Planned Missingness: A Sheep in Wolf's Clothing

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Charlene Zhang

IN PARTIAL FULTILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisors:
Paul R. Sackett
Nathan R. Kuncel

June, 2021

Acknowledgements

Four years ago, my entire knowledge about industrial/organizational (I/O) psychology could be summed up by the first google search page of the term. But now I have become a somewhat competent I/O psychologist, doing work that I love and am proud to do. For this, I have to first and foremost thank my advisors, Paul Sackett and Nathan Kuncel. Getting distracted by the albino squirrel outside of Elliott Hall during a College Board meeting with both of you will forever be one of my fondest memories during graduate school. Paul, I could not have asked for a better teacher and mentor. Every time I dropped by your office (often unannounced), I walked away with newfound knowledge, fresh perspectives, and immense support. You showed me what it is like to have accomplished so much as a scholar and still remain humble and curious, but also when to take pride in and stand up for your own work. I am deeply grateful to have had your guidance and confidence, and I look forward to our continued collaboration in the future. Nathan, your high expectations and affinity for asking the most difficult questions terrified me at first but have proved to be instrumental in my growth as a researcher. Your guidance through the two-year excruciating process of my very first publication will always hold a special place in my heart. Lastly, you win the Cupid award for giving me and Martin the opportunity to meet so far away from both of our homes.

I also must thank my committee members. Aaron Schmidt, Michael Cullen, and Nate Helwig. Aaron, thank you for always providing helpful feedback in seminars as well as on my dissertation. Also thanks for coming to my office to check on me upon hearing my screaming after Tetsuhiro Yamada scared me with the Michael Myers theme song (Tetsu, thanks for that). Michael, as I have told you many times before, working with you for the med school has been formative in transforming me from an academic into an applied researcher. You have been a model scientist-practitioner that I really look up to, and I will always be deeply grateful for the opportunity to have worked with you. Nate, thank you for pointing out that I was doing my simulation wrong when I still had time to fix it. You brought the quantitative expertise and perspective that I really appreciated. I also wanted to thank Richard Landers for giving the data science course and generously answering my many questions. My code %>% fundamentally changed.

I thank my boyfriend, Martin Yu, for being a big part of my support system. Martin, thank you for encouraging me through challenges that seemed impossible at the time, talking some sense into me when I was convinced that my dissertation was tanking, badmouthing reviewers who didn’t like my papers with me, working on our publications on the weekends even though you didn’t want to, and still pretending that you were interested in my dissertation even after hearing about it for the millionth time. My path would have been very different without you.

I am also grateful for my cohort for being the other crucial part of my support system: Dana Tomeh, Elena Auer, Sebastian Marin, Jiayin Qu, Khue Tran, Merrill Levitt, and Elise Anderson. Suffering through the seminars and/or comps together pretty much makes us family, right?! Dana and Elena, our gym-and-rant sessions were my favorite way to cope with stress and I am very thankful that you girls were always there for me. Sebastian, thanks for creating a safe space in our office with me along with Tetsu that was free of judgment and raisins. I miss our movie nights. Jiayin, you are one of the busiest people I have ever met yet you are always so generous with your time and ideas. Khue, I’m still hopeful that we will start our catering business together someday. Merrill, I think of you every time I come upon a study using the Great GATB. Elise, you always light up the room with your energy and I have definitely needed it at times.

I am also grateful to Phoebe Hessen for being a supportive friend, Markus Langer for being the landmark at our gatherings, Yagizhan Yazer for being my AAI bus buddy, Yilei Wang for being my Amazon virtual buddy, Sarah Tian for being my RAND buddy (with a year of lag), Brittany Marcus-Blank for introducing me to MPPAW, the Shewachs (Ori, Katie, Laila, and Holly) for being great hosts and convincing me to come to the program, Jeff Dahlke for being the unrelenting reminder that I was not a doctor yet, Geoff Lamb for feeding me sushi early on so I could pass my comps, Melissa Sharpe for always giving Arsinoe some camera time on our Zoom calls for my sake, and Sander Quint for having an endless number of stories.

On my path to becoming a practitioner, I have had wonderful mentors that guided me in one way or another. Thank you to Ron and Nate Page, Jinyan Fan, George Paajanen, Sean Robson, Maria Lytell, Bobby Naemi, Kathy Stewart, Kayo Sady, and Anne Hansen for the opportunities that you have given me and the wisdom that you have imparted on me.

