Evaluation of supplemental samples in longitudinal research with non-normal missing data

Jessica A. M. Mazen¹ · Xin Tong¹ · Laura K. Taylor²

Published online: 13 August 2018
© Psychonomic Society, Inc. 2018

Abstract

Missing data is a commonly encountered problem in longitudinal research. Methodological articles provide advice on ways to handle missing data at the analysis stage, however, there is less guidance for researchers who wish to use supplemental samples (i.e., the addition of new participants to the original sample after missing data appear at the second or later measurement occasions) to handle attrition. The purpose of this study is to evaluate the effects of using supplemental samples when analyzing longitudinal data that are non-normally distributed. We distinguish between two supplemental approaches: a refreshment approach where researchers select additional participants using the same criteria as the initial participants (i.e., random selection from the population of interest) and a replacement approach where researchers identify auxiliary variables that explain missingness and select new participants based on those attributes. Overall, simulation results suggest that the addition of refreshment samples, but not replacement samples, is an effective way to respond to attrition in longitudinal research. Indeed, use of refreshment samples may reduce bias of parameter estimates and increase efficiency and statistical power, whereas use of replacement samples results in biased parameter estimates. Our findings may be utilized by researchers considering using supplemental samples and provide guidance for selecting an appropriate supplemental sample approach.

Keywords Supplemental sample · Refreshment sample · Replacement sample · Missing data · Non-normal data · Longitudinal design

The popularity of longitudinal research is growing and over the past two decades there has been a considerable rise in attention paid to longitudinal theory, methodology, and application. Longitudinal research provides valuable insight into the predictors and outcomes associated with a variety of topics such as childhood abuse (Feiring, Simon, & Cleland, 2009), mental illness (Strickler, Whitley, Becker, & Drake, 2009), political violence (Taylor, Merrilees, Goeke-Morey, Shirlow, & Cummings, 2016), and war (Betancourt et al., 2010). Although longitudinal research may further our understanding in human development, it is often encumbered with methodological challenges. One such challenge is that missing data frequently arise. Researchers

use a variety of approaches to handle missingness including missing data analysis techniques such as full information maximum likelihood, retention and tracking techniques, planned missingness designs, and supplemental samples. In contrast to other approaches, little research has examined the consequences of using supplemental samples despite their wide implementation (e.g., Adler, 1994; Casswell et al., 2012; Fong et al., 2006). Consequently, there is a need to evaluate the efficacy of supplemental samples, and in this article, we fill the gap in the existing literature by investigating the effects of supplemental samples on model estimation in longitudinal research.

Missing data appear in all types of research designs. However, the problem of missing data is compounded in longitudinal designs because participants are repeatedly measured at multiple time points. In such designs, attrition occurs as participants may be absent at one or more time points. Attrition may be permanent (a participant drops out of the study and does not return) or intermittent (a participant is not available for one or more measurement occasions, but returns at later waves of data collection).

✉ Jessica A. M. Mazen
jm5ku@virginia.edu

¹ Department of Psychology, University of Virginia, Charlottesville, VA, USA

² School of Psychology, Queen's University Belfast, Belfast, UK

Attrition rates vary across the literature. Authors of a recent meta-analysis of 92 longitudinal studies examining personality traits reported a 44% average attrition rate across samples (Roberts, Walton, & Viechtbauer, 2006). Whereas, in a systematic literature review of attrition across 25 population-based longitudinal studies of the elderly attrition rates varied from 5 to 50% (Chatfield, Brayne, & Matthews, 2005). Attrition may be especially problematic in longitudinal study designs with at-risk populations (Betancourt et al., 2010; Goemans, van Geel, & Vedder, 2015; Kronenberg et al., 2010); in some cases it can be as high as 85% (Goemans et al., 2015). Thus, attrition is a commonly encountered problem when conducting longitudinal research, and in certain situations (e.g., working with at-risk populations) researchers may lose a substantial portion of the original sample. Therefore, many researchers face the challenge of dealing with missing data over time.

Missing data mechanisms refer to the process that causes missing data (Little & Rubin, 2002; Rubin, 1976). Missing completely at random (MCAR) defines missingness on the outcome variable Y as completely independent of other variables that influence Y . For example, suppose researchers investigating changes in math ability across elementary school encounter attrition. Missingness would be MCAR if the probability of missingness is unrelated to any variable that affects math scores (e.g., a student happened to be sick on the day of the math test). The missing at random (MAR) mechanism refers to missingness on Y that is related to an observed variable (or an auxiliary variable) that affects Y . For example, suppose test anxiety affects both math scores and the probability of having a missing score (e.g., students with anxiety tend to skip the test more than less anxious students). The MAR mechanism would hold if test anxiety were measured and included in the dataset as an auxiliary variable. However, if test anxiety were not measured, the missing data mechanism would be missing not at random (MNAR). The MNAR missing data mechanism defines missingness on Y that is related to one or more unobserved variables that influence Y , or to unobserved values in Y . Thus, the missing data mechanism is MNAR if the missingness on Y is a function of Y itself. For example, students who would have scored low on the math test were more likely to skip the test (see Thoemmes and Mohan (2015) for a graphical representation of the missingness mechanisms).

The prevalence of missing data in longitudinal designs has engendered considerable research into how best to approach the problem of attrition. Different missing data analysis techniques have been developed. One of the most common techniques to dealing with missing data is through the deletion of cases (e.g., listwise or pairwise deletion; Jeličić, Phelps, & Lerner, 2009). Case deletion methods have a number of disadvantages including producing biased parameter estimates under MAR or MNAR and lacking

efficiency in all situations (Acock, 2005; Peugh & Enders, 2004; Schafer & Graham, 2002). Therefore, although deletion techniques are still widely used, modern methods for handling missing data such as full information maximum likelihood (FIML) and multiple imputation (MI) are increasing in popularity. FIML and MI are recommended because they require less stringent assumptions than case deletion methods (Enders & Bandalos, 2001; Peugh & Enders, 2004; Schafer & Graham, 2002). Given sufficient sample size, FIML and MI produce unbiased estimates under both MCAR and MAR and are generally more powerful than deletion techniques. In addition to handling missingness at the analysis stage, researchers may try to combat the problem of missing data during data collection by limiting attrition through the use of retention (e.g., increased financial incentive over time) and tracking techniques (e.g., obtaining contact information of friends/family of participants; see Ribisl et al. (1996) for a review). Retention and tracking techniques can lower rates of attrition, but are time-intensive and expensive (Hobden, Forney, Durham, & Toro, 2011; Kuhns, Vazquez, & Ramirez-Valles, 2008; Nemes, Wish, Wraight, & Messina, 2002).

Planned missingness designs are another way researchers attempt to reduce attrition in longitudinal research. In these designs, researchers intentionally collect incomplete data from participants by randomly assigning a subset of participants to have missing items (e.g., three-form design; Graham, Hofer, & MacKinnon, 1996; Graham, Taylor, Olchowski, & Cumsille, 2006; Rhemtulla, Jia, Wu, & Little, 2014), missing measures (e.g., two-method measurement design; Garnier-Villarreal, Rhemtulla, & Little, 2014; Graham et al., 2006), or missing measurement occasions (e.g., wave-missing design; Graham, Taylor, & Cumsille, 2001; Mistler & Enders, 2012; Rhemtulla et al., 2014). For example, in a wave-missing design, participants are assigned *a priori* to miss entire waves of data collection. That is, researchers may plan for complete data at the first time point and 50% of participants to be missing at each of the four subsequent time points. Therefore, each participant would only be expected to provide data at three instead of five measurement occasions. Planned missingness designs allow researchers to collect more informative data, decrease cost and time of data collection, lessen practice effects, and reduce unplanned missing data by decreasing participant burden (Harel, Stratton, & Aseltine, 2015; Jorgensen et al., 2014; Little & Rhemtulla, 2013). Furthermore, planned missingness designs (in combination with modern missing data approaches) produce accurate and efficient parameter estimates (Garnier-Villarreal et al., 2014; Graham et al., 1996; Rhemtulla et al., 2014) and in some cases increase statistical power (Graham et al., 2006).

