**QUEEN'S
UNIVERSITY
BELFAST**

Evaluating supplemental samples in longitudinal research: Replacement and refreshment approaches

Taylor, L. K., Tong, X., & Maxwell, S. E. (2019). Evaluating supplemental samples in longitudinal research: Replacement and refreshment approaches. *Multivariate Behavioral Research*.
<https://doi.org/10.1080/00273171.2019.1628694>

Published in:
Multivariate Behavioral Research

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
[Link to publication record in Queen's University Belfast Research Portal](#)

Publisher rights
© 2019 Taylor & Francis Group, LLC.
This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Running head: SUPPLEMENTAL SAMPLES IN LONGITUDINAL RESEARCH

Evaluating supplemental samples in longitudinal research: Replacement and refreshment
approaches

Laura K. Taylor^{1,2}, Xin Tong³ & Scott E. Maxwell⁴

¹University College Dublin, ²Queen's University Belfast, ³University of Virginia, ⁴University of
Notre Dame

Correspondence should be sent to Laura K. Taylor, School of Psychology, University College
Dublin, Ireland. Email: laura.taylor@ucd.ie

Special thanks to Ke-Hai Yuan for his input during the initial stages of developing this project
for the Quantitative Minor in Psychology and to the comments and feedback we received at the
2014 Society for Research on Child Development (SRCD) Developmental Methodologies
Conference.

Abstract

Despite the wide application of longitudinal studies, they are often plagued by missing data and attrition. The majority of methodological approaches focus on participant retention or modern missing data analysis procedures. This paper, however, takes a new approach by examining how researchers may *supplement* the sample with additional participants. First, *refreshment* samples use the same selection criteria as the initial study. Second, *replacement* samples identify auxiliary variables that may help explain patterns of missingness, and select new participants based on those characteristics. A simulation study compares these two strategies for a linear growth model with five measurement occasions. Overall, the results suggest that *refreshment* samples lead to less relative bias, greater relative efficiency, and more acceptable coverage rates than *replacement* samples or not supplementing the missing participants in any way.

Refreshment samples also have high statistical power. The comparative strengths of the refreshment approach are further illustrated through a real data example. These findings have implications for assessing change over time when researching at-risk samples with high levels of permanent attrition.

Key words: planned missing data, supplemental sample, replacement sample, refreshment sample, longitudinal design

Evaluating supplemental samples in longitudinal research: Replacement and refreshment approaches

Developmental researchers conducting longitudinal research often confront attrition, or participants dropping out of the study during the course of data collection (Enders, 2013; Jeličić, Phelps, & Lerner, 2009). Rates of permanent attrition, that is, when individuals drop out and do not return leading to monotonic missing data patterns, may be particularly pronounced in high-risk samples (Capaldi & Patterson, 1987). Methodologists have developed a host of modern missing data techniques, such as full information maximum likelihood (FIML), multiple imputation, and a two-stage estimation approach which may help to handle attrition (Enders, 2010; Savalei & Bentler, 2009). Moreover, applied researchers have documented a range of retention strategies that may be employed during longitudinal studies, such as sending biannual newsletters, reminders of data collection and increasing financial incentives over time (see Capaldi & Patterson, 1987; Cotter, Burke, Loeber, & Navratil, 2002). This paper takes a new approach of investigating the benefits and limitations of adding supplemental samples, or late entrants, to the original sample in longitudinal research. This treatment of supplemental samples fits within the broader literature on planned missing data designs (see special issue of *International Journal of Behavioral Development*, 38(5), 2014), or data collection designs that intentionally collect incomplete data from participants, reducing costs and participant burden, often with minimal implications for bias or power (Graham et al., 2006; Jorgensen et al., 2013; Little & Rhemtulla, 2013; Mistler & Enders, 2012). Thus, this study examines the consequences for estimating parameters of latent growth curves when additional participants are introduced into intermediate measurement occasions in longitudinal study designs.

When facing permanent attrition in longitudinal studies, researchers may take one of two approaches to supplement, or add late entrants to, the original sample (Little, 1993), often without informed decision making guided by methodologists. On the one hand, researchers may add *refreshment* samples, using the same selection criteria as the initial study. On the other hand, researchers may consider using *replacement* samples which first identify auxiliary variables that may help explain patterns of missingness, and select new participants based on those characteristics. Both of these supplemental approaches fall within planned missingness designs, and lead to two research questions. First, if conditions permit adding participants across a longitudinal design, should supplemental samples be added? Second, if so, which is the best supplemental sample to use when researchers are interested in assessing change over-time? That is, should researchers attempt to *refresh* the original sample by following the initial recruitment technique, such as a random selection of the population, or instead try to *replace* participants lost through attrition?

Following a brief review of missing data mechanisms, planned and unplanned missingness, and missing data patterns in longitudinal research, the paper presents the details of the proposed project, including a description of supplemental samples, as well as a distinction between two approaches to adding participants: refreshment and replacement samples. The sampling goals, logic, and procedures guiding the addition of supplemental samples, along with the assumptions about missing data mechanisms, will be discussed for both of these approaches. A simulation study will be presented to evaluate the effect of adding supplemental samples based on a linear growth curve model; these findings will be compared to other common practices employed by substantive researchers, such as listwise deletion and using maximum-likelihood-based methods to analyze all available data from the original sample. These two comparisons

were chosen, respectively, to reflect the ‘worse’ and ‘best’ current practices that may be commonly used by substantive researchers. Following the simulation study, analysis of the National Youth Study (NYS; Elliot, Huizinga, & Menard, 1989) demonstrates how these two approaches may function with a real data example. The paper concludes with a discussion of the practical implications for substantive researchers working with at-risk populations or high levels of permanent attrition.

Missing Data Mechanisms

Missing data theory was originally introduced and defined by Rubin (1976) who described three types of missing data mechanisms. First, missing completely at random (MCAR) defines any missingness as purely haphazard; that is, probability of a value being missing is not related to the observed data or the missing value itself. Second, for missing at random (MAR), the missingness is determined by the value on an observed variable in the dataset or model covariates, but is not related to the missing value itself. Finally, missing not at random (MNAR) describes the direct relation between missing data and the value on the outcome of interest; that is, whether a value is missing depends on the value itself, even after taking into account other variables in the dataset.

The match among these missing data mechanisms (i.e., why data are missing), the study design, and statistical analysis will determine the success of avoiding bias (Little, 1995). For example, traditional forms of handling missing data, such as listwise deletion, are the default in some mainstream statistical software, such as SPSS, and assume data are MCAR (see Enders, 2010). If these programs are used to analyze data that follow a different missing data mechanism (i.e., MAR or MNAR), researchers introduce systematic bias into their findings. Although modern missing data analytical techniques, such as FIML and multiple imputation, have been

developed (e.g., Baraldi & Enders, 2010; Graham, 2009; Little & Rubin, 2002; Newman, 2003), these approaches often assume data are MCAR or MAR and may lead to biased results when data are MNAR. With this constraint in mind, we now proceed to discuss planned missingness designs that often yield MCAR or MAR data.

Planned Missing Data in Longitudinal Designs

For longitudinal researchers in the social and behavioral sciences, the development of modern methods dealing with missing data has led to a number of planned missingness designs, with many practical advantages including easing participant burden and saving costs (Little & Rhemtulla, 2013). For example, participants can be randomly assigned to conditions which employ multiform designs (i.e., scales with missing items), two-method measurement designs (i.e., missing measures), and/or wave-missing designs (i.e., missing measurement occasions). Two types of planned wave-missing designs are relevant to understanding the approach of supplemental samples in longitudinal data designs. First, the cohort-sequential design (Baltes & Nesselroade, 1979) includes a number of cross-sections of the population, or cohorts, which are followed for a limited number of waves. The cohort-sequential design is of particular interest to developmental researchers who are interested in change over time and developmental trajectories. Second, the three-form design and extensions of this approach (Graham, Hofer, & MacKinnon, 1996; Graham et al., 2001), randomly assign participants into groups at the initiation of the study; each group is then measured only at distinct time points over the course of the data collection. These planned missingness designs often sacrifice little power compared to a complete-data design (Enders, 2010; for exceptions, see Schoemann, Miller, Pornprasertmanit, & Wu, 2014; Rhemtulla, Savalei, & Little, 2016); moreover, planned missing data are typically by design MCAR or MAR, eliminating the risk of bias in parameter estimates.

There is great promise in these types of planned missingness; yet, despite the advantages, these designs are still plagued by unplanned missingness (Widaman, 2006). For example, reducing participant burden within a time point, or in the number of measurement occasions, does not ensure retention over time. The threat of attrition is particularly pronounced in longitudinal research and with at-risk samples.

Unplanned Missing Data

This section will first present two factors that influence unplanned missing data – longitudinal designs and at-risk samples, followed by a discussion of permanent attrition. In longitudinal research, missing data can occur not only at the item-level, but also at the measurement occasion. In most instances, these types of missing data will be MAR or MNAR. In developmental research with normative samples, retention (i.e., the percentage of participants from the initial wave that complete the study) rates of 90 to 95% are considered ideal (Pierce & Hartford, 2004). That is, losing between 5 to 10% of participants over the course of the study is a ‘gold standard’ that researchers strive to meet using established best practices for retention in longitudinal studies (e.g., Cotter, Burke, Loeber, & Navratil, 2002; Hartsough, Babinski, & Lambert, 1996; Scott, 2004). However, even with normative samples, attrition can be influenced by sample characteristics, such as age group (e.g., high school students move out for college, older adults move to retirement communities, etc.), as well as study design (e.g., questionnaire length, time between measurement occasions, etc.). These factors may contribute to attrition of about 20% in studies ranging from youth and school-based programs to clinical settings (see Hall, 1993; Kellam, Rebok, Ialongo, & Mayner, 1994; Mason, 1999) and even up to 47% attrition over the course of a number of years (Capaldi & Patterson, 1987).