Last but not least, I am grateful for my parents Hongjun Zhang and Hong Chen. Mom and dad, you have always encouraged me to pursue my passion and work hard, and provided me with the best you could. Although I still fail to fully explain to you what it is that I do, you have always believed in me and unconditionally supported my endeavors. I know that you are proud to call me Dr. Zhang, but I could not have done it without you two.

Abstract

There has been an extensive body of methodological literature supporting the effectiveness of planned missingness (PM) designs for reducing survey length. However, in industrial/organizational (I/O) psychology, it is still rarely applied. Instead, when there is a need to reduce survey length, the standard practice is to either reduce the number of constructs measured or to use short forms rather than full measures. The former is obviously unideal. The latter requires prioritizing the measurement of some items over that of others and can also quickly become time and labor intensive, as not all measures have established short forms.

This dissertation presents three studies that compare the relatively unused methodology of PM against the common practice of using short forms. First, the two approaches are compared in three archival datasets, finding that PM consistently yields more accurate correlational estimates than short forms. Second, a Monte Carlo simulation is conducted to explore how this comparison may be affected by data characteristics, including the number of constructs, construct intercorrelations, sample size, amount of missingness, as well as different types of short forms. Average of all conditions simulated, short forms produce slightly more accurate estimates than PM when empirically developed short forms are readily available for use. When a part of the sample needs to be used to first develop short forms, the two approaches perform equivalently. When the selection of items for short forms strays from being purely empirical, PM outperforms short forms. Lastly, a qualitative survey exploring social science researchers’ knowledge about PM finds that most are not familiar with PM or have an inaccurate understanding of the concept despite working with surveys frequently. A number of research contexts are identified for which PM may not be suitable.

Overall, the findings of this dissertation demonstrate that PM designs are technically effective in producing accurate estimates. Its effectiveness, along with its convenience, makes it a valuable survey design tool. It is apparent that the road to popularizing this technique within the I/O field will require much education in its understanding and application, and this dissertation serves as a first step in doing so.

Table of Contents

List of Tables

List of Figures

Overview

Surveys are ubiquitously used for research in Industrial/Organizational (I/O) psychology and in the social sciences broadly. As a study methodology, surveys are easy to implement and very useful for gathering self-report information. Typically, researchers collect responses on numerous psychological constructs, each measured by multiple items. The measurement of many multi-item constructs can result in an undesirably long estimated response time. Facing the need to reduce study length, common approaches include prioritizing constructs that are most central to the research questions of interest and dropping the others, and using shorter measurement scales. I conduct a series of studies that support an alternative solution, namely planned missingness, a survey design technique that has the potential to significantly lessen the burden of both researchers designing the study and respondents completing the study. However, before introducing the concept of planned missingness, the general missing data problem warrants a discussion.

Missingness in data is common in survey research. Unplanned missingness or survey nonresponse can occur because of inattentive responding, software malfunction, attrition, or intentional decision to skip items (Newman, 2014). In most cases, data missingness poses an inconvenience for data analysis as most data analytic procedures are not suited for treatment with missing data (Schafer & Graham, 2002).

Different Types of Data Missingness

Missingness in a dataset is of concern to researchers and statisticians because it can have an impact on estimates of population parameters. A parameter is biased when it is systematically over- or under-estimated. The effect of missingness on bias is influenced by three factors, such that

$$B_{mis} = f\,(R,\ M,\ T)$$

where *R* is the response rate, or the amount of missing data in a dataset; *M* is the missingness mechanism, or whether missingness is not at random, at random, or completely at random; and *T* is the treatment used for missing data (Newman, 2014; Newman & Cottrell, 2015).

Missingness can be classified according to level of missingness, causal patterns, or statistical relations. Level of missingness distinguishes missing data at the item-level, construct-level, and person-level (Newman, 2014). This manner of classifying missingness is straightforward, as a simple look at the data would suffice. When a respondent fails to answer one or more but not all of the items within a scale, missingness is at the item-level. When one fails to answer all items of a scale, missingness is at the construct-level. When one fails to respond to all constructs of a survey, missingness is at the person-level. Thus, item-level missingness is nested under construct-level missingness, which is nested under person-level missingness (Newman, 2014).