In the same vein, researchers who experience or expect to experience some level of attrition may use supplemental

samples to reduce the detrimental effects of unplanned missingness. Supplemental sample designs fit within the planned missingness design framework as researchers intentionally collect incomplete data from a subset of participants, similar to wave-missing designs. Although supplemental samples are widely used in practice, there is little research evaluating their performance and thus, little guidance for effectively implementing them in longitudinal designs. Therefore, the purpose of this study is to further study the effects of using supplemental samples. After an introduction to the different types of supplemental samples and a review of their current use, we present a Monte Carlo simulation study in a growth curve modeling framework. In the study, we investigate the consequences of adding two different types of supplemental samples on model parameter estimation. Further, we explore these effects across various conditions meant to replicate the wide variety of conditions researchers encounter when conducting longitudinal research. We conclude with recommendations for using supplemental samples in longitudinal designs and for future methodological research.

Supplemental samples

In the context of longitudinal research, a supplemental sample is a set of new participants added to the original sample after missing data appear at the second or later measurement occasions in order to reduce the detrimental effects of unplanned missingness. There are two approaches researchers can take when selecting a supplemental sample: a refreshment approach or a replacement approach. Using a refreshment approach, researchers select additional participants using the same criteria as the initial participants (i.e., random selection from population of interest). In this case, regardless of the type of missing data pattern (e.g., MCAR vs. MAR; permanent vs. intermittent attrition), researchers follow the initial study design utilizing the original sampling frame and procedures to recruit new participants. In contrast, using a replacement approach, researchers first identify auxiliary variables that explain the pattern of missingness in the data and then select new participants based on those attributes. Thus, researchers are attempting to replace participants from the original sample lost to attrition by over-selecting participants with the same characteristics for the supplemental sample. Note that it is not possible to use the replacement approach when the missing data mechanism is MCAR because the missingness is truly random and not related to any variables in the data set; therefore, no relevant auxiliary variables exist to guide the selection of the supplemental sample.

Returning to the above example, suppose the researchers investigating changes in math ability decide to add a supple-

mental sample to address the attrition present in their study. The first step would be to identify auxiliary variables to interpret the missing data mechanism. If no auxiliary variable can be identified, they may believe the missingness to be MCAR. Therefore, researchers would use a refreshment approach to supplement the sample by randomly selecting grade school children in the United States. However, if one or more auxiliary variables can be identified so the missing data mechanism was MAR, researchers would have a choice between using a refreshment or replacement approach. Choosing a replacement approach, researchers would first obtain the auxiliary variable, in this case test anxiety, and then over-select participants for the supplemental sample based on this variable. Thus, if students with greater test anxiety are more likely to be missing, researchers would over-select for children with high test anxiety.

Supplemental samples have been used in numerous studies in the field of psychology and beyond. For instance, International Tobacco Control Policy Evaluation Project (ITC Project) researchers use a longitudinal design to measure the impact of tobacco control policies across multiple countries and use a refreshment approach to supplement their sample ensuring adequate sample size at each wave (Fong et al., 2006). Additionally, supplemental samples are utilized to address attrition in numerous other large-scale studies including the Medicare Current Beneficiary Survey (MCBS; Adler, 1994), the International Alcohol Control Study (IAC; Casswell et al., 2012), the Survey of Health, Ageing and Retirement in Europe (SHARE; Börsch-Supan et al., 2013), and the English Longitudinal Study of Ageing (ELSA; Steptoe, Breeze, Banks, & Nazroo, 2013). Data generated from these projects has, thus far, resulted in over 2600 published articles (Centers for Medicare Medicaid Services, 2012; ELSA, 2016; IAC Study, 2016; ITC Project, 2015; SHARE, 2016). Therefore, the effects of including supplemental samples are bound to have far-reaching implications. Moreover, the utilization of supplemental samples is not unique to these five projects. Indeed, numerous other studies examining a wide-range of topics include supplemental samples to address unplanned missingness (see Beltrán-Sánchez, Drumond-Andrade, & Riosmena, 2015; Edwards, Cherry, & Peterson, 2000; Fillit, Guterman, & Brooks, 2000; Krause et al., 2002; Liao et al., 2015; Nicklas, 1995; Taylor & Lynch, 2011; Tubman, Windle, & Windle, 1996; Windle & Windle, 2001). Despite the prevalence of supplemental samples, little research has investigated the effects of using this approach on parameter estimates. Consequently, researchers who hope to use this approach are at a loss as to how supplemental samples can be best utilized in psychological studies.

Thus far, only one study has systematically studied the effects of utilizing supplemental samples on model parameters

under conditions commonly observed in psychological literature. Taylor, Tong, and Maxwell (2018) compared refreshment and replacement approaches to adding supplemental samples in growth curve modeling with MCAR and MAR data. Findings revealed that refreshment samples yielded unbiased parameter estimates similar to the complete data analysis, acceptable coverage rates, and narrower confidence intervals for the population mean slope parameter. However, using a replacement approach resulted in biased estimates and unacceptable coverage rates. Further, as the size of the replacement sample increased, the bias increased while the coverage rate decreased. Although these findings further our understanding of the effects of adding supplemental samples, there are several limitations that may limit the applicability of these findings to real-world studies.

First, this study only focused on normally distributed data. However, practical data are rarely normally distributed in the social and behavioral sciences (Micceri, 1989). They may have heavy tails or be contaminated. When data are non-normally distributed, traditional normal-distribution-based maximum likelihood (NML) methods may lead to incorrect parameter estimates and unexpected misleading statistical inferences. Therefore, under such circumstances, adding supplemental samples may or may not perform the same as when data are normally distributed. Second, Taylor et al. (2018) only added supplemental samples at one measurement occasion, whereas in practice, it is equally likely that supplemental samples will be added at multiple occasions; in some cases they are added at each measurement occasion (Fong et al., 2006). Thus, the utility of gradually adding supplemental samples at multiple time points compared to adding a larger sample at one occasion needs to be investigated. Third, in Taylor et al. (2018), missingness only represented permanent attrition, yet, both permanent and intermittent attrition are common in longitudinal designs. Thus, supplemental samples should be evaluated when both patterns of missingness are present. Fourth, Taylor et al. (2018) evaluated the two types of supplemental samples based on an unconditional multilevel model. Since covariates are often included in growth curve models to study the influence of other variables (e.g., individual characteristics) on growth parameters, we include a time-invariant covariate. Finally, as research demonstrates that high rates of attrition are not uncommon in longitudinal designs (Roberts et al., 2006), we increase the number of measurement occasions from five (Taylor et al., 2018) to eight to replicate high-attrition situations. Consequently, it is necessary to systematically study the influence of supplemental samples under these circumstances in order to provide more comprehensive guidelines to applied researchers and this article fulfills this aim based on a Monte Carlo simulation study.