Compounding the challenges of retention with normative samples, at-risk samples may experience greater instability and have fewer means to maintain contact across time. For example, at-risk families may be more vulnerable economically; they may live in transitional housing or with relatives, need to follow work, or be less likely to leave forwarding addresses or keep the same phone number. In the school system, at-risk youth have higher truancy rates, may be more likely to be sick, or drop-out of the formal schooling system altogether. In international research, at-risk communities may be more likely to be affected by forced displacement due to internal conflict and violence. Across such at-risk samples, ranging from economically-disadvantaged U.S. samples to child soldiers in a post-war setting, retention rates during the course of longitudinal research range from 50 to 86% (e.g., Cummings et al., 2013; Betancourt et al., 2010; Browning, Burrington, Leventhal, & Brooks-Gunn, 2008; Kronenberg et al., 2010). Thus, over the years of data collection, these studies indicate that when working with longitudinal and at-risk populations, researchers may lose up to half of the original sample.

Finally, two patterns of attrition can be found in longitudinal research. Intermittent attrition describes the pattern in which individuals may miss one or more measurement occasions, but do not necessarily leave the study for good. On the other hand, permanent attrition occurs when participants drop out of the study and do not return, resulting in monotonic missing data patterns. Examples of this pattern of permanent attrition with at-risk populations may include severe illness or death with elderly populations, moving out of a domestic violence shelter, or entering a witness-protection program. These are extreme cases, however, they highlight patterns of permanent attrition in which existing strategies to increase retention would be ineffective. If researchers can identify factors or auxiliary variables that fully explain these

patterns of unplanned missingness, the data are MAR and can be analyzed using appropriate missing data analysis methods.

To date, the majority of the study of unplanned missingness has focused on the analysis stage, such as modern missing data analytical techniques (Jeličić et al., 2009; Nicholson & Deboek, 2017). However, there has been less guidance for researchers who might contemplate adding participants during the course of the study (for exceptions, see Deng, Hillygus, Reiter, Si, & Zeng, 2013; Graham & Donaldson, 1993; Si, Reiter, & Hillygus, 2015).

Supplemental Samples in Longitudinal Research

Supplemental samples, or late entrants (Little, 1993), are groups of participants who are intentionally included in the study after the first time point of measurement. For example, researchers may consider adding new sets of participants to diversify the sample (Conger & Conger, 2002), as additional funding is infused into the project (Cummings et al., 2013), or because conditions on the ground change to allow for additional recruitment (Betancourt et al., 2010; Lucas, 2005). Although supplemental samples are used in both U.S. and international settings, and despite the fact that reporting attrition has become common place, there is no agreed upon way to discuss late entrants in developmental psychology. This practice calls for greater attention not only to how/when/why participants leave a longitudinal study, but also to how/when/why participants are added in later waves.

To advance the understanding, the current study, informed by reporting practices in the study of population trends (Lucas, 2005; van Soest & Kapteyn, 2011), distinguishes between two types of supplemental samples. First, a *refreshment* approach follows the original sampling procedures, such as a random selection from the population. That is, researchers would not necessarily conduct prior analyses to determine predictors of missingness, but rather rely on the

original sampling frame and procedures to recruit late entrants. Second, a *replacement* approach is an attempt to over-sample from the profile of participants who have dropped out of the study. That is, researchers do exploratory attrition analyses after data collection to identify auxiliary variables (*A*) that are correlated to the missing data. These auxiliary variables would then be used to select new participants later in the study. In this manner, researchers are conceptually trying to *replace* those participants lost in attrition by over-selecting for those characteristics in later waves. Focusing on the higher-risk aspects of the sample that relate to attrition may be appealing as omitting elusive or less cooperative participants may bias results (Thornberry, Bjerregaard, & Miles, 1993). Moreover, “the practice of [replacing] the sample of a longitudinal study with new recruits who fit the profiles of those who drop out has much of the appeal of weighted and stratified repeated cross-sectional sampling without compromising the advantages of repeatedly sampling from participants who do not drop out” (Lee & Neimeier, 2005, p. 5). Thus, this study will explore both refreshment and replacement approaches to adding participants to on-going longitudinal research.

A number of policy reports and scientific studies have described the use of different types of supplemental samples. Investigating aging processes, the English Longitudinal Study of Ageing (ELSA) focused on people aged 50 and over (ELSA, 2015). The initial sample was drawn from households that had previously responded to the Health Survey for England (HSE), and at Wave 3 a refreshment sample was added from the HSE to include those who had now aged-in to the older cohort. Moreover, the research team noted that “further refreshment samples may also be added at future waves” (p. 1). On the other end of the life course, studying the impact of childhood poverty across four low-income countries over 15 years, the Young Lives Study recruited participants through households (Barnett et al., 2013). Even with low refusal

rates, less than 2%, the researchers note that “in such cases, replacement sampling was used” (p. 2). That is, when participants did not return, new participants were added to the sample.¹

Supplemental samples are also added to smaller scale developmental research projects. For example, an additional cohort of 107 grade-matched students from single-parent homes, approximately 25% of the original sample, was added at Time 3 of a five-year panel study with yearly assessments of 7th to 11th graders (Conger & Conger, 2002). Although this example differs from the definitions of refreshment and replacement approaches noted above – the late entrants were drawn from a new population (e.g., mother-headed households vs. two-parent families) and added to the original sample – it highlights the need for methodological rigor in describing how projects add participants during longitudinal research.

Supplemental samples can also occur at the group, not only the individual participant, level. For example, the High School and Beyond Longitudinal Study (HS&B) is a national longitudinal survey of U.S. high school students (D’Amico, 2011). Describing the sampling procedure, it was noted that “1,120 schools were selected in the original sample and 810 of these schools participated in the survey. An additional 200 schools were drawn in a replacement sample” (p. 1). These studies are some examples that explicitly discuss adding supplemental samples to longitudinal designs; however, this practice is not yet commonly or consistently documented in developmental psychological research.

Without clear methodological guidance on the implications of adding supplemental samples, researchers may struggle with selecting the appropriate approach to add late entrants. A brief thought experiment might highlight the difficulties in decision making that substantive

¹ According to our definition of the two types of supplemental samples, the Young Lives Study used a *refreshment* approach. That is, the field teams followed the original sampling procedures; for example, the Ethiopia country report states: “there were some refusals and changes of mind during the second visit and such households were replaced by another eligible one using the standard procedure” (Alemu et al., 2003, p. 21).

researchers face. Imagine that a research team interested in trajectories of aggression has collected a random sample of $N = 100$ adolescents (50% male) from a freshman class in a high-risk high school at Time 1. By sophomore year (Time 2), only 80 of those original adolescents are still at the high-school; the other 20 have either moved away, dropped out, or cannot be reached by school administrators. Conducting preliminary analyses on the two time points, the research team realizes that all 20 who have dropped out were male. They still have enough money in their budget to collect a full sample of 100 students, so as they plan for Time 3, knowing that changes in aggression through adolescence may vary for males and females (Boxer et al., 2013; Karrifer-Jaffee, Foshee, Ennett, & Suchindran, 2013; Maldonado-Molina, Jennings, & Komro, 2010), what would be recommended?

Possible scenarios observed across substantive research suggest possible responses could be to: (a) only follow the remaining part of the original sample ($N = 80$, 37.5% male), (b) randomly select an additional 20 students (10 female, 10 male) in the same grade (e.g., junior year) as the original cohort at Time 3 ($N = 100$, 40% male), or (c) use the auxiliary variable, gender, to oversample from the attrited group, male, to add the additional students at Time 3 and thus replace the 20 attrited males with 20 additional males instead of adding both females and males ($N = 100$, 50% male). Of course, this thought experiment could have included a continuous auxiliary variable, such as exposure to violence, which also relates to the outcome of interest and permanent attrition. The goal of this brief example, however, is merely to illustrate the difficult choices that researchers are faced with when trying to model developmental change, particularly with at-risk samples. Consequently, the purpose of our study is to provide substantive researchers with guidelines in the application of supplemental samples.

Monte Carlo Simulation Study

A simulation study is conducted to compare the impact of the two types of supplemental approaches, refreshment and replacement samples, in analyzing MCAR and MAR data in longitudinal research. R code for simulation and analyses available at lauraktaylor.wordpress.com.