Depending on the underlying causes of missingness, missing data can take a few different forms with regards the way they are distributed throughout a dataset (Enders, 2010; Schafer & Graham, 2002). A univariate pattern occurs when there is some missingness in only one variable or one set of variables, while the rest is completely observed. Enders (2010) has made the distinction between univariate pattern, referring to missingness in only one variable, and unit nonresponse pattern, referring to uniform missingness in a set of variables. For example, if respondents are instructed to skip items two to four if they answer “no” on item one, and there is no additional, unexpected missingness, this would constitute a univariate pattern of missingness. If missingness on one item or a set of items automatically lead to subsequent missingness on the next item or set of items, a monotone pattern of missingness emerges. An example would be attrition in longitudinal data where respondents do not reappear after missing one timepoint of data collection. Lastly, an arbitrary pattern (or sometimes called general pattern) occurs when any value of any variable could be missing. This is more common when there is inattentive responding. While the classification of missingness data patterns as univariate, monotone, arbitrary is useful in helping us understand the contextual causes for missingness, a comprehensive and accurate understanding of how missing data came about is not always feasible. Furthermore, when determining the appropriate treatment for missing data, the “distribution of missingness” or “the probabilities of missingness” is of more critical interest (Schafer & Graham, 2002, p. 151). This distribution of missingness can be represented with a matrix with the same dimensions as the dataset, constituting of elements of 1 or 0 for each value to denote whether the value is missing. This missingness matrix will be denoted as R going forward, following the convention of Schafer and Graham (2002). In other words, the statistical relationship between whether a data point is missing and the value of that data point will have important implications for how we should deal with missingness.

Rubin (1976) classified missingness in terms of how it relates to the data itself into missingness at random (MAR), missingness not at random (MNAR), and missingness completely at random (MCAR). As denoted by Schafer and Graham (2002), if the complete dataset is denoted as Y_{com}, part of which may be missing (Y_{mis}) and the rest is observed (Y_{obs}), then

$$Y_{\text{com}} = (Y_{\text{mis}}, Y_{\text{obs}})$$

such that the would-be complete dataset can be partitioned into the observed parts of the data and the not observed or missing parts of the data. Missingness is classified as MAR when the probabilities of missingness does not depend on the missing data but does depend on the observed data, such that

$$P(R|Y_{\text{com}}) = P(R|Y_{\text{obs}})$$

In other words, patterns of missingness (whether a datapoint is missing or observed) is associated with at least some of the observed data. Therefore, despite being named random, MAR is actually nonrandom missingness. A common example of this is in longitudinal data collection. If the first survey collects course grades of first-year college students, and the follow-up survey a year later collects grades of their second-year courses, whether second-year course grades are missing or observed would be dependent on their first-year course grades (whether they passed first-year courses). This constitutes a case of MAR.

Another interesting example in the industrial/organizational (I/O) psychology literature is direct range restriction, also known as Thorndike's Case 2 (Sackett & Yang, 2000; Thorndike, 1949). In predictive validity studies, a predictor (e.g., cognitive ability) is used to evaluate applicants and make selection decisions. Then a criterion of interest (e.g., job performance) is measured to assess the predictive validity of the predictor. Thus, only criterion data of individuals selected are available, resulting in the range restriction problem (Sackett & Yang, 2000). Traditionally, the attenuated validity due to range restriction is corrected upwards to estimate the true value. However, range restriction can also be viewed a missing data problem (Mendoza, 1993; Pfaffel et al., 2016; Wiberg & Sundström, 2009). Whether a criterion datapoint is missing is dependent on the observed predictor scores, constituting a case of MAR mechanism.

More stringent conditions define MCAR such that the probabilities of missingness do not depend on the missing data or the observed data. In other words, probabilities of missingness are truly random such that

$$P(R|Y_{\text{com}}) = P(R)$$

MCAR can occur when respondents accidentally miss questions or the software used to administer questionnaires malfunctions and skips questions. In addition, instances where researchers implement planned missingness for purposes of shortening survey lengths or preventing respondent fatigue, missingness is MCAR, as each respondent is given a randomly chosen subset of items. In these cases, missing data are not expected to be related to either observed data or the data that are missing.

In contrast, when the distribution of R depends on Y_{mis}, it constitutes a case of MNAR. The most common context in which MNAR operates is nonrandom dropout in longitudinal data collection, where the dropout rate is associated with the variable(s) measured.