Evaluation of supplemental samples through a simulation study

Study design

A simulation study is conducted to compare the influence of adding different supplemental samples in analyzing longitudinal data. Latent growth curve modeling is commonly used for longitudinal research as it can directly investigate intraindividual change over time and interindividual differences in intraindividual change; thus, we study the performance of supplemental samples in the growth curve modeling framework. A typical form of a conditional linear growth curve model can be expressed as

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

$$\mathbf{b}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}_i + \mathbf{u}_i,$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$ is a $T \times 1$ random vector with y_{ij} representing an observation for individual i at time j ($i = 1, \dots, N$; $j = 1, \dots, T$). Here N is the sample size and T is the total number of measurement occasions. The matrix $\Lambda = ((1, 1, \dots, 1)', (0, 1, \dots, T-1)')$ is a $T \times 2$ factor loading matrix determining the growth trajectories, $\mathbf{b}_i = (b_{iL}, b_{iS})'$ is a 2×1 vector of random effects with b_{iL} representing the random intercept and b_{iS} representing the random slope, and $\mathbf{e}_i$ is a vector of intraindividual measurement errors. A $2 \times q$ vector $\mathbf{x}_i$ is a vector of q covariates which are used to predict the random effects $\mathbf{b}_i$ for each individual. The regression coefficients $\boldsymbol{\beta}_0 = (\beta_{0L}, \beta_{0S})'$ and $\boldsymbol{\beta}_1 = (\beta_{1L}, \beta_{1S})'$ are fixed effects. In our study, we simulate one covariate ($q = 1$), but our conclusions can be generalized to models with more than one covariate and more complicated models such as quadratic growth curve models if the model is not misspecified. The residual vector $\mathbf{u}_i$ represents the random component of $\mathbf{b}_i$.

To systematically evaluate the effect of adding supplemental samples in longitudinal studies, in the simulation, nine possible influential factors are studied including sample size, number of measurement occasions, distribution of the population, variance of measurement errors, missing data mechanism, correlation between the latent slope and auxiliary variable, missing data rate, supplemental sample type, and the size and timing of the supplemental sample. Population parameter values for $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $cov(\mathbf{u}_i)$ are specified a priori and kept constant across conditions.¹ We set $\boldsymbol{\beta}_0 = (6, 0.3)'$ as past longitudinal research on a set of real data (Tong & Zhang, 2012) demonstrate that these values are reasonable. The covariate $\mathbf{x}_i$ is generated from

¹Pilot studies showed that different values of $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $cov(\mathbf{u}_i)$ lead to the same conclusions in the evaluation of the performance of supplemental samples.

a normal distribution $N(10, 1.5)$ and its coefficient $\beta_1 = (1, .1)'$. The covariance matrix for the residual vector $\mathbf{u}_i$ is set as an identity matrix, thus, the latent intercept b_{iL} and slope b_{iS} are uncorrelated. Although the latent intercept and slope may be correlated in real data, past research (Taylor et al., 2018) demonstrates that different values of the correlation between latent intercept and slope do not affect the pattern of adding supplemental samples. Thus, we fixed the correlation to be zero in this study.

Conditions

We explain the simulation conditions regarding the nine manipulated factors in detail below.

First, four different sample sizes are considered: $N = 50$, 200, 500, and 1000. Second, the number of measurement occasions T is either 4 or 8.

Third, the influence of the population distribution is evaluated by manipulating the distributions of intraindividual measurement errors.² In total, six distributional conditions are considered: normal distribution $N(0, \sigma_e^2)$, normal distribution $N(0, \sigma_e^2)$ with 2% outliers, normal distribution $N(0, \sigma_e^2)$ with 8% outliers, gamma distribution $\Gamma_{(1,1)}(0, \sigma_e^2)$, log-normal distribution $LN_{(0,1)}(0, \sigma_e^2)$, and t distribution $t_{(5)}(0, \sigma_e^2)$.³ Outliers are generated from normal distributions with the same variance but shifted means $N(6, \sigma_e^2)$. These conditions are selected because a wide variety of non-normal data are commonly encountered in psychological research (Micceri, 1989) including the occurrence of outliers, large skewness, and large kurtosis. We want to evaluate the performance of supplemental samples in the presence of differing amounts of outliers (0, 2, and 8%), and varying levels of skewness (normal, gamma, and log-normal distributions) and kurtosis (normal, t , gamma, and log-normal distributions). Note that the t distribution has higher than normal kurtosis while the gamma distribution and the log-normal distribution have both larger skewness and larger kurtosis so they deviate more from a normal distribution.

Fourth, σ_e^2 is set to be 1 or 3 to investigate the influence of measurement errors and thus the reliability of $\mathbf{y}_i$. Given the current setting, the reliability of $\mathbf{y}_i$ ranges from .45 to .98.

Fifth, MCAR and MAR missing data mechanisms are investigated. Data are completely observed at the first

measurement occasion and may be missing from the second measurement occasion. For the MCAR mechanism, missing data are generated such that all observations after the first measurement occasion have the same probability to be missing creating both intermittent and permanent patterns of missingness. For the MAR mechanism, an auxiliary variable score (aux_i) is generated for each individual equal to $aux_i = r \times b_{iS} + \epsilon_i$, where r is the correlation between the auxiliary variable and latent slope, b_{iS} is the slope component of the random effects, and ϵ_i are generated from the normal distribution $N(0, \sqrt{1-r^2})$. A cutoff score for each measurement occasion is determined as the $(1 - mp_j/c)$ th quantile of aux_i , where mp_j represents the expected missing rate for the j th measurement occasion and c is a predetermined proportion (e.g., .80 in our simulation). Then for each measurement occasion, $100c\%$ of observations greater than the cutoff score are missing. This procedure guarantees that we generate expected proportions of missing data as well as produce a pattern of missingness that mirrors both permanent and intermittent attrition. Past research (Taylor et al., 2018) investigating the performance of supplemental samples has only considered permanent attrition. In this paper, we add intermittent attrition to more closely resemble real world situations. For the $N \times mp_j$ missing cases at the j th measurement occasion, the proportion of intermittent attrition is $(1 - c^{T-j})$ and the proportion of the permanent attrition is c^{T-j} . Thus, for all datasets with missing values, both MCAR and MAR, missingness represents both permanent and intermittent attrition.

Sixth, the correlation between the latent slope and auxiliary variable is set to be either $r = .3$ or $.8$, representing a relatively small and large correlation, respectively.

Seventh, the influence of the missing data rate (MR; MR = 3, 5, 8, and 15%) is considered by varying the proportion of observations missing between each time point. For example, if the MR = 3%, 3% of the observed data at time T will be missing at time $T + 1$. Therefore, 0 $(1 - (1 - .03)^0 = 0)$ observations are missing at time 1, 3% $(1 - (1 - .03)^1 = .03)$ of the observations are missing at time 2, 5.9% $(1 - (1 - .03)^2 = .059)$ of observations are missing at time 3, and 8.7% $(1 - (1 - .03)^3 = .087)$ are missing at time 4, whereas 19.2% $(1 - (1 - .03)^7 = .192)$ are missing at time 8. In addition to the MCAR and MAR conditions, complete data are analyzed to serve as a benchmark when judging the effect of the missing data rate. We expect that the missing data rate will be associated with bias when using replacement samples such that higher missing data rates equate to greater bias (Taylor et al., 2018).

Eighth, to evaluate the effect of supplemental samples in analyzing longitudinal data with missing values two types of supplemental samples are explored: refreshment samples and replacement samples. Refreshment samples

²We assume the random effects are normally distributed. So the non-normal distribution of observed variables $\mathbf{y}_i$ is controlled by the distribution of measurement errors $\mathbf{e}_i$.