Data Generation

We first generate complete data using software R (R Development Core Team, 2011) from a structural linear growth model in which five time points are equally spaced for each participant (see Figure 1 for the path diagram of the data generating model). In the model, $\mathbf{e}_i = (e_{i1}, e_{i2}, e_{i3}, e_{i4}, e_{i5})'$ represents intraindividual measurement errors for the i th individual, with mean $\mu_e = 0$ and covariance matrix $\Sigma_e = I\sigma_e^2$. The residual vector $\mathbf{u}_i = (u_{Li}, u_{Si})'$ represents the random component of the latent coefficients $(L_i, S_i)'$, with mean 0 and covariance matrix

$$D = \begin{pmatrix} \sigma_L^2 & \sigma_{LS} \\ \sigma_{LS} & \sigma_S^2 \end{pmatrix},$$

where L represents the latent intercept and S represents the latent slope. Both $\mathbf{e}_i$ and $\mathbf{u}_i$ are generated from multivariate normal distributions with corresponding means and covariance matrices. Given $\mathbf{e}_i$, $\mathbf{u}_i$ and true values of model parameters, we generate outcome variables $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5})'$ by

$$\mathbf{y}_i = \Lambda \mathbf{b}_i + \mathbf{e}_i,$$

$$\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{u}_i,$$

where $\Lambda = ((1,1,1,1,1)', (0,1,2,3,4)')$ is the factor loading matrix which can represent the growth trajectories, $\mathbf{b}_i = (L_i, S_i)'$ is a vector of the random intercept and the random slope, and $\boldsymbol{\beta} = (\beta_L, \beta_S)'$ is the mean of $\mathbf{b}_i$ and is called fixed effects.

INSERT FIGURE 1 HERE

The original sets of complete data are retained for comparison purposes. In other words, we report not only results for datasets affected by attrition, but we also report results that would occur for a complete dataset where no attrition has occurred. We do so not because it is realistic to assume a complete absence of attrition, but instead to provide a baseline against which to evaluate other methods. We also retain a dataset with listwise deletion; this allows for a comparison of performance when adding supplemental samples relative to shifting from listwise to a modern missing data approach. We propose that this is a relevant comparison in terms of appreciating the magnitude of impact of adding a supplemental sample.

In all datasets, once a participant has a missing value at Time t , s/he has a missing value at Time $t+1$ (i.e., monotonic missingness). Table 1 summarizes the missing proportion ($r\%$) and retention percentages from the original sample over the course of five time points for the proportion of missingness in each condition.

INSERT TABLE 1 HERE

Based on the complete data, we generate MCAR and MAR data. For both MCAR and MAR, missing values are simulated based on the missing proportion ($r\%$) at each time point. That is, for MCAR, of the observed data at Time 1, $r\%$ are randomly selected to have missing values at Time 2; of the observed data at Time 2, $r\%$ are randomly selected to be omitted at Time 3, and so on. For MAR, between each time point, $r\%$ are selected to be missing based on the value of an auxiliary variable. The auxiliary variable (A) is created to be related to the slope parameter S . Specifically, A is generated as $A = aS + \epsilon$. Let $var(\epsilon) = var(S) = 1$.

$$cor(A, S) = \frac{cov(aS + \epsilon, S)}{\sqrt{var(aS + \epsilon)var(S)}} = \frac{a \cdot cov(S, S)}{\sqrt{(a^2 + 1)var(S)var(S)}} = \frac{a}{\sqrt{a^2 + 1}} \quad (1)$$

So, $a = \frac{cor(A, S)}{\sqrt{1 - cor(A, S)^2}}$. The correlation between A and the slope parameter S varies across conditions, but a is calculated based on Equation 1 when generating the auxiliary variable

following $A \sim N(a\beta_S, a^2 + 1)$, where β_S is the mean of S . We selected the r th percentile of this distribution as the cutoff value to generate missing values.

Supplemental samples, both replacement and refreshment, are generated separately and independently, and added at Time 3. The refreshment sample is created following the same procedures as the original sample. The replacement sample is determined by selecting participants that surpassed a threshold on the auxiliary variable. Practically, this entailed generating complete data for one individual including calculating $A = aS + \epsilon$. If the value of A for this individual was larger than the mean of A , or $a\beta_S$, the individual is saved into the replacement sample. If the value of A is less than $a\beta_S$, the data for this individual is dropped.² These steps are repeated until the replacement sample reaches the specified supplemental sample in each condition. The same missing data mechanism and proportion of the original sample is applied to the remaining time points for both the refreshment and replacement samples.

Conditions Investigated in the Simulation Study

The simulation design varies the real missing data mechanism (MCAR vs. MAR) with permanent (i.e., monotonic) attrition. We consider the possible steps researchers might take after the second wave of data collection. For the MCAR condition, even if the researchers conduct preliminary analyses, there would not be a relevant auxiliary variable that would predict later attrition;³ therefore, for MCAR, *only* the refreshment approach is used to supplement the original sample. However, in the MAR case we investigate two possible scenarios. In the first, the researchers proceed at Time 3 with the refreshment approach, regardless of whether they

² For analyses, the data from Time 1 and 2 were deleted for all individuals in the replacement and refreshment samples. In other words, we generated scores at all five time points for these individuals, but they are not included in the analyses until Time 3 because they are in the supplemental, not the original, sample.

³ Acknowledging the role of multiplicity, we recognize that even with MCAR, if 20 potential auxiliary variables were examined, one may relate significantly to attrition using $\alpha = .05$. Even so, if missingness is MCAR, any conceivable auxiliary variable would necessarily have a zero population correlation with missingness.

conducted attrition analyses between Time 1 and Time 2. In the second scenario for MAR, we suppose that the attrition analyses reveal an auxiliary variable collected at Time 1 that predicts missingness at Time 2. With this knowledge, the replacement approach is used to over-sample new recruits with relevant levels on the auxiliary variable. That is, three types of datasets are generated: MCAR with refreshment, MAR with refreshment, and MAR with replacement.

Across each type of dataset (MCAR with refreshment, MAR with refreshment, and MAR with replacement), four primary conditions⁴ are manipulated: (1) sample size ($N = 50, 100, 300, 500, 1000$), (2) missing proportion between each time point ($r = 3\%, 5\%, 8\%, 15\%$), (3) correlation between the auxiliary variable and latent slope ($\text{cor}(A,S) = .3, .5, .8$), and (4) size of refreshment(RF)/replacement(RP) sample at Time 3 ($n=1x, 2x, 3x$ the number of missing participants from Time 1 to Time 2); that is, RF(1) would be a refreshment sample 1 times the number of missing participants at Time 2, while RP(3) would be a replacement sample 3 times the number of missing participants at Time 2. For each condition, 500 replications are simulated.

These conditions and levels are selected to represent a range of options that may be most useful to practical researchers. For example, (1) sample sizes are realistic of short-term diary studies, developmental psychological research, and large-scale longitudinal studies (Aldridge & Roesch, 2008; Nuttall, Valentino, Borkowski, 2012; Cummings et al., 2013); (2) a missing proportion of .03 would yield an approximate 90% retention over the five time points, thus approaching the gold standard for normative samples, and .15 missing proportion represents approximately 50% retention over the five time points, consistent with the lower range in

⁴ Two additional conditions were originally included: (5) correlation between the latent intercept and slope ($\text{cor}(L,S) = 0, .3$) and (6) variance of measurement errors ($\text{var}(e) = 1, 3$). The combination of these levels and conditions were chosen to make the reliabilities of the manifest variables between 0.25 and 0.95. Varying the correlation between the latent intercept and slope or the measurement error variance did not affect the pattern of findings; researchers may consider omitting these conditions from future simulations.

existing longitudinal research with high-risk samples; (3) correlations between the auxiliary variable and latent slope ranging from weak (.3), to medium (.5), to strong (.8); and (4) levels for supplementary sample that correspond to number of original participants missing at Time 2 ($n=x$), given which, researchers could (a) choose to supplement as many as were missing ($n=1x$), or anticipating similar losses across the remaining three time points, could (b) choose to supplement three times as many ($n=3x$).

Data Analysis and Evaluation Criteria

A two-stage maximum likelihood estimation approach (Savalei & Bentler, 2009) is used to analyze all of the models, except the listwise deletion condition. The two-stage approach was selected because it allows for incorporation of auxiliary variables, has good coverage, and in smaller samples it outperforms FIML for both MCAR and MAR data (Savalei & Bentler, 2009; see also Savalei & Falk, 2014). Practically, the two-stage maximum likelihood has been incorporated in an R package *rsem* (Yuan & Zhang, 2012) so that applied researchers may easily use it.⁵

The impact of adding refreshment and replacement samples across conditions is evaluated using multiple criteria, including relative bias, relative efficiency, coverage rate (CR) of the 95% confidence interval, and statistical power or Type I error rate. Let θ denote a parameter and also its population value in the simulation, and $\hat{\theta}_r, r = 1, \dots, 500$ denote its estimates from the r th simulation replication. Parameter estimate $\hat{\theta}$ is calculated as the average of parameter estimates of 500 simulation replication

⁵ If the auxiliary variable is incorporated in the model (e.g., score at Time 1), then FIML can be used directly. Moreover, for large samples, results from FIML and the two-stage method should yield the same parameter estimates.

$$\hat{\theta} = \frac{1}{500} \sum_{r=1}^{500} \hat{\theta}_r.$$

Relative bias is calculated as

$$Relative\ bias = \begin{cases} \frac{\hat{\theta} - \theta}{\theta} & \theta \neq 0 \\ \hat{\theta} & \theta = 0 \end{cases}.$$

Relative efficiency is the ratio of the squared parameter empirical standard error of complete data to incomplete data

$$Relative\ efficiency = \frac{ESE_{\hat{\theta}, complete}^2}{ESE_{\hat{\theta}, missing}^2}.$$

A value of 1 indicates that the parameter estimates are as efficient as those for the complete data.

