e2120377119. https://doi.org/10.1073/pnas.2120377119

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, *22*(16), 2553–2575. https://doi.org/10.1002/sim.1475

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*, 222. https://doi.org/10.1186/2193-1801-2-222

Dynan, K. E., & Rouse, C. E. (1997). The underrepresentation of women in economics: A study of undergraduate economics students. *The Journal of Economic Education*, *28*(4), 350–368. https://doi.org/10.1080/00220489709597939

Elsherif, M. M., Middleton, S. L., Phan, J. M., Azevedo, F., Iley, B. J., Grose-Hodge, M., Kapp, S. K., Gourdon-Kanhukamwe, A., Grafton-Clarke, D., Yeung, S. K., Shaw, J. J., Hartmann, H., & Dokovova, M. (2022). Bridging Neurodiversity and Open Scholarship: How shared values can Guide best practices for research integrity, social justice, and principled education. *MetaArXiv*. https://doi.org/10.31222/osf.io/k7a9p

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

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*(1), 1–16. https://doi.org/10.1037/a0022640

Enders, C. K. (2016). A review of handbook of missing data methodology. *Journal of Educational and Behavioral Statistics*, *41*(5), 554–556. https://doi.org/10.3102/1076998616646650

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy*, *98*, 4–18. https://doi.org/10.1016/j.brat.2016.11.008

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, *25*(1), 88.

Fairclough, D. L., Peterson, H. F., & Chang, V. (1998). Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine*, *17*(5–7), 667–677. https://doi.org/10.1002/(sici)1097-0258(19980315/15)17:5/7<667::aid-sim813>3.0.co;2-6

Fielding, S., Fayers, P. M., McDonald, A., McPherson, G., & Campbell, M. K. (2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, *6*(1), 57. https://doi.org/10.1186/1477-7525-6-57

Findley, M. G., Jensen, N. M., Malesky, E. J., & Pepinsky, T. B. (2016). Can results-free review reduce publication bias? The results and implications of a pilot study. *Comparative Political Studies*, *49*(13), 1667–1703. https://doi.org/10.1177/0010414016655539

Ford, K.-K. (2001). "First, do no harm"—The fiction of legal parental consent to genital-normalizing surgery on intersexed infants. *Yale Law & Policy Review*, *19*(2), 469–488. https://heinonline.org/HOL/P?h=hein.journals/yalpr19&i=479

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research*, *47*(1), 1–25. https://doi.org/10.1080/00273171.2012.640589

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

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Grol, R., & Wensing, M. (2004). What drives change? Barriers to and incentives for achieving evidence-based practice. *The Medical Journal of Australia*, *180*(S6), S57–S60. https://doi.org/10.5694/j.1326-5377.2004.tb05948.x

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics*, *43*(3), 316–353. https://doi.org/10.3102/1076998617738087

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153. https://doi.org/10.2307/1912352

Heitjan, D. F., & Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, *50*(3), 207–213. https://doi.org/10.1080/00031305.1996.10474381

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, *15*(5), 615–625. https://doi.org/10.1097/01.ede.0000135174.63482.43

Howell, D. C. (2007). The treatment of missing data. In W. Outhwaite & S. Turner (Eds.), *The Sage handbook of social science methodology* (pp. 208–224). SAGE Publications.

Hutchins, L. F., Unger, J. M., Crowley, J. J., Coltman, C. A., & Albain, K. S. (1999). Underrepresentation of patients 65 years of age or older in cancer-treatment trials. *New England Journal of Medicine*, *341*(27), 2061–2067. https://doi.org/10.1056/NEJM199912303412706

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials—A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*, 1–10. https://doi.org/10.1186/s12874-017-0442-1

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(4), 1195–1199. https://doi.org/10.1037/a0015665

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Kanim, S., & Cid, X. C. (2020). Demographics of physics education research. *Physical Review Physics Education Research*, *16*(2), 020106. https://doi.org/10.1103/PhysRevPhysEducRes.16.020106

Karahalios, A., Baglietto, L., Carlin, J. B., English, D. R., & Simpson, J. A. (2012). A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology*, *12*, 96. https://doi.org/10.1186/1471-2288-12-96

Katzmarzyk, P. T., Denstel, K. D., Martin, C. K., Newton, R. L., Jr., Apolzan, J. W., Mire, E. F., Horswell, R., Johnson, W. D., Brown, A. W., Zhang, D., & PROPEL Research Group. (2022). Intraclass correlation coefficients for weight loss cluster randomized trials in primary care: The PROPEL trial. *Clinical Obesity*, e12524. https://doi.org/10.1111/cob.12524