³ $\Gamma_{(1,1)}(0, \sigma_e^2)$ denotes the rescaled log of the standard normal distribution $\Gamma(1, 1)$ with mean 0 and variance σ_e^2 . $LN_{(0,1)}(0, \sigma_e^2)$ denotes the rescaled log of the standard normal distribution with mean 0 and variance σ_e^2 . $t_{(5)}(0, \sigma_e^2)$ denotes the rescaled $t(df=5)$ distribution with mean 0 and variance σ_e^2 .

may be used with both missing data patterns (MCAR and MAR), however, replacement samples cannot be used when missingness is MCAR because given the missing data mechanism, no relevant auxiliary variables can be identified and there is no way to obtain new samples which have the same characteristics to replace the ones that are missing. Refreshment samples are generated identically to the original data. To generate data for the replacement sample, a set of data are generated in the same way as the original data with the addition of generating aux_i for each individual. Next, the value of aux_i is compared to the theoretical mean of the auxiliary variable, if the value is greater, the observation is included in the replacement sample, otherwise the observation is discarded. This process is repeated until the replacement samples reach the appropriate size. After the supplemental samples are generated they are added to the original dataset. There are no missing values in the supplemental sample for the time point the supplemental sample is added, but for subsequent time points the missing data rate for the supplemental samples is identical to the original data. Thus, participants in a replacement sample are more likely to attrite over time than those in a refreshment sample. We expect that using replacement samples will result in biased parameter estimates whereas using refreshment samples will produce estimates comparable to the complete data (Taylor et al., 2018). Further, we hypothesize that using supplemental samples will increase efficiency and power, and decrease the average confidence interval width.

Finally, the size of the supplemental samples and measurement occasion(s) in which the supplemental samples are added is investigated. Three variations of the size and timing of adding supplemental samples are examined: $1 \times$ the number of missing observations at the second measurement occasion added at the third measurement occasion, $(T-2) \times$ the number of missing observations at the second measurement occasion added at the third measurement occasion, and $1 \times$ the number of missing observations at the second measurement occasion added at the third measurement occasion and every subsequent measurement occasion. The three variations of size and timing of the supplemental sample addition are denoted as follows: 'small one-time' indicates that a supplemental sample $1 \times$ the number of missing observations at $T = 2$ is added at $T = 3$, 'large one-time' indicates that a supplemental sample $(T-2) \times$ the number of missing observations at $T = 2$ is added at $T = 3$, and 'repeated' indicates that a supplemental sample $1 \times$ the number of missing observations at $T = 2$ is added at $T = 3$ and every subsequent time point. For example, if there are 4 measurement occasions, a large one-time refreshment sample indicates that a refreshment sample $2 \times$ the number of missing observations at $T = 2$ is added at $T = 3$, whereas

repeated refreshment samples indicates that a refreshment sample $1 \times$ the number of missing observations at $T = 2$ is added at $T = 3$ and every subsequent time point (see Fig. 1 for a visual representation of the different size/timing conditions).

Overall, 7152 simulation conditions were considered. For each condition, a total of 500 data sets were generated and analyzed using the free software R (R Core Team, 2016).

Analysis

As the generated data are non-normally distributed and have missing values, we use a two-stage robust procedure

Fig. 1 Example of three different conditions varying the size of the supplemental sample and measurement occasion(s) in which the supplemental samples are added. Each rectangle depicts an example data matrix where white represents complete data and black missingness. In **a** a small one-time supplemental sample, $1 \times$ the number of missing observations at the second wave is added at the third wave. In **b** a large one-time supplemental sample, $T-2 \times$ the number of missing observations at the second wave is added at the third wave. In **c** repeated supplemental samples, $1 \times$ the number of missing observations at the second wave are added at the third wave and every subsequent wave. Note that the missingness in supplemental samples is not marked out

for structural equation modeling (SEM; Yuan & Zhang, 2012) to analyze the generated data. Previous research demonstrates that the two-stage robust procedure for SEM leads to less biased parameter estimates and more reliable test statistics compared to traditional NML procedures when data are non-normal and contain missing values (Tong, Zhang, & Yuan, 2014). In the first stage of the analysis, robust M-estimates of the saturated mean vector and covariance matrix are generated for all variables using the expectation robust algorithm with Huber-type weights. The robust estimates are used in place of the sample mean and covariance matrix in the second stage of the analysis when estimating the model parameters. For MAR data, the auxiliary variable is included in the first stage of the estimation process, thus allowing for the auxiliary variable to account for missingness in the substantive variables. The R package 'rsem' (Yuan & Zhang, 2012) is used to summarize missing data patterns, obtain the robust saturated mean vector and covariance matrix, and finally yield the parameter estimates and robust standard error estimates (simulation code is available in the supplemental materials available on our website: <http://people.virginia.edu/~jm5ku/>).

Robust procedures differ from NML procedures in how much weight is given to each observation during the estimation process. In NML, all cases are given equal weight. In robust procedures, the weight of each observation is based on its distance from the center of the majority of the data with observations far away from the center assigned smaller weights. The Huber-type weight function is used to assign observations different weights. The tuning parameter φ allows for different weighting schemes. In this study, φ is set at 0.1, meaning observations with $d_i > c_{0.10}$ are assigned weights smaller than 1, where d_i represents the Mahalanobis distance from case i to the center of the majority cases, and $c_{0.10}$ is the critical value corresponding to the 90th percentile of a chi-distribution. Although higher φ values better protect estimates against data contamination, they also result in estimates that are less efficient when the data come from a population that is truly normally distributed. When $\varphi = 0$, no cases are down-weighted and the procedure is identical to the NML procedure. It is suggested that $\varphi = 0.1$ can balance efficiency with protection against data anomalies (Zhong & Yuan, 2011).

Robust procedures are advantageous when analyzing data with missing values for multiple reasons. First, it is difficult to determine the distributional properties of the data when missing values are present. For instance, when there are missing values, observed data may demonstrate non-normality even when the underlying population is normally distributed, or alternatively, the observed data may look normally distributed whereas the true population distribution is non-normal. Additionally, compared to NML, multiple studies demonstrate that robust procedures,

overall, result in less biased, more efficient, and more reliable parameter estimates when data contamination, non-normality, or missing values are observed (Tong et al., 2014; Zhong & Yuan, 2011; Zu & Yuan, 2010). An example of R programming code for the two-stage robust procedure is provided in the Appendix to facilitate the application of the method by substantive researchers.

To better evaluate the performance of the supplemental samples, we compare analyzing the data with either the additional refreshment or replacement supplemental samples to analyzing the original data containing missing values but with no supplemental samples added (NSup), as well as with the original data where cases with missing values have been deleted (listwise deletion).

Evaluation criterion

Multiple outcomes are used to evaluate and compare the use of supplemental samples across the various conditions including absolute average bias, relative efficiency, power, and average confidence interval width. Bias is calculated as the estimation minus the true population parameter value and the absolute average bias is the absolute value of bias averaged across all converged replications. The relative efficiency is the ratio of the squared parameter empirical standard error (i.e., the standard deviation of the estimates across all replications) of complete data to incomplete data ($RE_{\hat{\theta}} = \frac{ESE_{\hat{\theta}, complete}^2}{ESE_{\hat{\theta}, missing}^2} \times 100$). Relative efficiency compares the missing data with the corresponding data without missingness, where a value of 100 indicates that the parameter estimates are as efficient as those for the complete data. Statistical power is the proportion of replications for which the 95% confidence interval does not contain zero. The confidence interval width is calculated by subtracting the lower confidence interval boundary from the upper confidence interval boundary, and the average confidence interval width is the confidence interval width averaged across all converged replications.