Importance of Distinguishing between MAR, MCAR, and MNAR

Whether the missingness is MAR, MCAR, or MNAR will have important implications for the extent to which and how the dataset can be used to infer population distribution. When the population distribution is interpreted as an infinitely repeated sampling distribution, the complete sample dataset is associated with a certain probability among all possible samples. On the other hand, any sampled statistics can also be interpreted as the likelihood given what is known about the population parameters from the sample. When data are MNAR, the missing data will not represent a proper sampling distribution or a proper likelihood. It is for this reason that many of the available techniques for treating missing data are not appropriate for cases of MNAR. For example, Collins, Schafer, and Kam (2001) demonstrate that methods for treating missingness that assume MAR and MCAR may yield biased population estimates when used on data with MNAR (Collins et al., 2001).

In cases of MAR or MCAR, however, the distribution of Y_{obs} can be written as a function of the distribution of Y_{com} and Y_{mis}, such that

$$P(Y_{obs}; \theta) = \int P(Y_{com}; \theta)\, dY_{mis}$$ (Rubin, 1976; Schafer & Graham, 2002)

with $P(Y_{com}; \theta)$ denoting the population distribution with unknown parameter θ from which complete data (Y_{com}) are randomly sampled, then the sampling distribution of observed data ($P(Y_{obs}; \theta)$) is the definite integral of $P(Y_{com}; \theta)$ with respect to missing data (Y_{mis}). This equation will yield a proper sampling distribution only when missingness is MCAR but will yield a proper likelihood as long as missingness is MAR (Rubin, 1976). Therefore, different techniques for handling missing data are appropriate for different missingness mechanisms depending on the underlying assumptions. Generally, while distribution-based treatment methods (e.g., listwise and pairwise deletion) only yield valid estimates under MCAR and become biased under MAR, likelihood-based methods (e.g., maximum likelihood) yield unbiased estimates under either MAR or MCAR (Schafer & Graham, 2002).

In situations of MNAR, treatment methods generally require an explicit understanding or modeling of the causes of missingness. It is for this reason that the MNAR mechanism is sometimes also referred to as "non-ignorable," meaning that the non-response model is not ignorable, while MAR and MCAR are "ignorable". It is important to note that while MAR, MCAR, and MNAR are often discussed as discrete categories, their distinctions are nowhere as clear because rarely is naturally-occurring missingness in data purely MAR, MCAR, or MNAR. Other than planned missingness which collects MCAR data by definition, any and all unintended missing data are on "a continuum between MAR and MNAR" (Graham, 2009, p. 567). Even when data are collected with a planned missingness design, additional unintended missingness can occur due to a variety of reasons.

Nevertheless, determining the nature of missingness is important for the techniques used to treat it. It is especially important to make the differentiation between ignorable and non-ignorable missingness as both MAR and MCAR but not MNAR can be appropriately treated with a class of techniques discussed in a later section. Unfortunately, this poses somewhat of a Catch-22. As previously defined, MNAR is present in datasets where missingness is related with the missing data. Sensitivity analyses can be used to statistically suggest missingness nonignorability (e.g., Molenberghs & Verbeke, 2006; Verbeke et al., 2001). When parameters under the assumptions of nonignorability exert little effect on inferences of interest, ignorable and therefore missingness at random may be satisfied (Troxel et al., 2004). However, Molenberghs and colleagues (2008) demonstrate that it is not possible to distinguish MAR and MNAR empirically because MNAR models cannot be verified by observed data. For every MNAR model fit, there is a MAR counterpart model with identical fit. To definitively diagnose MNAR, therefore, the missing values are needed, whose missing status composes the MNAR. Thus, it is important to consider the substantive variables involved in the study and whether there are any theoretical explanations for why missing values might be related to the non-responses. In some cases, such conceptual rationale can be easily established. For example, if a longitudinal study measures the level of depression in participants over time, it is reasonable to postulate that individuals with more severe levels of depression are less likely to respond to surveys. One way to diagnose MNAR is to conduct follow-up surveys to non-respondent. For example, Fielding et al. (2008) set up a reminder system in their collection of quality of life data, which retrieved about 50% of the missing data. The non-missing data were then compared with the originally missing data to establish probable MNAR mechanism. Although useful, this approach is not always possible and even when done, only provides an indicator of potential MNAR.

Tests of MCAR primarily rely on the conceptual property of random missingness, such that cases with missing data and cases without missing data should belong to the same population and therefore have equal means and covariances (Kim & Bentler, 2002; Rubin, 1976). While a series of independent *t* tests can be performed to test each variable, Little (1988) provides a global, multivariate statistical procedure for testing whether the missingness mechanism in a dataset is MCAR. Mean differences are computed across subgroups of respondents that share missingness pattern. The resulting weighted sum of the mean differences is tested for its statistical significance against the null hypothesis that missingness is completely at random.