Keller, B. T., & Enders, C. K. (2021). *Blimp user's guide* (Version 3) [Computer software]. www.appliedmissingdata.com/multilevel-imputation.html

King, K. M., Pullmann, M. D., Lyon, A. R., Dorsey, S., & Lewis, C. C. (2019). Using implementation science to close the gap between the optimal and typical practice of quantitative methods in clinical science. *Journal of Abnormal Psychology*, *128*(6), 547. https://doi.org/10.1037/abn0000417

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics – Physics Education Research*, *5*(1), 010101. https://doi.org/10.1103/PhysRevSTPER.5.010101

Kost-Smith, L. E., Pollock, S. J., Finkelstein, N. D., Cohen, G. L., Ito, T. A., Miyake, A., Singh, C., Sabella, M., & Rebello, S. (2010). Gender differences in physics 1: The impact of a self-affirmation intervention. *AIP Conference Proceedings*, *1289*(1), 197–200. https://doi.org/10.1063/1.3515197

Lang, K. M., & Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, *19*, 284–294. https://doi.org/10.1007/s11121-016-0644-5

Ledgerwood, A., Hudson, S. K. T. J., Lewis, N., Jr., Maddox, K., Pickett, C., Remedios, J., Cheryan, S., Diekman, A. B., Dutra, N. B., Goh, J. X., Goodwin, S. A., Munakata, Y., Navarro, D. J., Onyeador, I. N., Srivastava, S., & Wilkins, C. L. (2022). The pandemic as a portal: Reimagining psychological science as truly open and inclusive. *Perspectives on Psychological Science*, *17*(4), 937–959. https://doi.org/10.1177/17456916211036654

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. https://doi.org/10.1093/aje/kwp425

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., Hogan, J. W., & Carpenter, J. R. (2021). Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology*, *134*, 79–88. https://doi.org/10.1016/j.jclinepi.2021.01.008

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods*, *26*(4), 466–485. https://doi.org/10.1037/met0000381

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125. https://doi.org/10.2307/2290705

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*, 141–165. https://doi.org/10.1037/met0000096

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*(2), 157–181. https://doi.org/10.1037/met0000233

Lun, Z., & Khattree, R. (2022). A general approach for imputation of non-normal continuous data based on copula transformation. *Communications in Statistics – Simulation and Computation*, 1–28. https://doi.org/10.1080/03610918.2022.2025839

Manly, C. A., & Wells, R. S. (2015). Reporting the use of multiple imputation for missing data in higher education research. *Research in Higher Education*, *56*(4), 397–409. https://doi.org/10.1007/s11162-014-9344-9

Mather, N., Wendling, B. J., & Woodcock, R. W. (2001). *Essentials of WJ III tests of achievement assessment*. John Wiley & Sons.

Mavridis, D., Chaimani, A., Efthimiou, O., Leucht, S., & Salanti, G. (2014). Addressing missing outcome data in meta-analysis. *Evidence-Based Mental Health*, *17*(3), 85–89. https://doi.org/10.1136/eb-2014-101900

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558. https://doi.org/10.1214/ss/1177010269

Mertens, G., & Krypotos, A.-M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, *59*(1), 338–352. https://doi.org/10.5334/pb.493

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, *5*(4), 297–310. https://doi.org/10.1080/19312458.2011.624490

Nassiri, V., Molenberghs, G., Verbeke, G., & Barbosa-Breda, J. (2020). Iterative multiple imputation: A framework to determine the number of imputed datasets. *The American Statistician*, *74*(2), 125–136. https://doi.org/10.1080/00031305.2018.1543615

National Academy of Sciences. (2011). *Expanding underrepresented minority participation: America's science and technology talent at the crossroads*. National Academies Press. https://doi.org/10.17226/12984

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2021). Practical strategies for handling breakdown of multiple imputation procedures. *Emerging Themes in Epidemiology*, *18*(1), 5. https://doi.org/10.1186/s12982-021-00095-3

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, *41*, 143–153. https://doi.org/10.1177/0165025415618275

Nissen, J. M., Jariwala, M., Close, E. W., & Dusen, B. V. (2018). Participation and performance on paper- and computer-based low-stakes assessments. *International Journal of STEM Education*, *5*(1), 21. https://doi.org/10.1186/s40594-018-0117-4

Nissen, J. M., & Shemwell, J. T. (2016). Gender, experience, and self-efficacy in introductory physics. *Physical Review Physics Education Research*, *12*(2), 020105. https://doi.org/10.1103/PhysRevPhysEducRes.12.020105