Results

This section presents findings for these three approaches: MCAR with refreshment, MAR with refreshment, and MAR with replacement. It is not possible to use the replacement approach when the missing data mechanism is MCAR since the missingness is not related to any variables, and so no auxiliary variables exist to guide the selection of the supplemental sample. The average latent slope parameter is chosen as the estimate of interest based on the presumption that researchers using growth curve models are most often interested in assessing change over

time. Simulation results for the other parameters show the same pattern in evaluating the supplemental samples so those results are omitted in this article but are available in the supplemental materials available on our website: <http://people.virginia.edu/~jm5ku/>. Further, the results for the listwise deletion condition are as expected and in line with past research. Specifically, the listwise deletion condition demonstrated bias when data are MAR and was less efficient than other conditions. The improvement in relative efficiency and statistical power from listwise deletion to the Nsup condition is in magnitude comparable to the improvement from Nsup to the conditions with supplemental samples. Thus, we omit the results from the listwise deletion condition in the rest of the results section. The two-stage robust procedure for SEM with missing data is used to obtain the bias, relative efficiency, statistical power, and confidence interval width for the estimated slope parameter to evaluate the performance of the supplemental samples. In general, findings are similar across the five different non-normal distributions when the appropriate data analysis technique is applied. Thus, the lognormal distribution condition is chosen to demonstrate the results across the other conditions.

Convergence

There are no convergence issues for the vast majority of conditions. When $T = 4$, all model estimations across all conditions and across the 500 replications converged. When $T = 8$, there are no convergence issues if sample size is large. When sample size is small ($N = 50$), out of the 128 conditions, 19 conditions didn't fully converge across the 500 replications due to the high missing proportion (a missing rate of 68% at the last measurement occasion led to only ~16 cases left in the data set). Because these conditions are very extreme and only some replications did not converge across the 500 replications, we believe that convergence is not an issue in most practical situations. Information about the converge rate for each condition is provided in the supplemental materials available on our website: <http://people.virginia.edu/~jm5ku/>.

Absolute average bias

In general, when missingness is MCAR, bias for the parameter estimates is close to zero for refreshment sample approaches and Nsup, and is comparable to the bias when the complete data are analyzed. Results are consistent across all conditions (e.g., different N s, σ_e^2 , missing data rates, and time points). For MAR data, the pattern of bias is strongly influenced by the supplemental sample approach. When a refreshment approach or Nsup is used, bias is low across conditions (e.g., different N s, σ_e^2 , and missing data rates)

and is comparable to that for the complete data. On the other hand, when a replacement approach is used, bias is greater than that for the complete data, and bias grows as the missing data rate and size of the replacement sample increases. Specifically, the large one-time replacement sample and the repeated replacement sample approaches show more bias than the small one-time replacement approach. For example, in the log-normal distribution condition ($N = 1000$, $T = 8$, $\sigma_e^2 = 3$, and $r = .8$), when the missing data rate is .03 the bias is .011 and .099 for the small one-time replacement sample and the large one-time replacement sample, respectively, whereas when the missing data rate is .08, the bias for the small one-time replacement sample and the large one-time replacement sample is .048 and .258, respectively. This pattern is observed across σ_e^2 conditions. The correlation between the auxiliary variable and the latent slope influences bias for the replacement samples as well. That is, a higher correlation (.8) between the auxiliary variable and the latent slope results in considerably higher bias than a lower correlation (.3). For instance, as demonstrated in Fig. 2 ($N = 1000$, $T = 8$, distribution = log-normal, $\sigma_e^2 = 3$, and MR = .08), using repeated replacement samples results in less bias when $r = .3$ (.082) than when $r = .8$ (.253). This is observed across distributions and σ_e^2 conditions.

Relative efficiency

When the missing data mechanism is MCAR, efficiency is greatest when the refreshment approach is used, and increasing the size of the refreshment sample results in higher efficiency. Specifically, the small one-time refreshment sample approach has slightly lower efficiency than the large one-time refreshment sample approach and the repeated refreshment sample approach. When $T = 4$, the large one-time refreshment and the repeated refreshment sample approaches have efficiency comparable to that for the complete data. Nsup is less efficient than the refreshment approach, and efficiency for Nsup decreases slightly when missingness increases. For example, when the missing data rate is .15 ($N = 1000$, $T = 4$, distribution = log-normal, and $\sigma_e^2 = 3$) the relative efficiency is 103 for the repeated refreshment sample approach, 98 for the one-time refreshment sample approach, and 91 for Nsup. The pattern when there are more timepoints is similar. When $T = 8$, the large one-time refreshment sample approach and repeated refreshment sample approach are most efficient, more efficient than complete data, and efficiency increases slightly as missingness increases. Nsup and the small one-time refreshment sample approach are second in efficiency and are comparable to the complete data. The pattern for efficiency in the MCAR conditions is similar across the other conditions (e.g., different N s, σ_e^2 , time points,

Fig. 2 Absolute average bias for the population mean slope parameter ($N = 1000$, $T = 8$, distribution = lognormal, $\sigma_e^2 = 3$, and $MR = .08$). MCAR = missing completely at random; MAR = missing at random; RF = refreshment supplemental sample; RP = replacement supplemental sample; (1) = $1 \times$ the missing number from Time 1 to Time 2 added

at Time 3; (T-2) = $6 \times$ the missing number from Time 1 to Time 2 added at Time 3; (M) = $1 \times$ the missing number from Time 1 to Time 2 added at Time 3, 4, 5, 6, 7, and 8; r = correlation between the auxiliary variable and latent slope

and distributions). As demonstrated in Fig. 3, the pattern of efficiency is similar for the MAR condition; the replacement approach demonstrates a comparable pattern as the refreshment approach.

Statistical power

Across all datasets, statistical power increases as sample size increases. When data are MCAR, the refreshment

Fig. 3 Relative efficiency for the population mean slope parameter ($N = 1000$, $T = 8$, distribution = lognormal, $\sigma_e^2 = 3$, and $MR = .08$). MCAR = missing completely at random; MAR = missing at random; RF = refreshment supplemental sample; RP = replacement supplemental sample; (1) = $1 \times$ the missing number from Time 1 to Time 2 added

at Time 3; (T-2) = $6 \times$ the missing number from Time 1 to Time 2 added at Time 3; (M) = $1 \times$ the missing number from Time 1 to Time 2 added at Time 3, 4, 5, 6, 7, and 8; r = correlation between the auxiliary variable and latent slope

sample approach consistently demonstrates high power, comparable to that for the complete data. This pattern is observed across missing data rate conditions. When $T = 4$, power for NSup is generally similar to the refreshment sample approach and complete data, but in the highest missing data rate condition (.15) power for NSup is slightly lower than the refreshment approach and complete data. For instance, when the missing data rate is .15 ($N = 1000$, $T = 4$, distribution = log-normal, and $\sigma_e^2 = 3$), refreshment sample approaches demonstrate comparable power to complete data (.902) (small one-time refreshment sample = .894 and large one-time refreshment sample = .910) whereas NSup produces slightly lower power (.874). When $T = 8$, NSup and the small one-time refreshment sample approach demonstrate power comparable to complete data whereas the repeated refreshment sample and large one-time refreshment sample approaches generally have slightly more power.