Coverage rate (CR) of the 95% confidence interval is calculated as

$$CR = \frac{\#(\hat{l}_r < \theta < \hat{u}_r)}{500},$$

where $\#(\hat{l}_r < \theta < \hat{u}_r)$ is the total number of replications with confidence intervals covering the true parameter θ . Good 95% confidence intervals should give coverage probabilities close to 0.95. Statistical power or Type I error is calculated by

$$power/Type\ I\ error = \frac{\#(\hat{l}_r > 0) + \#(\hat{u}_r < 0)}{500},$$

where $\#(\hat{l}_r > 0)$ is the total number of replications with the lower limits of confidence intervals larger than 0, and $\#(\hat{u}_r < 0)$ is the total number of replications with the upper limits of confidence intervals smaller than 0. If $\theta = 0$, it is Type I error rate. Otherwise, it is statistical power.

Results

Analysis of the three types of dataset (MCAR with refreshment, MAR with refreshment, and MAR with replacement) permits comparisons between the two missing data mechanisms, MCAR and MAR, as well as between the refreshment and replacement approach for MAR. To demonstrate the range of impact of the supplemental samples, the three types of missing dataset will also be compared to the complete dataset and the listwise deletion dataset.

Given developmental psychologists' interest in change over time, here we describe the results for average latent slope parameter (β_S) in this article. Results on other parameters ($\beta_L, \sigma_L^2, \sigma_S^2, \sigma_{LS}$) show the same patterns when adding different types of supplemental samples. To further simplify the presentation of findings, for both MAR and MCAR we include figures for $\text{var}(\epsilon) = 3$ and for the MAR condition we include the figures for $\text{cor}(A, S) = .8$. Again, the pattern of results remains the same across the other levels of these factors. Results and figures for all parameters, factors, and levels are available as supplementary files at lauraktaylor.wordpress.com.

Refreshment approach. As noted above, the refreshment approach to supplemental samples was used with both MCAR and MAR datasets, respectively. In the case of MCAR, a number of patterns in the dependent variables emerge. First, particularly for the larger samples ($N > 200$), the relative bias remains low across conditions, regardless of whether a refreshment sample has been added or not (Figure 2a). Even at the higher end of attrition, the refreshment samples performed roughly equivalent to the complete data analyses. That is, because of the missing data mechanism and the two-stage estimation method, there is little bias in the estimate of the average slope parameter.

INSERT FIGURE 2A HERE

Second, for relative efficiency, the MCAR with refreshment samples perform better than the listwise deletion method for the original dataset with no supplemental samples and better than the MCAR with no refreshment (Figure 2b). Moreover, as the size of the refreshment sample increases, the relative efficiency improves and approaches performance of the complete dataset.

INSERT FIGURE 2B HERE

Third, for the coverage rate, or the proportion of replications that captured the population value of the slope, Bradley (1978) established guidelines for researchers to evaluate robustness of the Type I error rate. He recommended that any empirical Type I error rate no higher than .075 should be considered as robust relative to a nominal Type I error rate of .05. This would correspond to a coverage rate of .925. The MCAR with refreshment samples had acceptable coverage rates (Figure 2c), except for a limited set of cases which fell below the acceptable cut-off value (see supplemental files: $N=50$, $\text{cor}(A,S) = 0$, MCAR-RF(1) when $r=5\%$ and 15% , and MCAR-RF(2) and MCAR-RF(3) when $r=5\%$, 8% and 15%).

INSERT FIGURE 2C HERE

Fourth, because of the larger sample size after supplemental samples are added, statistical power is also improved (Figure 2d). Across these conditions, the statistical power to detect the latent slope parameter improves as the number of late entrants increases, particularly for the higher rates of attrition. For example, for the mid-range sample ($N=500$) when the latent intercept and slope are not correlated and the missing proportion is high ($r=15\%$), statistical power increases from approximately .70 to .90 comparing MCAR with no refreshment to MCAR-RF(3).

INSERT FIGURE 2D HERE

For MAR with refreshment, the overall pattern of findings for the relative bias, relative efficiency, coverage rate, and statistical power is largely the same as for the MCAR condition. First, as with MCAR, relative bias remains low with or without a refreshment sample. That is, MAR with refreshment samples performed roughly equivalent in regard to relative bias to the complete data analyses, even at the higher end of attrition (Figure 3a). Second, as the size of the refreshment sample increases, so does the relative efficiency for the MAR; that is, the MAR with refreshment also outperform the MAR with no refreshment samples (Figure 3b). Third, the coverage rate was above Bradley's cut-off criteria of .925 in all conditions of MAR with refreshment (Figure 3c). Fourth, considering the low relative bias and acceptable coverage rates, the improvements in statistical power for MAR with refreshment follow the pattern in the MCAR condition (Figure 3d); that is, compared to MAR with no refreshment, the gains in statistical power were more pronounced for the larger refreshment samples. Overall, adding a refreshment sample to MCAR and MAR datasets results in low relative bias, improved relative efficiency, good coverage rates, and improvements in statistical power.

INSERT FIGURE 3A HERE

INSERT FIGURE 3B HERE

INSERT FIGURE 3C HERE

INSERT FIGURE 3D HERE

Replacement approach. Next, the condition was investigated assuming that the researchers conducted attrition analyses after Time 2 which accurately identified an auxiliary variable that was related to missingness. The auxiliary variable was then used to guide the selection of the replacement sample in attempt to 'match' the profile of the participants who had

dropped out of the study. In this case of MAR with replacement samples, a different pattern of findings was found in the simulation.

First, for the MAR with replacement condition, the relative bias increased as the number of late entrants added at Time 3 increased (Figure 3a). This trend was more pronounced when there was greater attrition. For example, when there is no correlation between the latent intercept and slope with a missing proportion of $r=15\%$ for a sample of $N=500$, the relative bias increased as the replacement samples increased, from approximately $\text{MAR-RP}(1) = .8$ to $\text{MAR-RP}(2) = 1.0$ to $\text{MAR-RP}(3) = 1.5$. These findings are not surprising, however, because adding the replacement sample, which was higher on the auxiliary variable, changes the distribution of the original sample and is therefore less accurate than the refreshment approach.

Second, for MAR with replacement, the simulation suggests that as more late entrants are added, the relative efficiency decreases (Figure 3b). For example, the $\text{MAR-RP}(3)$ is less efficient across the various levels of the factors compared to the $\text{MAR-RP}(1)$.

Third, the coverage rates for MAR with replacement largely fell well *below* Bradley's criterion (Figure 3c). As the size of the replacement sample at Time 3 increased, the coverage rates worsened. For example, when there was no correlation between the latent intercept and slope and the missing proportion was high ($r=15\%$), in the $\text{MAR-RP}(3)$ for an original sample size of $N=500$, the coverage rate dropped to approximately .25. This value can be interpreted that there is only a 25% chance that the correct value was contained in what is nominally a 95% confidence interval. That is, for MAR, as more late entrants were added following the replacement strategy, there was a decrease in the proportion of cases that captured the true value of the mean slope parameter in the 95% confidence interval. When the coverage rate is lower than Bradley's cut-off criterion, results should be interpreted with caution.

Finally, considering greater bias and lower coverage rates, interpretation of statistical power is called into question for the MAR with replacement datasets (Figure 3d). That is, statistical power should only be interpreted when there is little bias with acceptable coverage rates. In this condition, although statistical power is higher in the MAR with replacement datasets compared to the MAR with refreshment datasets, the center of the interval is misplaced. What appears to be great statistical power is not meaningful and may lead to uninterpretable results because the estimates are so biased for the MAR with replacement conditions.

An Analytical Interpretation of the Simulation Results

We expand on these results with an analytic interpretation. Let $\mathbf{y}$ represent the population of the outcome variable from Time 1 to Time T . The manifest variables $\mathbf{y}_i, i = 1, \dots, N$, constitute a sample from $\mathbf{y}$ and may contain missing values. We use the two-stage procedure to estimate the growth curve model because it works better for small samples and is convenient to incorporate auxiliary variables (Savalei & Bentler, 2009). In the first stage, we obtain the saturated maximum likelihood estimate of the population mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{y}$ by maximizing the log-likelihood, $l(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}_1, \dots, \mathbf{y}_N, \mathbf{A})$, given observed outcome variables and potential auxiliary variables. In the second stage, we fit the growth curve model to the estimated population mean and covariance matrix $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ under the structural equation modeling framework. Let $\mu(\boldsymbol{\theta})$ and $\Sigma(\boldsymbol{\theta})$ be the structural model satisfying $\boldsymbol{\mu} = \mu(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma} = \Sigma(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents all the free parameters in the model. The estimated $\hat{\boldsymbol{\theta}}$ are obtained by minimizing

$$F_{ML}(\boldsymbol{\theta}) = [\hat{\boldsymbol{\mu}} - \mu(\boldsymbol{\theta})]' \Sigma^{-1}(\boldsymbol{\theta}) [\hat{\boldsymbol{\mu}} - \mu(\boldsymbol{\theta})] + tr[\hat{\boldsymbol{\Sigma}} \Sigma^{-1}(\boldsymbol{\theta})] - \log|\hat{\boldsymbol{\Sigma}} \Sigma^{-1}(\boldsymbol{\theta})| - T.$$

The above procedure will result in consistent parameter estimates if the sample $\mathbf{y}_i, i = 1, \dots, N$, is *randomly* selected from the population $\mathbf{y}$. When there are no supplemental samples

and when refreshment samples are used, the random sampling assumption is satisfied. However, replacement samples are not selected based on the random sampling procedure. In fact, they follow truncated distributions of the population distribution because they are selected to replace the missing data which corresponds to certain values on the auxiliary variable. For replacement samples, the data distribution becomes a mixture of the original population distribution and the truncated one, based on the mixing proportions of the original sample and the replacement sample, respectively. As a result, the estimated μ and Σ from the first stage of the estimation method are biased from the true mean and covariance matrix of the original population. Consequently, model parameters θ are also incorrectly estimated in the second stage. As the size of the replacement sample increases, the mixing proportion of the truncated distribution increases too. Thus, the data distribution deviates more from the original population distribution, leading to more biased parameter estimates. Our simulation results are consistent with these theoretical interpretations.