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Oberman, H. I., van Buuren, S., & Vink, G. (2021). *Missing the point: Non-convergence in iterative imputation algorithms* (*preprint*). https://arxiv.org/pdf/2110.11951.pdf

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahimet, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Pedersen, A., Mikkelsen, E., Cronin-Fenton, D., Kristensen, N., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157–166. https://doi.org/10.2147/CLEP.S129785

Pownall, M., Azevedo, F., Aldoh, A., Elsherif, M. M., Vasilev, M. R., Pennington, C. R., Robertson, O. M., Vel Tromp, M., Liu, M., Makel, M. C., Tonge, N. A., Moreau, D., Horry, R., Shaw, J. J., Tzavella, L., McGarrigle, R., Talbot, C. V., & Parsons, S. (2021). Embedding open and reproducible science into teaching: A bank of lesson plans and resources. *Scholarship of Teaching and Learning*. [Advance online publication]. https://doi.org/10.1037/stl0000307

Pownall, M., Azevedo, F., König, L. M., Slack, H. R., Evans, T. R., Flack, Z., Grinschg, S., Elsherif, M. M., Gilligan-Lee, K. A., de Oliveira, C. M. F., Gjoneska, B., Kalandadze, T., Button, K., Ashcroft-Jones, S., Terry, J., Albayrak-Aydemir, N., Děchtěrenko, F., Alzahawi, S., Baker, B. J., … FORRT. (2022). The impact of open and reproducible scholarship on students' scientific literacy, engagement, and attitudes towards science: A review and synthesis of the evidence. *Perspectives on Psychological Science*. https://doi.org/10.31222/osf.io/9e526

Proctor, E. K., Landsverk, J., Aarons, G., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy in Mental Health*, *36*(1), 24–34. https://doi.org/10.1007/s10488-008-0197-4

Puthillam, A., Montilla Doble, L. J., Delos Santos, J. I., Elsherif, M. M., Steltenpohl, C. N., Moreau, D., Pownall, M., & Kapoor, H. (2022, August 1). Guidelines to improve internationalization in psychological science. *PsyArXiv*. https://doi.org/10.31234/osf.io/2u4h5

Randall, M., Stern, A., & Su, Y. (2021). *Five ethical risks to consider before filling missing race and ethnicity data*. Urban Institute. https://www.urban.org/research/publication/five-ethical-risks-consider-filling-missing-race-and-ethnicity-data

Raykov, T. (2011). On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 419–429. https://doi.org/10.1080/10705511.2011.582396

Rhemtulla, M., & Hancock, G. R. (2016). Planned missing data designs in educational psychology research. *Educational Psychologist*, *51*(3–4), 305–316. https://doi.org/10.1080/00461520.2016.1208094

Rhemtulla, M., & Little, T. (2012). Tools of the trade: Planned missing data designs for research in cognitive development. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, *13*(4), 425–438. https://doi.org/10.1080/15248372.2012.717340

Rhodes, W. (2015). Improving disparity research by imputing missing data in health care records. *Health Services Research*, *50*(4), 939. 10.1111%2F1475-6773.12336

Rioux, C., & Little, T. D. (2021). Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development*, *45*, 51–58. https://doi.org/10.1177/0165025419880609

Rombach, I., Gray, A. M., Jenkinson, C., Murray, D. W., & Rivero-Arias, O. (2018). Multiple imputation for patient reported outcome measures in randomised controlled trials: Advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC Medical Research Methodology*, *18*, 87. https://doi.org/10.1186/s12874-018-0542-6

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15. https://doi.org/10.1177/096228029900800102

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

Shih, M., & Sanchez, D. T. (2009). When race becomes even more complex: Toward understanding the landscape of multiracial identity and experiences. *Journal of Social Issues*, *65*, 1–11. https://doi.org/10.1111/j.1540-4560.2008.01584.x

Singer, S., & Smith, K. A. (2013). Discipline-based education research: Understanding and improving learning in undergraduate science and engineering: discipline-based education research. *Journal of Engineering Education*, *102*(4), 468–471. https://doi.org/10.1002/jee.20030

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, *338*, b2393. https://doi.org/10.1136/bmj.b2393

Sterner, W. R. (2011). What is missing in counseling research? Reporting missing data. *Journal of Counseling & Development*, *89*, 56–62. https://doi.org/10.1002/j.1556-6678.2011.tb00060.x