In the MAR conditions, overall, using replacement samples produces greater power than refreshment samples, and using any supplemental sample results in greater power than NSup. In general, the repeated replacement sample and the large one-time replacement sample approaches have slightly higher power than complete data whereas other supplemental sample approaches demonstrate comparable or slightly lower power than the complete data. For example, as shown in Fig. 4 ($N = 1000$, $T = 8$, distribution = log-normal, $\sigma_e^2 = 3$, $r = .8$, and $MR = .08$), the large one-time

replacement sample approach produces greater power (.998) than large one-time refreshment sample approach (.930), and both supplemental sample approaches have greater power than NSup (.908). However, we would like to note that the power values for data with replacement samples should not be considered as the observed high power is caused by the large bias produced by using replacement samples. That is, when bias is large the confidence interval is less likely to contain zero compared to when the bias is small. Generally, power was not affected by distribution or σ_e^2 .

Average confidence interval width

With MCAR data, average confidence interval widths for the estimated average slope parameter for data with missing values are generally similar to that for the complete data. When $T = 4$, in the highest missing data rate condition (.15), the NSup condition has a slightly larger width than refreshment sample approaches, meaning that the refreshment sample approaches lead to more precise parameter estimates. When $T = 8$, the large one-time refreshment and repeated refreshment sample approaches have slightly smaller widths than the small one-time refreshment sample approach and NSup. When data are MAR, average confidence interval widths are similar across conditions when missing data rates are low, but differences between conditions grow as the missingness increases. The large

Fig. 4 Statistical power for the population mean slope parameter ($N = 1000$, $T = 8$, distribution = lognormal, $\sigma_e^2 = 3$, and $MR = .08$). MCAR = missing completely at random; MAR = missing at random; RF = refreshment supplemental sample; RP = replacement supplemental sample; (1) = $1 \times$ the missing number from Time 1 to Time 2 added

at Time 3; (T-2) = $6 \times$ the missing number from Time 1 to Time 2 added at Time 3; (M) = $1 \times$ the missing number from Time 1 to Time 2 added at Time 3, 4, 5, 6, 7, and 8; r = correlation between the auxiliary variable and latent slope

one-time refreshment, large one-time replacement, repeated refreshment, and repeated replacement sample approaches have similar widths to that for the complete data and stay consistent across missing data rates. The small one-time refreshment and one-time replacement sample approaches, and NSup have slightly greater widths, indicating that these approaches have lower precision in parameter estimates compared to the larger supplemental conditions. As Fig. 5 demonstrates ($N = 1000$, $T = 8$, distribution = log-normal, $\sigma_e^2 = 3$, $r = .8$, and $MR = .08$), the average confidence interval width for the small one-time refreshment and one-time replacement sample approaches is .131 whereas the average width ranges from .110–.111 for the large one-time refreshment, large one-time replacement, repeated refreshment, and repeated replacement sample approaches. In general, widths slightly increase as the missing data rate increases.

An analytical interpretation of the results

Let $\mathbf{y}$ represent the population of T random variables with $E(\mathbf{y}) = \boldsymbol{\mu}$ and $Cov(\mathbf{y}) = \boldsymbol{\Sigma}$. The manifest variables y_i , $i = 1, 2, \dots, N$, constitute a sample from $\mathbf{y}$ and contain missing values. The mean vector and covariance matrix corresponding to the observations in $\mathbf{y}_i$ are denoted as $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, respectively.

We use the two-stage robust procedure to estimate the growth curve model. The first stage of the robust procedure is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by solving estimating equations. The estimating equations defining the robust M-estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by

$$\sum_{i=1}^N \omega_{i1}(d_i) \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\mu}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

$$\sum_{i=1}^N \frac{\partial \text{vec}'(\boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\sigma}} \mathbf{W}_i \text{vec} \left[\omega_{i2}(d_i) (\mathbf{y}_i - \boldsymbol{\mu}_i) (\mathbf{y}_i - \boldsymbol{\mu}_i)' - \omega_{i3}(d_i) \boldsymbol{\Sigma}_i \right] = 0,$$

where d_i is the Mahalanobis distance, defined by

$$d_i^2 = d^2(\mathbf{y}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i),$$

$\omega_{i1}(d_i)$, $\omega_{i2}(d_i)$, and $\omega_{i3}(d_i)$ are non-increasing weight functions of d_i , $\mathbf{W}_i = 0.5(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})$ with $\otimes$ being the Kronecker product, $\text{vec}(\cdot)$ is the operator that transforms a matrix into a vector by stacking the columns of the matrix, and $\boldsymbol{\sigma} = \text{vec}(\boldsymbol{\Sigma})$ is the vector of stacking the columns of the lower triangular part of $\boldsymbol{\Sigma}$.

In the second stage of the robust method, we fit $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ by the growth curve model under the structural equation modeling framework. Let $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the

Fig. 5 Average confidence interval width for the population mean slope parameter ($N = 1000$, $T = 8$, distribution = lognormal, $\sigma_e^2 = 3$, and $MR = .08$). MCAR = missing completely at random; MAR = missing at random; RF = refreshment supplemental sample; RP = replacement supplemental sample; (1) = $1 \times$ the missing number from

Time 1 to Time 2 added at Time 3; (T-2) = $6 \times$ the missing number from Time 1 to Time 2 added at Time 3; (M) = $1 \times$ the missing number from Time 1 to Time 2 added at Time 3, 4, 5, 6, 7, and 8; r = correlation between the auxiliary variable and latent slope

structural model satisfying $\mu = \mu(\theta)$ and $\Sigma = \Sigma(\theta)$, where θ represents all the free parameters in the model. The estimates $\hat{\theta}$ are obtained by minimizing

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

The above procedure will result in consistent parameter estimates if the sample y_i , $i = 1, 2, \dots, N$, is random selected from the population y . When there is no supplemental samples and when refreshment samples are used, the random sampling assumption is satisfied. Therefore, as shown in the simulation results, these approaches lead to slightly biased or even unbiased parameter estimates. However, replacement samples are not selected based on the random sampling procedure. Instead, they follow truncated distributions of the population distribution because they are selected to replace the missing data which corresponds to certain values in the auxiliary variable. When replacement samples are added to the original sample, the overall population distribution becomes a mixture of the original population distribution and the truncated distribution, with the mixing proportions being the proportions of the original sample and the replacement sample, respectively. As a result, the estimated μ and Σ in the first stage of the robust 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. When the size of the replacement samples increases, the mixing proportion of the truncated distribution increases too. Then, the data distribution deviates more from the original population distribution, leading to more biased parameter estimates.

Note that although we analytically interpret the pattern of the results based on the two-stage robust method, we expect to see the same pattern for 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 how big the influence of adding replacement samples could be in empirical data analyses.

Discussion

Adding supplemental samples is a widely used technique to address attrition in longitudinal research. Yet, little research has investigated how supplemental samples affect model estimation. Thus, our study evaluated the benefits and limitations of using supplemental samples across a range of conditions commonly encountered by researchers with the goal of providing methodological recommendations for

how supplemental samples may be best incorporated into longitudinal designs.