Although we apply this interpretation to the two-stage maximum likelihood method, the same pattern would pertain to other estimation methods because adding replacement samples changes the data distribution. Without any correction, it is impossible to obtain unbiased parameter estimates. Our simulation study shows the degree of distortion from adding replacement samples in empirical data.

Real Data Example

This section presents a real data example of adding refreshment and replacement samples to a longitudinal study; R code and data are available at lauraktaylor.wordpress.com. The dataset is Cohort 1 (age 11, N = 239) of the NYS which is a representative longitudinal dataset of US youth from 1976 to 1980, available for download at

<https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/88>. This cohort was selected because the total sample size was in the middle of the range of values in our simulation.

The outcome of interest was youth attitudes in support of social deviance as there was linear change across the five annual waves of data (Figure 4; Raudenbush & Chan, 1993). Attitudes toward deviance were assessed with a Likert-type scale for 9 items, such as *cheat on school tests* and *hit or threaten someone without any reason*, with responses ranging from 1 “very wrong” to 4 “not wrong at all.” Higher scores indicated more support for deviant behaviors among peers and the average internal consistency was good ($\alpha = .84$). The log transformation of scores was used due to the positively skewed responses.

INSERT FIGURE 4 HERE

For the current analyses, a random subsample of 120 participants (47% male, 53% female) were selected as the original sample; this subsample approach allowed for the inclusion of supplemental samples of different sizes at Time 3. This random subsample participated in five measurements. The mean age for the random subsample across the five measurement occasions was 11.53 ($SD = 2.65$), 11.91 ($SD = 2.60$), 13.09 ($SD = 3.69$), 14.25 ($SD = 4.23$), and 14.68 ($SD = 4.43$) years old, respectively.

In this subsample of the real dataset, at Time 2, 10 participants, 3 boys and 7 girls, had missing values. The subsample dataset followed a pattern of permanent attrition and had a missing data rate at each wave of 0%, 8.33%, 8.33%, 12.5%, and 12.5%. That is, in the random subsample of 120 participants at Time 1, the actual number of missing values was 10, 10, 15, and 15 at Times 2-5, respectively.

Following the procedures for refreshment and replacement samples, a supplemental sample was added at Time 3. The refreshment and replacement samples were selected from the

remaining participants in the cohort not included in the original subsample. For the refreshment strategy, based on the attrition of $n=10$ students from Time 1 to Time 2, a random selection of 10 ($n=1x$), 20 ($n=2x$) and 30 ($n=3x$) adolescents were selected from the original cohort, respectively. For the replacement strategy, mimicking what researchers would do after Time 2 data collection, attrition analyses identified gender as a relevant auxiliary variable for missingness. That is, girls were more likely than boys to drop out from Time 1 to Time 2.⁶ Therefore, having identified gender as an auxiliary variable, for the MAR replacement condition, there was an over selection of girls for the Time 3 replacement sample. Following the different levels, the selection from the remaining participants for the replacement samples were: 3 boys and 7 girls ($n=1x$), 6 boys and 14 girls ($n=2x$), and 9 boys and 21 girls ($n=3x$), respectively.

The linear slope was estimated across the various strategies (Table 2), with gender included as an auxiliary variable in analyses. For the random subsample ($N = 120$), the average estimated linear slope was $\beta = .85, p < .05$ (Figure 4).⁷ Overall, the refreshment strategy estimated a lower slope ($\beta = .81$) while the replacement strategy estimated a higher slope ($\beta = .88$). These estimates may be similar to the original subsample because of the relatively small rate of attrition in this real data example. Given the difference between the refreshment and replacement samples, based on the simulation results, we trust the results from adding x_3 refreshment sample. Moreover, following the steps that researchers would take with real data, gender was identified as an auxiliary variable that was related to the missingness mechanism at Time 2. However, given that the results from refreshment and replacement samples are similar,

⁶ In the original subsample, the bi-serial correlation between gender and the estimated latent slope is .17 which is at the lower end of the conditions in our simulation. There were no other variables in the dataset that were identified as auxiliary variables, i.e., that were correlated with whether or not a participant returned at Time 2.

⁷ Figure 4 depicts the plots of individual growth trajectories of the original sample, and the estimate of the average slope parameter. It should be noted that for individuals who participated in only one time point, i.e., original sample that attrited after Time 1 or supplemental sample that attrited after Time 3, individual growth trajectories cannot be computed.

we can conclude that gender was not a real auxiliary variable and the missingness mechanism is in fact MCAR or MNAR.

INSERT TABLE 2 HERE

Discussion

Adding to the existing literature that outlines best practices for participant retention and analytical strategies that can accurately estimate parameters despite missing data, the current study defined two supplemental strategies, refreshment and replacement. A simulation study demonstrated the implications of using these approaches, specifically in regard to the accuracy and precision in model estimation and increased statistical power. A real data example further highlighted the practicality of conducting attrition analyses and responding with supplemental samples.

Refreshment Approach

When attrition analyses suggest that the missing data mechanism is plausibly MCAR, a refreshment sample was used. All analyses for the MCAR without and with refreshment used the two-stage maximum likelihood approach, except for the listwise deletion dataset, which yielded low relative bias. For all conditions of MCAR with refreshment samples, the bias was low and the coverage rate was above Bradley's criterion. Compared to the no refreshment approach that also used two-stage estimation, the advantages of adding a refreshment sample were also observed with regard to the relative efficiency, coverage rate, and statistical power because of the enhanced sample size.

The simulation study also revealed advantages of adding a refreshment sample even when the missing data mechanisms is known to be MAR, or in the practical case when attrition analyses identified a relevant auxiliary variable that related to missingness, indicating that data

were not MCAR. As with MCAR, apart from the listwise deletion dataset, MAR analyses used the two-stage maximum likelihood approach. For MAR data with and without refreshment samples, there was low bias. Moreover, the MAR with refreshment had appropriate coverage rates, allowing researchers to benefit from the increased precision associated with the relative efficiency and higher statistical power. Together, these findings suggest that the refreshment approach was the preferred mechanism for both MCAR and MAR cases.

Refreshment samples may have other benefits for longitudinal and intervention research. For example, previous research has used refreshment samples to identify potential bias due to selective attrition (Frick, Goebel, Schechtman, Wagner, & Yitzhaki, 2006; Kruse et al., 2009). Rather than retaining new participants in subsequent waves, these cross-sectional refreshment samples may also be used to correct biases related to attrition (Deng et al., 2013; Si et al., 2015). Toward evaluating the impact of an intervention program, a simulation study gained insight into the missing data mechanism by including a refreshment sample, or “a random sample of subjects with initially missing data” (Graham & Donaldson, 1993, p. 119). Building on Little and Rubin’s early work (1987), differential attrition across experimental conditions can be estimated without bias using an expectation-maximization (EM) algorithm, when the attrition mechanism is accessible through comparisons with late entrants (Hirano, Imbens, Ridder, & Rubin, 2001). Thus, adding a refreshment sample may not only yield benefits for estimation of latent growth curve model parameters, but also provide better insight into the missing data mechanism itself. In the current paper, we expand upon the range of supplemental samples that researchers could utilize, empirically comparing the impact of both refreshment and replacement approaches.

Replacement Approach

Given the conceptual appeal of replacing participants that permanently drop out of a study (Lee & Neimeier, 2005; Thornberry et al., 1993), researchers may consider over-selecting new recruits that have a similar profile with regard to the auxiliary variable that relates to missingness. However, the current simulation study showed that the replacement approach can produce very misleading results, even if an accurate auxiliary variable is identified. That is, by over-selecting participants that represent the risk factors related to attrition, and not taking potential sampling bias into account, the replacement approach actually introduces more bias into the model estimation; in essence, this approach changes the underlying population. Therefore, it is not surprising that the bias was more pronounced with lower relative efficiency as the size of the replacement sample increased. Relatedly, the coverage rate was poorest for the MAR with replacement across conditions, compared to the refreshment approach for both MAR and MCAR. Given the relative bias, relative efficiency, and coverage rates, any apparent advantage of higher statistical power was offset; that is, the confidence interval around the parameter estimate was less likely to include the true mean if the MAR replacement approach was used.

These empirical findings are echoed by two additional notes of caution. First, due to sampling variability, it is possible that researchers may incorrectly identify an auxiliary variable. For example, if a number of potential auxiliary variables were tested, by chance using $\alpha = .05$, one in 20 would result in a significant finding. In smaller samples, testing a greater number of potential auxiliary variables, researchers could identify a spurious auxiliary variable and then over-sample from that group. This potential introduction of bias applies to both MCAR and MAR cases, and may be an additional motivation to avoid replacement samples. Moreover, including too many auxiliary variables may create problems if they are collinear with themselves

or with the substantive variables (Yuan & Lu, 2008). Moreover, given that there is no mathematical test for whether data are MAR or MNAR, researchers can only try to find an auxiliary variable which is correlated with whether the outcome observations are missing or not (Enders, 2010; Pearl, 2018). In practice, researchers commonly assume that they know why data are missing and treat data as MAR, but there may be additional uncertainty around choosing a replacement sample that mimics the cause of missingness. Applying supplemental samples to cases of MNAR in simulated data may be a promising area for future research.