Techniques for Treating Ignorable Missingness

While researchers may habitually edit data post-hoc (e.g., listwise or pairwise deletion) as a simple and convenient way of removing data missingness, these popular procedures have problems of their own (Newman, 2014; Newman & Cottrell, 2015). Although no method for treating missing data is perfect, some (namely maximum likelihood and multiple imputation) are generally better than others (e.g., Little & Rubin, 2019; Newman, 2014). I briefly review the most common treatment methods below.

Listwise Deletion. In opposition with the "fundamental principle of missing data analysis" (Newman, 2009, p. 11) that unambiguously advises researchers to use all available data, listwise deletion, or otherwise known as complete-case analysis, involves removing all data of participants who are missing any amount of data before conducting statistical analysis. Newman (2014) explicitly prohibits the use of listwise deletion to treat missingness.

The most obvious issue with listwise deletion is that all item-level and construct-level missingness are exacerbated into person-level missingness. Rather than treating missingness, this practice introduces more missingness to the data. By doing so, total sample size and therefore statistical power is reduced. This procedure of throwing away some of the data that have been obtained by using respondents' energy and time can be wasteful and even viewed by some as unethical (Rosenthal, 1994). More seriously, listwise deletion introduces bias into both the sample tested as well as the parameter estimates (Newman, 2014). Because analyses are only conducted on participants who completed the survey in its entirety, such a sample is inherently not representative of the population or even the original sample. Under nonrandom missingness (MAR or

MNAR), listwise deletion yields biased parameter estimates (Arbuckle & Marcoulides, 1996; Enders & Bandalos, 2001; Muthén et al., 1987; Newman, 2014; Wothke, 2000). Given that a strict case of MCAR is so implausible in practical, empirical settings, listwise deletion is generally not recommended.

Pairwise Deletion. Pairwise deletion, also known as available-case analysis, is similar to listwise deletion in the manner that it essentially ignores the missingness of the data. It calculates different parameters with the available data that are not missing for the particular parameter. Thus, the sample size associated with each of the parameter estimates may be different. For example, the mean and standard deviation of each variable are calculated using the observed values for that variable, and the correlation between each pair of variables is calculated based on the number of cases that have observed values for both variables. Additional statistical analyses are then computed based on these pairwise values.

Because each parameter is estimated from a different set of cases using a pairwise deletion approach, it is often difficult to compute estimates of uncertainty like standard errors. Further, parameter estimates generated under missingness mechanisms of MAR and MNAR have been found to be biased (Arbuckle & Marcoulides, 1996; Muthén et al., 1987; Wothke, 2000), although the biases have been found to be small in empirical examples (Graham, 2009; Schafer & Graham, 2002).

Whether pairwise deletion is appropriate depends on the specific pattern of missingness in the data. Newman (2014) summarizes that when there is no construct-level missingness and only item-level missingness, and the proportion of respondents that exhibit some level of item-level missingness is below 10%, pairwise deletion performs just as well as maximum likelihood and multiple imputation procedures practically. Alternatively, in such a case where there is only item-level missingness, each construct is measured by multiple items, and each respondent has reported some data on each construct, each individual's mean across items that measure the same construct can be used to represent that construct (Newman, 2014).

Single Imputation. Single imputation is another technique infamous for being inappropriate at treating missing data (Newman, 2014). Generally, each missing value is filled in by a single fixed value. The specific procedure from which the fixed value is obtained can vary, the simplest of which is mean substitution (Wilks, 1932). The mean of each variable across all respondents can be calculated and used to fill in all missing values within that variable, thus retaining maximum sample size (Schafer & Graham, 2002).

Because single mean imputation or unconditional mean imputation fills in all missing data with a constant, the procedure artificially reduces the variance of the variable and changes its distribution. As a result, multivariate correlation magnitudes are also downwardly biased. Further, when the dataset is MCAR, mean substitution inherently violates the assumption of missing data being random, introducing bias (Newman, 2014). Another major disadvantage of single imputation is that because different parameter estimates are associated with different sample sizes, an accurate standard error usually cannot be calculated for hypothesis testing (Little & Rubin, 2019). Using the complete dataset sample size for partially imputed data results in downwardly biased standard errors and increases the probability of Type I errors (Newman, 2014).

More sophisticated single imputation methods attempt to preserve the distribution of the variables. For example, hot deck imputation, replaces each missing datum with the value from a real respondent who answers similarly on other variables (Ford, 1983); conditional mean imputation fills in missing data with the predicted values calculated from a regression equation (Buck, 1960). However, systematic biases in correlation estimates remain unresolved (Enders, 2010).