To gain a better understanding of the effects of using supplemental samples, we compared refreshment and replacement approaches to adding supplemental samples and compared those with complete data (without missing values), NSup (original data with no supplemental sample added) and listwise deletion (original data but cases contain missing values are deleted). Bias, relative efficiency, statistical power, and average confidence interval width were used as evaluation criteria. Our findings related to NSup and listwise deletion were as expected. Listwise deletion produced biased estimates when data were MAR, and lacked power and efficiency in all conditions. Whereas, NSup produced estimates similar to complete data under both MCAR and MAR, and acceptable power and efficiency. Conversely, our findings related to supplemental samples provide insight into the effects of adding supplemental samples in longitudinal designs. The use of refreshment samples resulted in low levels of bias that were similar to complete data and NSup. Whereas using replacement samples led to biased estimates, and this bias grew as both the size of the replacement sample and the missing rate increased. The positive impact of adding refreshment samples was primarily observed with regard to relative efficiency and power. Indeed, the addition of refreshment samples often resulted in increased efficiency and power, and these benefits increased as the size of the refreshment sample increased. For instance, in the four time point log-normal distribution condition ($MR = .15$), the gain in power using a refreshment sample $2 \times$ the missing number from Time 1 to Time 2 is similar to the gain in power moving from listwise deletion to NSup. Similarly, the parameter estimate is as efficient as it would have been with complete data when adding the refreshment sample whereas the estimate is 91 and 61% as efficient with NSup and listwise deletion, respectively. A comparable effect was observed with regards to the confidence interval width: use of large refreshment samples resulted in decreased confidence interval width and thus more precise estimations.

When appropriate data analysis techniques such as the two-stage robust procedure were used to deal with non-normality,⁴ the effects of adding supplemental samples were

⁴Note that we used the two-stage robust procedure for the model estimation because the two-stage procedure has been proven to outperform ML based methods when data contain missing values and are non-normally distributed (Tong et al., 2014). ML based methods (even with robust standard errors or test statistics) may lead to biased parameter estimates. However, the pattern of adding supplemental samples based on ML methods is similar to the pattern based on the robust methods, so we can draw the same conclusion. We did not present the results from ML methods in this article because we do not suggest those methods when the non-normality is suspected. But those results are available on our website.

not affected by different distributions of the data. Furthermore, the type of attrition (permanent versus intermittent) does not appear to affect the performance of supplemental samples. In addition, sample size and variance of measurement errors did not have a noticeable influence on the pattern of results. One exception being the combination of low sample size ($N = 50$) and high missing rate (.08 or .15) which sometimes produced unpredictable results. The correlation between the auxiliary variable and the latent slope influenced the bias of the supplemental samples. Specifically, using replacement samples resulted in greater bias in the higher correlation condition. The effect of the missing data rate was most evident when comparing across methods in that the discrepancy in evaluation criterion between supplemental sample approaches and NSup grew with greater missingness. Further, the missing data rate influenced specific outcomes—when using a replacement approach, the bias grew as the missing data rate increased. Additionally, efficiency decreased as the missing data rate grew when using most methods, but, unexpectedly, increased when using large supplemental samples. We also examined the utility of gradually adding supplemental samples at multiple time points compared to adding a larger sample at one occasion. Findings suggest that the timing of adding supplemental samples (i.e., at one occasion compared to over multiple occasions) does not influence model estimation, rather it is the overall size of the sample that affects results. Indeed, the benefits of utilizing supplemental samples are magnified as the size of the sample increases.

In sum, although using replacement samples increase power and efficiency, they produce biased estimates. Thus, replacement samples should not be used. However, refreshment samples appear to be a viable solution to increasing statistical power and efficiency in longitudinal studies. The decision to use refreshment samples depends on many factors such as the expected effect size, missing data rate, and cost or difficulty of obtaining the supplemental sample. For example, if the attrition rate is expected to be relatively low and there is great cost or difficulty associated with recruiting a refreshment sample, researchers may want to consider mainly using other strategies to handle missing data (e.g., retention and tracking techniques). In contrast, if the expected effect size is relatively small and researchers have the resources to collect data from a refreshment sample, this could increase power and improve researchers' chances of correctly rejecting a null hypothesis.

Our simulation results are consistent with findings from Taylor et al. (2018). Furthermore, although conceptually attempting to replace participants lost in attrition may make sense, methodologically, the simulation results are reasonable. The two-stage robust procedure for SEM uses

all available data to estimate the parameters; missing values are dealt with during the model-fitting process. Thus, the replacement sample is not truly replacing participants whose data were missing, rather the replacement sample is adding additional participants with specific characteristics resulting in an unrepresentative sample of the population.

We would like to note that although only one type of model (i.e., linear growth curve model) was considered in this study, we believe that supplemental samples will perform similarly with other models used in longitudinal designs. Further, only MCAR and MAR missing data mechanisms were considered. Researchers may wish to investigate how using supplemental samples affect parameter estimates when data are MNAR as this may be the missing data mechanism present in some longitudinal studies. Moreover, previous research has shown that supplemental samples can provide additional information regarding the attrition process (i.e., the missing data mechanism; Bhattacharya, 2008; Deng, Hillygus, Reiter, Si, & Zheng, 2013; Hirano, Imbens, Ridder, & Rubin, 2001). Thus, supplemental samples may be utilized to distinguish MAR data from MNAR data in growth curve models. We also would like to point out that replacement samples have been used in previous studies which may have led to biased parameter estimates and misleading statistical conclusions, however, researchers may be able to correct this bias using bias correction techniques (e.g., Du, Liu, & Wang, 2017). Unfortunately, as discussed in Stolzenberg and Relles (1997), no technique or combination of techniques offer universal rescue to the selection bias problem. Although adding sampling weights to the replacement samples could be an easy solution, different weighting schemes may be adopted and are currently not guaranteed to be exact (Cortes, Mohri, Riley, & Rostamizadeh, 2008). Bootstrapping methods may lead to more accurate estimations than adding sampling weights. As such, future research should investigate how researchers who used a replacement approach can correct the bias problem after data collection is complete.

In sum, these findings deepen our understanding of the effects of supplemental samples on parameter estimates in longitudinal designs. As non-normal data are a better representation of real-world data, our results better replicate the outcomes of using supplemental samples in applied work. Further, this study extends previous research by demonstrating how the performance of supplemental samples is affected by various situations often found when conducting research in the social and behavioral sciences. Results from this study may be utilized by researchers considering using supplemental samples and provide guidance for selecting an appropriate supplemental sample approach.

Appendix

R code for the two-stage robust procedure for a linear growth model with a time-invariant covariate

```
install.packages('rsem')
library(rsem)

dat <- read.table(file = "data.txt",
header = TRUE)
N <- nrow(dat)

# Specify model: A linear growth model
with a time-invariant covariate
gcmmodel <- 'i =~ 1*V1 + 1*V2 + 1*V3 + 1*V4
              +1 *V5 + 1*V6 + 1*V7 + 1*V8
              s =~ 0*V1 + 1*V2 + 2*V3 + 3*V4
              + 4*V5 + 5*V6 + 6*V7 + 7*V8
              i ~ Cov
              s ~ Cov'

# Obtain missing data pattern
pat <- rsem.pattern(dat)

# Proportion of data to be down-weighted
phi <- 0.1

# Obtain robust mean and covariance matrix
musig <- rsem.emmusig(pat, varphi=phi, max.it
= 100000)

# Run model
res.lavaan <- growth(gcmmodel,
                    sample.cov=musig$sigma,
                    sample.mean=musig$mu,
                    sample.nobs=N,
                    mimic='EQS')

summary(res.lavaan)
```