Second, the current study used a replacement approach without adjusting for sampling weights. Because the replacement approach intentionally over-samples based on the identified auxiliary variable, a weighted adjustment ought to be taken into consideration. Challenges to using sampling weights, however, include a trade-off between bias reduction and increase in variance of the weight (Höfler, Pfister, Lieb & Wittchen, 2005) and the need for precise and stable estimates about the probability of the auxiliary variable(s) in the population; such estimates may be difficult to generate for smaller or unique samples. In addition, sampling weights need to be revised over time (Gouskova, Heeringa, McGonagle, & Schoeni, 2008) or customized to a particular research question (NYLS, 1997). As Cortes, Mohri, Riley and Rostamizadeh (2008) stated, “several techniques have been commonly used to estimate the reweighting quantities. But, these estimate weights are not guaranteed to be exact” (p. 41). Therefore, despite the potential promise of sampling weights, there appears to be no consistent solution to the threat of selection bias (Stolzenberg & Relles, 1997). Based on our pilot study, adding sampling weights that are commonly used in practice (e.g., Höfler et al, 2005, p. 293; Mazen & Tong, 2018) does not reduce bias caused by the replacement samples. Other bias

correction techniques, such as parametric bootstrap, non-parametric bootstrap methods, and inverse probability as weights, need to be studied further for longitudinal designs.

Limitations

There are a number of ways that the limitations of the current study could be addressed in future research. First, all data were generated following the normal distribution; however, in social and behavioral sciences, non-normal distributions may be more common. Although maximum likelihood (ML) estimation has been incorporated in all statistical software, the robustness of ML is threatened when missing data are combined with non-normal data (Savalei, 2010) and may vary across the type of non-normal distribution (Yuan, Yang-Wallentin, & Bentler, 2012). More specifically, simulation studies have shown that the robustness of the ML approach is violated if the proportion of missing data increases to greater than 10% with non-normal data (Savalei, 2010; Savalei & Falk, 2014). Therefore, future research should investigate if the current findings hold across non-normal distributions (Yuan, 2009), and may also consider dichotomous or ordinal auxiliary variables in addition to continuous data.

Second, for practical research, it may be difficult to predict when conditions on the ground may change, or when additional funding may become available, which can facilitate adding a supplemental sample. Therefore, it would also be useful for future research to investigate the implications for adding supplemental samples at multiple measurement occasions, rather than just at a single time point (i.e., half-way through data collection) as modeled in the current study. In this direction, it would be useful to compare how these forms of supplemental samples may approximate variants of the three-form design.

Third, beyond the scope of the current study, participant reactivity is another issue that is relevant to longitudinal research with at-risk populations or covering topics that are sensitive or

personal in nature. For example, later points of assessment may be affected by participants' responses to earlier time points; that is, "the act of being surveyed can affect behavior and confound estimates of parameters that initially motivated the data collection" (Zwane et al., 2011, p. 1821). Exponential decay functions offer a possible modelling method to account for the nonlinearity caused by measurement reactivity (Fritz, 2014). In addition, as one reviewer astutely noted, participants "settle down" and provide more consistent and presumably more accurate responses over time. Concerns for this threat to validity would equally apply across both supplemental approaches, however. Future research might study the robustness of the current findings regarding supplemental samples to model nonlinear change or when the participant reactivity issue exists.

Finally, also beyond the scope of the current study, additional bias correction methods should be considered with the use of replacement samples. A number of simulation and empirical studies have investigated the use of sampling weights (e.g., Deng et al., 2013; Hirano et al., 2001; Langkamp, Lehman, & Lemeshow, 2010; Magee, Robb, & Burbridge, 1998); however, our pilot study showed that the commonly applied sampling weights did not work for the replacement samples.⁸ Although different techniques may be applied to correct the sample selection bias (Stolzenberg & Relles, 1997), as far as we are aware, no method offers a general solution. Thus, it is necessary to systematically compare different approaches to identify the best method to correct the bias caused by replacement samples.

Conclusion

⁸ Little and Rubin (2002) discussed design weights which are inversely proportional to the probability of the individual being in the dataset. That is, the weight for each individual is denoted as $1/\pi_i$, where π_i is the probability that the individual is included in the sample. In a related study, we used inverse probability weighting method; however, the bias was not corrected by applying the weights to the data.

Conceptually, this study offers a new approach to planned missing data designs in longitudinal research. Supplemental samples can extend and complement the three-form design (Graham et al., 1996; Graham et al., 2001), cohort-sequential design (Baltes & Nesselroade, 1979), and other planned wave-missing designs (Rhemtulla, et al., 2014). The simulation confirms that adding either a refreshment or a replacement sample leads to different results; therefore, in practice, we suggest researchers follow the guidance offered by the simulation study. Compared to no supplemental approach, adding a *refreshment* sample to the original participants in longitudinal research can improve the accuracy and precision of parameter estimates for data that are either MCAR or MAR. Even in the case of MAR with a highly correlated auxiliary variable, the refreshment approach was advantageous over a replacement approach, particularly for smaller samples. Extending previous research (Deng et al., 2013; Graham & Donaldson, 1993; Si et al., 2015), the current study suggests that researchers should seek additional funding to strategically incorporate refreshment samples over time.

Practically, this study highlights the need for researchers working with longitudinal samples to more accurately document and describe how and when late entrants are added. In a similar way that attrition and retention have received increasing attention in statistical reporting in developmental psychology journals (Jeličić et al., 2009), reviewers should be encouraged to evaluate studies based on their acknowledgement and treatment of supplemental samples. Currently there are no standardized expectations or norms for discussing late entrants in the social and behavioral sciences; to varying degrees of clarity and precision, research teams are left to describe such procedures on an ad-hoc basis. This paper demonstrates the need for clearly defined terminology and consistent reporting mechanisms, perhaps adapted from the study of population trends (Lucas, 2005; van Soest & Kapteyn, 2011), around supplemental samples.

Without clear definitions and guidelines about different supplemental approaches, there may be confusion or ambiguity about the sampling strategy used. For example, while the detailed descriptions in the country-level reports of the Young Lives Study suggest that the original sampling frame was used, i.e., refreshment approach, this was described as a ‘replacement’ strategy. The current paper offers clear and easily applied definitions of two supplemental approaches. This type of more consistent, wide-spread reporting may help methodologists to refine recommendations related to supplemental samples at the intersection of planned and unplanned missing data.

However, additional research is needed to better understand the general conditions under which supplemental samples will enhance longitudinal designs. For example, the current study assumed the comparability of units in the refreshment sample and the stability of population characteristics; that is, we assume that the refreshment sample is representative of a population that does not change over time in unobservable ways (Deng et al., 2013). However in practice, refreshment samples may be also plagued by implementation difficulties (Chapman, 2003; Vehovar, 2003). For example, previous research has documented that refreshment samples may be more similar to early responders or those retained in the longitudinal study, than the initial overall sample (Chapman & Roman, 1985). If these assumptions are not met and practical issues are not overcome, the use of refreshment sample may be undermined.

Moreover, for data that are MAR, comparing refreshment samples to replacement samples with appropriate bias correction method is warranted. At the same time, the practicality of accurate sample weights applied at the data analysis phase or guiding the selection of the replacement sample, as with propensity score matching (Dorsett, 2010), relies on accurate

information about the original sampling frame and population characteristics. The advantages of both types of supplemental samples in practice, therefore, may be limited (Vehovar, 1999).

Overall, this paper aimed to suggest guidance for a broad range of behavioral science researchers who are conducting longitudinal research and facing permanent attrition. Findings suggest that even for smaller studies, *refreshment* samples have better coverage rates than *replacement* samples or not supplementing the missing participants in any way. Moreover, refreshment samples also have high power. Therefore, if conditions on the ground or the funding climate changes, researchers can supplement longitudinal samples using the refreshment approach to assess change. Conducting preliminary analyses to identify auxiliary variables may be useful for later estimation, but not for selecting replacement participants. In fact, relying on these variables to select a supplemental strategy response may increase bias and decrease the coverage rate, particularly if bias correction techniques are not taken into account. These findings have implications for developmental researchers working with at-risk populations or samples with high levels of permanent attrition.