Syed, M. (2021). Infant and Child Development: A journal for open, transparent and inclusive science from prenatal through emerging adulthood. *Infant and Child Development*, *30*, e2215. https://doi.org/10.1002/icd.2215

Tierney, W., Hardy, J., Ebersole, C. R., Viganola, D., Clemente, E. G., Gordon, M., Hoogeveen, S., Haaf, J., Dreber, A., Johannesson, M., Pfeiffer, T., Huang, J. L., Vaughn, L. A., DeMarree, K., Igou, E. R., Chapman, H., Gantman, A., Vanaman, M., Wylie, J., … Uhlmann, E. L. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology*, *93*, 104060. https://doi.org/10.1016/j.jesp.2020.104060

Tierney, W., Hardy, J. H., Ebersole, C. R., Leavitt, K., Viganola, D., Clemente, E. G., Gordon, M., Dreber, A., Johannesson, M., Pfeiffer, T., & Uhlmann, E. L. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes*, *161*, 291–309. https://doi.org/10.1016/j.obhdp.2020.07.002

Topor, M., Pickering, J. S., Barbosa Mendes, A., Bishop, D. V. M., Büttner, F. C., Elsherif, M. M., Evans, T. R., Henderson, E. L., Kalandadze, T., Nitschke, F. T., Staaks, J., Van den Akker, O., Yeung, S. K., Zaneva, M., Lam, A., Madan, C. R., Moreau, D., O'Mahony, A., Parker, A. J., … Westwood, S. J. (2020). An integrative framework for planning and conducting Non-Intervention, Reproducible, and Open Systematic Reviews (NIRO-SR) [Preprint]. *MetaArXiv*. https://doi.org/10.31222/osf.io/8gu5z

Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., Mary, C., Daly, P., & Najarian, M. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011): User's Manual for the ECLS-K: 2011 Kindergarten-Fourth Grade Data File and Electronic Codebook Public Version (NCES 2015-074)*. National Center for Education Statistics. https://nces.ed.gov/ecls/kindergarten2011.asp

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (Vol. 10, pp. 173–196). Routledge.

van Buuren, S. (2018). *Flexible imputation of missing data* (Second ed.). CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Multivariate imputation by chained equations. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.1177/0962280206074463

van Buuren, S., & Oudshoorn, C. G. M. (2000). *Multivariate imputation by chained equations: MICE V1.0 users's manual*. TNO Prevention and Health, Public Health.

van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P. E., Hall, A., Kosie, J., Kruse, E., Ritchie, S. J., Valentine, K. D., van't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *PsyArXiv*. https://doi.org/10.31234/osf.io/hvfmr

Van Dusen, B., & Nissen, J. (2020). Associations between learning assistants, passing introductory physics, and equity: A quantitative critical race theory investigation. *Physical Review Physics Education Research*, *16*. https://doi.org/10.1103/PhysRevPhysEducRes.16.010117

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., & Egger, M. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE Initiative): Explanation and elaboration. *PLoS Medicine*, *4*(10), e297. https://doi.org/10.1371/journal.pmed.0040297

von Hippel, P. T. (2009). 8. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

White, I. R., Daniel, R., & Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis*, *54*(10), 2267–2275. https://doi.org/10.1016/j.csda.2010.04.005

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Widaman, K. F. (2006). Iii. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, *71*(3), 42–64. https://doi.org/10.1111/j.1540-5834.2006.00404.x

Woods, A. D., Davis-Kean, P., Halvorson, M. A., King, K. M., Logan, J. A. R., Xu, M., Bainter, S., Brown, D. M. Y., Clay, J. M., Cruz, R. A., Elsherif, M. M., Gerasimova, D., Joyal-Desmarais, K., Moreau, D., Nissen, J., Schmidt, K., Uzdavines, A., & Vasilev, M. R. (2021). *Missing data and multiple imputation decision tree*. https://doi.org/10.31234/osf.io/mdw5r

Wu, W., & Jia, F. (2021). Applying planned missingness designs to longitudinal panel studies in developmental science: An overview. *New Directions for Child and Adolescent Development*, *2021*(175), 35–63. https://doi.org/10.1002/cad.20391

Zuberi, T. (2001). *Thicker than blood: How racial statistics lie*. University of Minnesota Press.

Zuberi, T., & Bonilla-Silva, E. (Eds.). (2008). *White logic, white methods: Racism and methodology*. Rowman & Littlefield Publishers.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2023). Best practices for addressing missing data through multiple imputation. *Infant and Child Development*, e2407. https://doi.org/10.1002/icd.2407