References

- Acocck, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, 67(4), 1012–1028.
- Adler, G. S. (1994). A profile of the Medicare Current Beneficiary Survey. *Health Care Financing Review*, 15(4), 153.
- Beltrán-Sánchez, H., Drumond-Andrade, F. C., & Riosmena, F. (2015). Contribution of socioeconomic factors and health care access to the awareness and treatment of diabetes and hypertension among older Mexican adults. *Salud Pública de México*, 57, s06–s14.
- 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.
- Bhattacharya, D. (2008). Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics*, 144(2), 430–446.
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbacher, J., Malter, F., & Zuber, S. (2013). Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4), 992–1001.
- Casswell, S., Meier, P., MacKintosh, A. M., Brown, A., Hastings, G., Thamarangsi, T., & You, R. Q. (2012). The International Alcohol Control (IAC) Study - evaluating the impact of alcohol policies. *Alcoholism: Clinical and Experimental Research*, 36(8), 1462–1467.
- Centers for Medicare Medicaid Services (2012). Medicare Current Beneficiary Survey (MCBS) bibliography. Retrieved from <https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/Bibliography.html>
- Chatfield, M. D., Brayne, C. E., & Matthews, F. E. (2005). A systematic literature review of 749 attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies. *Journal of Clinical Epidemiology*, 58(1), 13–19.
- Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., & Zheng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28(2), 238–256.
- Du, H., Liu, F., & Wang, L. (2017). A Bayesian “fill in” method for correcting for publication bias in meta-analysis. Manuscript submitted for publication.
- Edwards, A. B., Cherry, R. L., & Peterson, J. (2000). Predictors of misconceptions of Alzheimer's disease among community dwelling elderly. *American Journal of 759 Alzheimer's Disease*, 15(1), 27–35.
- ELSA (2016). Research outputs. Retrieved from. <http://www.elsa-project.ac.uk/publications/case/related>
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430–457.
- Feiring, C., Simon, V. A., & Cleland, C. M. (2009). Childhood sexual abuse, stigmatization, internalizing symptoms, and the development of sexual difficulties and dating aggression. *Journal of Consulting and Clinical psychology*, 77(1), 127.
- Fillit, H. M., Guterman, E. M., & Brooks, R. L. (2000). Impact of donepezil on caregiving burden for patients with Alzheimer's disease. *International Psychogeriatrics*, 12(3), 389–401.
- Fong, G. T., Cummings, K. M., Borland, R., Hastings, G., Hyland, A., Giovino, G. A., & Thompson, M. E. (2006). The conceptual framework of the International Tobacco Control (ITC) policy evaluation project. *Tobacco Control*, 15(3), iii3–iii11.
- Garnier-Villareal, M., Rhemtulla, M., & Little, T. D. (2014). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development*, 38(5), 411–422.
- Goemans, A., van Geel, M., & Vedder, P. (2015). Over three decades of longitudinal research on the development of foster children: A meta-analysis. *Child Abuse Neglect*, 42, 121–134.
- 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.
- Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in the analysis of change. In Collins, L. M., & Sayer A.G. (Eds.) *New methods for the analysis of change* (pp. 335–353). Washington: American Psychological Association.
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11(4), 323.

- Harel, O., Stratton, J., & Aseltine, R. (2015). Designed missingness to better estimate efficacy of behavioral studies-application to suicide prevention trials. *Journal of Medical Statistics and Informatics*, 3(1), 2.
- Hirano, K., Imbens, G. W., Ridder, G., & Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6), 1645–1659.
- Hobden, K., Forney, J. C., Durham, W. K., & Toro, P. (2011). Limiting attrition in longitudinal research on homeless adolescents: What works best *Journal of Community Psychology*, 39(4), 443–451.
- IAC Study (2016). Publications. Retrieved from http://www.iacstudy.org/?page_id=23
- ITC Project (2015). ITC results. Retrieved from http://www.itcproject.org/itc_results
- 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, 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.
- Krause, N., Liang, J., Shaw, B. A., Sugisawa, H., Kim, H.-K., & Sugihara, Y. (2002). Religion, death of a loved one, and hypertension among older adults in Japan. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 57b(2), S96–S107.
- 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.
- Kuhns, L. M., Vazquez, R., & Ramirez-Valles, J. (2008). Researching special populations: Retention of Latino gay and bisexual men and transgender persons in longitudinal health research. *Health Education Research*, 23(5), 814–825.
- Liao, C. C., Li, C. R., Lee, S. H., Liao, W. C., Liao, M. Y., Lin, J., & Lee, M. C. (2015). Social support and mortality among the aged people with major diseases or ADL disabilities in Taiwan: A national study. *Archives of Gerontology and Geriatrics*, 60(2), 317–321.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*, (2nd ed.). Wiley-Interscience: Hoboken, N.J.
- Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, 7(4), 199–204.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166.
- Mistler, S. A., & Enders, C. K. (2012). Planned missing data designs for developmental research. In Laursen, B., Little, T. D., & Card, N. A. (Eds.) *Handbook of developmental research methods* (pp. 742–754). New York: Guilford Press.
- Nemes, S., Wish, E., Wraight, B., & Messina, N. (2002). Correlates of treatment follow-up difficulty. *Substance Use Misuse*, 37(1), 19–45.
- Nicklas, T. A. (1995). Dietary studies of children: The Bogalusa Heart Study experience. *Journal of the American Dietetic Association*, 95(10), 1127–1133.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525–556.
- R Core Team (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. <https://www.R-project.org/>
- 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.
- Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, 19(1), 1–25.
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177.
- SHARE (2016). Publications. Retrieved from <http://www.share-project.org/publications.html>
- Stephens, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: The English Longitudinal Study of Ageing. *International Journal of Epidemiology*, 42(6), 1640–1648.
- Strickler, D. C., Whitley, R., Becker, D. R., & Drake, R. E. (2009). First-person accounts of long-term employment activity among people with dual diagnosis. *Psychiatric Rehabilitation Journal*, 32(4), 261.
- Taylor, L. K., Merrilees, C. E., Goeke-Morey, M. C., Shirlow, P., & Cummings, E. M. (2016). Trajectories of adolescent aggression and family cohesion: The potential to perpetuate or ameliorate political conflict. *Journal of Clinical Child and Adolescent Psychology*, 45(2), 114–28.
- Taylor, L. K., Tong, X., & Maxwell, S. E. (2018). *Evaluating supplemental samples in longitudinal research: Replacement and refreshment approaches*. Manuscript submitted for publication.
- Taylor, M. G., & Lynch, S. M. (2011). Cohort differences and chronic disease profiles of differential disability trajectories. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(6), 729–738.
- Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 631–642.
- Tong, X., & Zhang, Z. (2012). Diagnostics of robust growth curve modeling using Student's *t* distribution. *Multivariate Behavioral Research*, 47, 493–518.
- Tong, X., Zhang, Z., & Yuan, K.-H. (2014). Evaluation of test statistics for robust structural equation modeling with nonnormal missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 553–565.
- Tubman, J. G., Windle, M., & Windle, R. C. (1996). The onset and cross-temporal patterning of sexual intercourse in middle adolescence: Prospective relations with behavioral and emotional problems. *Child Development*, 67(2), 327–343.
- Windle, M., & Windle, R. C. (2001). Depressive symptoms and cigarette smoking among middle adolescents: Prospective associations and intrapersonal and interpersonal influences. *Journal of Consulting and Clinical Psychology*, 69(2), 215.
- Yuan, K.-H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, 77, 803–826.
- Zhong, X., & Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. *Multivariate Behavioral Research*, 46, 229–265.
- Zu, J., & Yuan, K.-H. (2010). Local influence and robust procedures for mediation analysis. *Multivariate Behavioral Research*, 45, 1–44.