References

- Alemu T, Asgedom G, Liebenberg J, Mekonnen, A., Seager, J., Tefera, B., & Woldehanna, T. (2003). *Young Lives Preliminary Country Report: Ethiopia*. London: Young Lives.
Retrieved from: https://www.younglives.org.uk/sites/www.younglives.org.uk/files/YL-CR1-Ethiopia-2003-revised_2.pdf
- Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1-59). New York: Academic Press.
- 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>
- Barnett, I., Ariana, P., Petrou, S., Penny, M.E., Thuc Duc, L., ... Boyden, J. (2013). Cohort Profile: The Young Lives Study. *International Journal of Epidemiology, 42*(3), 701-708. <https://doi.org/10.1093/ije/dys082>
- Betancourt, T. S., Borisova, I. I., Williams, T. P., Brennan, R. T., Whitfield, T. H., de la Soudiere, M., ... & Gilman, S. E. (2010). Sierra Leone's former child soldiers: A follow-up study of psychosocial adjustment and community reintegration. *Child Development, 81*(4), 1077-1095. <https://doi.org/10.1111/j.1467-8624.2010.01455.x>
- Boxer, P., Rowell, L., Dubow, E. F., Landau, S. F., Gvirsman, S. D., Shikaki, K., & Ginges, J. (2013). Exposure to violence across the social ecosystem and the development of aggression: A test of ecological theory in the Israeli–Palestinian conflict. *Child Development, 84*(1), 163–177. <https://doi.org/10.1111/j.1467-8624.2012.01848.x>
- Bradley, J. V. (1978). Robustness? *The British Journal of Mathematical & Statistical Psychology, 31*(2), 144-152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>

- Browning, C. R., Burrington, L. A., Leventhal, T., & Brooks-Gunn, J. (2008). Neighborhood structural inequality, collective efficacy, and sexual risk behavior among urban youth. *Journal of Health and Social Behavior*, 49(3), 269-285.
<https://doi.org/10.1177/002214650804900303>
- Capaldi, D. M., & Patterson, G. R. (1987). An approach to the problem of recruitment rates for longitudinal research. *Behavioral Assessment*, 9(2), 169-178.
- Chapman, D. W. (2003). To substitute or not to substitute – that is the question. *Survey Statistician*, 48, 32–34.
- Conger, R. D., & Conger, K. J. (2002). Resilience in Midwestern families: Selected findings from the first decade of a prospective, longitudinal study. *Journal of Marriage and Family*, 64(2), 361-373. <https://doi.org/10.1111/j.1741-3737.2002.00361.x>
- Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). In Y. Freund, L. Györfi, G. Turán, & T. Zeugmann (Eds.), *Algorithmic Learning Theory*. ALT 2008. (Lecture Notes in Computer Science, 5254). Springer, Berlin, Heidelberg.
- Cotter, R. B., Burke, J. D., Loeber, R., & Navratil, J. L. (2002). Innovative retention methods in longitudinal research: A case study of the developmental trends study. *Journal of Child and Family Studies*, 11(4), 485-498. <https://doi.org/10.1023/A:1020939626243>
- Cummings, E.M., Merrilees, C.E., Taylor, L.K., Shirlow, P., Goeke-Morey, M., & Cairns, E. (2013). Longitudinal relations between sectarian and non-sectarian community violence and child adjustment in Northern Ireland. *Development and Psychopathology*, 25(3), 615-627. <https://doi.org/10.1017/S0954579413000059>

- D'Amico, A. M. (2011). Appendix A.2. High School and Beyond Longitudinal Study. In *National Center for Education Statistics Sources*. Retrieved from:
http://nces.ed.gov/programs/digest/d11/app_a2_9.asp
- Dancey, C., & Reidy, J. (2004). *Statistics without maths for psychology: Using SPSS for Windows*. London: Prentice Hall.
- Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., & Zeng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28(2), 238-256. <https://doi.org/10.1214/13-STS414>
- Dorsett, R. (2010). Adjusting for nonignorable sample attrition using survey substitutes identified by propensity score matching: An empirical investigation using labour market data. *Journal of Official Statistics*, 26 (1), 105–125. Retrieved from:
<http://www.jos.nu/Articles/abstract.asp?article=260105>
- Elliot, D. S., Huizinga, D., & Menard, S. (1989). *Multiple problem youth: Delinquency, substance use, and mental health problems*. New York: Springer-Verlag. doi: 10.1007/078-1-4613-9637-6
- Enders, C.K. (2010). *Applied missing data analysis*. New York, NY: Guildford Press.
- Enders, C.K. (2013). Dealing with missing data in developmental research. *Child Development Perspectives*, 7(1), 27-31. <https://doi.org/10.1111/cdep.12008>
- English Longitudinal Study of Ageing (ELSA) (2015). User Guide for the Wave 0 (HSE) Core Datasets Retrieved from:
http://www.esds.ac.uk/doc/5050%5Cmrdoc%5Cpdf%5C5050_Wave_0_User_Guide.pdf
- Frick, J. R., Goebel, J., Schechtman, E., Wagner, G. G., and Yitzhaki, S. (2006). Using analysis of Gini (ANOGI) for detecting whether two subsamples represent the same universe: The

- German SocioEconomic Panel Study (SOEP) experience. *Sociological Methods Research*, 34(4), 427–468. <https://doi.org/10.1177/0049124105283109>
- Fritz, M. S. (2014). An exponential decay model for mediation. *Prevention science*, 15(5), 611–622. <https://doi.org/10.1007/s11121-013-0390-x>.
- Gouskova, E., Heeringa, S. G., McGonagle, K., & Schoeni, R. F. (2008, June). Technical Report Panel Study of Income Dynamics Revised Longitudinal Weights 1993-2005. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI. Retrieved from: <https://psidonline.isr.umich.edu/data/weights/Long-weights-doc.pdf>
- Graham, J. W. & Donaldson, S. I. (1993). Evaluating interventions with differential attribution: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, 78(1), 119-128. <http://dx.doi.org/10.1037/0021-9010.78.1.119>
- Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31(2), 197-218. https://doi.org/10.1207/s15327906mbr3102_3
- Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In *New methods for the analysis of change*, by J. W. Graham, B. J. Taylor, & P. E. Cumsille (335-353). Washington, DC: American Psychological Association. <http://dx.doi.org/10.1037/10409-011>
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data in psychological research. *Psychological Methods*, 11(4), 323-343. <http://dx.doi.org/10.1037/1082-989X.11.4.323>

- 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>
- Hartsough, C. S., Babinski, L. M., & Lambert, N. M. (1996). Tracking procedures and attrition containment in a long-term follow-up of a community-based ADHD sample. *Journal of Child Psychology and Psychiatry*, 37(6), 705-713. <https://doi.org/10.1111/j.1469-7610.1996.tb01462.x>
- Hirano, K., Imbens, G. W., Ridder, G. and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6), 1645–1659. <https://doi.org/10.1111/1468-0262.00260>
- Höfler, M., Pfister, H., Lieb, R., & Wittchen, H. (2005). The use of weights to account for non-response and drop-out. *Social Psychiatry and Psychiatric Epidemiology*, 40(4), 291-299. <https://doi.org/10.1007/s00127-005-0882-5>
- 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.
- Jorgensen, T. D., Rhemtulla, M., Schoemann, A., McPherson, B., Wu, W., & Little, T. D. (2014). Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies. *International Journal of Behavioral Development*, 38(5), 397-410. doi:10.1177/0165025414531094
- Karriker-Jafee, K. J., Foshee, V. A., Ennett, S. T., Suchindran, C. (2013). Associations of neighborhood and family factors with trajectories of physical and social aggression during adolescence. *Journal of Youth and Adolescence*, 42(6), 861-877. <https://doi.org/10.1007/s10964-012-9832-1>

- Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology & Psychiatry*, 35(2), 259-281. doi:10.1111/j.1469-7610.1994.tb01161.x
- Kronenberg, M. E., Hansel, T. C., Brennan, A. M., Osofsky, H. J., Osofsky, J. D., & Lawrason, B. (2010). Children of Katrina: Lessons learned about postdisaster symptoms and recovery patterns. *Child Development*, 81(4), 1241-1259. <https://doi.org/10.1111/j.1467-8624.2010.01465.x>
- Kruse, Y., Callegaro, M., Dennis, J., Subias, S., Lawrence, M., DiSogra, C., and Tompson, T. (2009). Panel conditioning and attrition in the AP-Yahoo! News Election Panel Study. In 64th conference of the *American Association for Public Opinion Research (AAPOR)*, Hollywood, FL. Retrieved from: https://www.researchgate.net/publication/253431812_Panel_Conditioning_and_Attrition_in_the_AP-Yahoo_News_Election_Panel_Study
- Langkamp, D. L., Lehman, A., & Lemeshow, S. (2010). Methods: Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics*, 10(3), 205-210. <https://doi.org/10.1016/j.acap.2010.01.005>
- Leventhal, T., & Brooks-Gunn, J. (2011). Changes in neighborhood poverty from 1990 to 2000 and youth's problem behaviors. *Developmental Psychology*, 47(6), 1680-1698. <https://doi.org/10.1037/a0025314>
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421), 125-134. <https://doi.org/10.1080/01621459.1993.10594302>

- Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431), 1112-1121.
<https://doi.org/10.1080/01621459.1995.10476615>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
- Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, 7(4), 199-204. doi:10.1111/cdep.12043
- Lucas, R. E. (2005). Time does not heal all wounds: A longitudinal study of reaction and adaptation to divorce. *Psychological Science*, 16(12), 945-950.
<https://doi.org/10.1111/j.1467-9280.2005.01642.x>
- Magee, L., Robb, A., & Burbridge, J. (1998). On the use of sampling weights when estimating regression models with survey data. *Journal of Econometrics*, 84(2), 251-271.
[https://doi.org/10.1016/S0304-4076\(97\)00086-9](https://doi.org/10.1016/S0304-4076(97)00086-9)
- Maldonado-Molina, M. M., Jennings, W. G., & Komro, K. A. (2010). Effects of alcohol on trajectories of physical aggression among urban youth: An application of latent trajectory modeling. *Journal of Youth and Adolescence*, 39(9), 1012-1026.
<https://doi.org/10.1007/s10964-009-9484-y>
- Mason, M. J. (1999). A review of procedural and statistical methods for handling attrition and missing data in clinical research. *Measurement & Evaluation in Counseling & Development (American Counseling Association)*, 32(2), 111-118.
- Mazen, J. & Tong, X. (2018). Bias correction for replacement samples in longitudinal research. Developmental Methods Conference. Whitefish, MT. Available from:

http://www.developmentalmethods.org/wp-content/uploads/2018/08/DM2018_ConferenceProgram_FINAL.pdf

Mistler, S. A., & Enders, C. K. (2012). Planned missing data designs for developmental research. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 742–754). New York, NY: Guilford Press.

National Youth Longitudinal Survey (NYLS, 1997). *Sample Weights & Design Effects*. Retrieved from: <https://www.nlsinfo.org/content/cohorts/nlsy97/using-and-understanding-the-data/sample-weights-design-effects>

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, 6(3), 328-362. <https://doi.org/10.1177/1094428103254673>

Nicholson, J., Deboek, P. R., & Howard, W. (2017). Attrition in Developmental Psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, 41(1), 143-153. <https://doi.org/10.1177/0165025415618275>

Nuttall, A., Valentino, K., & Borkowski, J (2012). Maternal history of parentification, maternal warm responsiveness, and children's externalizing behavior. *Journal of Family Psychology*, 26(5), 767-775. <http://dx.doi.org/10.1037/a0029470>

Pearl, J. (2018, January 10). Facts and Fiction from the “Missing data Framework”. Retrieved from: <http://causality.cs.ucla.edu/blog/index.php/2018/01/10/facts-and-fiction-in-the-missing-data-framework/#comments>

- Pierce, B., & Hartford P. (2004). White Paper on Evaluation of Sampling Design Options for the National Children's Study. Rockville, MD: National Children's Study Program Office National Institute for Child Health and Human Development.
- R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from <http://www.R-project.org> (ISBN 3-900051-07-0)
- Raudenbush, S. W. & Chan, W. (1993). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology*, 61(6), 941-951.
- Rhemtulla, M., Jia, F., Wu, W., & Little, T. D. (2014). Planned missing designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavioral Development*, 38(5), 423-434. doi:10.1177/0165025413514324
- Rhemtulla, M., Savalei, V. & Little, T. D. (2016). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, 81(1), 60-89. doi:10.1007/s11336-014-9422-0
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
<https://doi.org/10.1093/biomet/63.3.581>
- Savalei, V. (2010). Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 1-22.
<https://doi.org/10.1080/10705510701758091>
- Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 477-497. <https://doi.org/10.1080/10705510903008238>

- Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 280-302.
<https://doi.org/10.1080/10705511.2014.882692>
- Schoemann, A. M., Miller, P., Pornprasertmanit, S., & Wu, W. (2014). Using Monte Carlo simulations to determine power and sample size for planned missing designs. *International Journal of Behavioral Development*, 38(5), 471-479.
<https://doi.org/10.1177/0165025413515169>
- Scott, C. K. (2004). A replicable model for achieving over 90% follow-up rates in longitudinal studies of substance abusers. *Drug and Alcohol Dependence*, 74(1), 21-36.
<https://doi.org/10.1016/j.drugalcdep.2003.11.007>
- Si, Y., Reiter, J. P., & Hillygus, D. S. (2015). Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples, *Political Analysis*, 23(1), 92-112. <https://doi.org/10.1093/pan/mpu009>
- Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, 62(3), 494-507. doi:10.2307/2657318
- Thornberry, T. P., Bjerregaard, B., & Miles, W. (1993). The consequences of respondent attrition in panel studies: A simulation based on the Rochester Youth Development Study. *Journal of Quantitative Criminology*, 9(2), 127-158. <https://doi.org/10.1007/BF01071165>
- van Soest, A., & Kapteyn, A. (2011). Mode and context effects in measuring household assets. In M. Das, P. Ester, & L. Kaczmirek (Ed.), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 291-323). New York, NY, US: Routledge/Taylor & Francis Group, New York, NY.

- Vehovar, V. (2003). Field substitutions redefined. *Survey Statistician*, 48, 32–34. Available from <http://isi-iass.org/home/wp-content/uploads/N48.pdf>.
- Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics*, 15(2), 335–350.
- Widaman, K. F. (2006). Best practices in quantitative methods for developmentalists: 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>
- Yuan, K. H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis*, 100(9), 1900–1918. <https://doi.org/10.1016/j.jmva.2009.05.001>
- Yuan, K. H. & Lu, L. (2008). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, 43(4), 621–652. <https://doi.org/10.1080/00273170802490699>
- Yuan, K. H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, 77(4), 803–826. <https://doi.org/10.1007/s11336-012-9282-4>
- Yuan, K. H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML Versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research*, 41(4), 598–629. <https://doi.org/10.1177/0049124112460373>
- Zwane, A. P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., ... & Duflo, E. (2011). Being surveyed can change later behavior and related parameter estimates. *Proceedings of the National Academy of Sciences*, 108(5), 1821–1826. <https://doi.org/10.1073/pnas.1000776108>

Table 1

Retention percentage of the original sample at each measurement occasion based on the proportion of missingness (r).

r	Time 1	Time 2	Time 3	Time 4	Time 5
3%	100%	97%	94%	91%	89%
5%	100%	95%	90%	86%	81%
8%	100%	92%	85%	78%	72%
15%	100%	85%	72%	61%	52%

Table 2

Real Data Model Estimation Results: Trajectories of Adolescent Attitudes toward Social Deviance (N = 120)

	Original	Refreshment Sample			Replacement Sample		
		RF(1)	RF(2)	RF(3)	RP(1)	RP(2)	RP(3)
Ave(intercept)	11.36* (0.21)	11.32*	11.31*	11.31*	11.37*	11.36*	11.36*
		(0.21)	(0.20)	(0.20)	(0.21)	(0.21)	(0.20)
Ave(slope)	0.85* (0.10)	0.83*	0.83*	0.81*	0.88*	0.87*	0.88*
		(0.09)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)
Var(intercept)	2.87* (0.82)	2.77*	2.76*	2.76*	2.87*	2.94*	2.89*
		(0.81)	(0.81)	(0.80)	(0.82)	(0.84)	(0.82)
Var(slope)	0.54* (0.19)	0.55*	0.52*	0.52*	0.60*	0.59*	0.66*
		(0.19)	(0.18)	(0.17)	(0.21)	(0.20)	(0.20)
Cov(intercept,slope)	0.44 (0.24)	0.46*	0.46*	0.42	0.48*	0.51*	0.43
		(0.23)	(0.22)	(0.22)	(0.25)	(0.24)	(0.25)

Note: * $p < .05$; standard errors reported in parenthesis. RF denotes a refreshment and RP a replacement sample, with the number in parentheses indicating the size of the supplemental sample at Time 3.

Figure 1. Model used to generate complete cases across all conditions, and as the estimation model to assess the impact of the adding the supplemental samples. This represents a linear growth model (LGM) across five time points, equally space for each participant; errors (e_i) followed a univariate normal distributions.

Figure 2a. Plots of relative bias for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing completely at random (MCAR) simulation study and $\text{var}(e) = 3$. Missing proportion (r) is the amount of missing data from one time point to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and list-wise deletion (MCAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all

estimations except for the listwise deletion analyses.

Figure 2b. Plots of relative efficiency for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing completely at random (MCAR) simulation study and $\text{var}(e) = 3$. Missing proportion (r) is the amount of missing data from one time point to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment

sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and listwise deletion (MCAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure 2c. Plots of coverage rates for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing completely at random (MCAR) simulation study and $\text{var}(\epsilon) = 3$. Missing proportion (r) is the amount of missing data from one time point to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and listwise deletion (MCAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure 2d. Plots of statistical power for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing completely at random (MCAR) simulation study and $\text{var}(e) = 3$. Missing proportion (r) is the amount of

missing data from one time point to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and listwise deletion (MCAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure 3a. Plots of relative bias for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing at random (MAR) simulation study and $\text{var}(e) = 3$. Missing proportion (r) is the amount of missing data from one time point to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment and RP a replacement

parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and list-wise deletion (MAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure 3b. Plots of relative efficiency for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing at random (MAR) simulation study and $\text{var}(e) = 3$. Missing proportion (r) is the amount of missing data

from one time point to the next,

following a pattern of permanent attrition (see Table 1). RF denotes a refreshment and RP a replacement sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and list-wise deletion (MAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure 3c. Plots of coverage rates for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing at random (MAR) simulation study and $\text{var}(e) = 3$. Missing proportion (r) is the amount of missing data from one time point to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment and RP a replacement sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and listwise deletion (MAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure
3d.
Plots of

statistical power for estimates of the average latent slope parameter (β_S) in the simulated datasets associated with the condition of missing at random (MAR) simulation study and $\text{var}(\epsilon) = 3$. Missing proportion (r) is the amount of missing data from one time point

to the next, following a pattern of permanent attrition (see Table 1). RF denotes a refreshment and RP a replacement sample, with the number in parentheses indicating the size of the supplemental sample at Time 3. Complete case (Complete) and listwise deletion (MAR-LIST) are provided as comparisons for the supplemental approaches. The two-stage estimation approach was used in all estimations except for the listwise deletion analyses.

Figure 4. Individual plots of the random sample of 120 adolescents across five measurement occasions. On average, the log transformation of attitudes toward deviance go up with age.