Person mean imputation/prorated scale score. Person mean imputation is a more rare type of single imputation, used when there is some data available representing each construct for each respondent, or when missingness exists only at the item-level. Thus, missing values of each respondent are replaced by the average of the available items. While this technique has received less empirical attention, biased parameter estimates, such as coefficient alphas, have been found (Enders, 2003).

More complicated methods of single imputation exist that involve regression modeling and corrections. However, when faced with MAR or MCAR, statisticians unanimously recommend multiple imputation over single imputation (Newman, 2014; Schafer & Graham, 2002).

Multiple Imputation (MI). The advantage of multiple imputation over single imputation lies in the word “multiple.” By using an unbiased single imputation method (e.g., stochastic regression imputation where each predicted value from a regression equation is augmented with a normally distributed residual term and used to fill in a missing value) and performing the imputation routine multiple times, variations across different imputations are taken into consideration when calculating standard errors or other degrees of uncertainty, making hypothesis testing more accurate. Whereas *SE*s are usually downwardly biased in single imputation methods because sample size is inflated by using the complete dataset sample size when in fact some of the data have been imputed, *SE*s in multiple imputation procedures are corrected upwardly by including the between-imputations variance (Newman, 2014; Schafer & Graham, 2002). As such, the pooled MI parameter estimates are unbiased when missingness is either MAR or MCAR (Newman, 2014).

In a typical procedure of multiple imputation, missing data are imputed *m* times. Generally, it is recommended in imputing, all variables that are included in the survey design should be used as predictors, as leaving out variables may lead to biased estimates (Rubin, 1996). Statistical analyses are computed using each of the *m* partially imputed datasets independently (Royston, 2004). Thus, rather than filling in a missing datum with a single value, MI procedures replace the missing value with a set of *m* plausible values, retaining the uncertainty around the datum (Yuan, 2000). Results across the *m* datasets are then pooled, with the overall estimate being simply the average of the *m* estimates, and the uncertainty calculated as a function of both the average within-imputation variance and between-imputations variance using Rubin's (1987) formula, such that

$$S.E. = \sqrt{\frac{1}{M}\sum_{m=1}^{M} S.E._m^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{m=1}^{M}(b_m-\bar{b})^2}$$

Maximum Likelihood (ML). The techniques that have been discussed thus far either attempt to ignore missingness in the dataset (listwise and pairwise deletion) or try to recover the missing data (single and multiple imputation). Maximum likelihood techniques are less conceptually straightforward. Rather than estimating the missing data to model the would-have-been complete dataset, ML procedures select the parameter estimates that maximize the probability of the observed data (Newman, 2014).

Based on the relationships between different values of the parameter estimates and the likelihood or probability of the observed data, a likelihood function is generated. The parameter estimates that maximize the likelihood function based on the partially missing data are then selected. Thus, when missingness is MAR or MCAR, parameter estimates yielded by ML routines will be unbiased and standard errors will be accurate (Dempster et al., 1977; Enders & Bandalos, 2001; Finkbeiner, 1979). Due to the increasing availability of software and different statistical packages designed to allow for ML approaches, including LISREL (Joreskog & Sorbom, 1983), AMOS (Arbuckle, 1995), SPSS, and R, this technique has grown in popularity in recent years.

Two major ML techniques are elaborated below: full-information maximum likelihood (FIML) and expectation maximization (EM).

Full-information maximum likelihood (FIML). FIML aims to identify population parameters that have the highest probability of yielding a particular sample of data by computing a case-wise likelihood function using only items that are observed for each case (respondent) and accumulating and maximizing likelihood functions across all cases (Arbuckle & Marcoulides, 1996; Enders, 2010; Enders & Bandalos, 2001).

Expectation maximization (EM). The EM algorithm divides a maximum likelihood procedure into two steps. In the expectation (E) step, elements in the initial estimates of mean and covariance matrix are used to create a function for the expectation of the log-likelihood. In the maximization (M) step, parameter estimates are generated to maximize the expected log-likelihood function generated in the E-step. The updated parameter estimates are then fed back into the E-step. This iterative process continues until elements no longer change between consecutive iterations, and the likelihood function has converged (Enders, 2010).

Although MI and ML techniques have been repeatedly found to be equivalent (e.g., Collins et al., 2001; Graham et al., 2012), each has practical advantages and disadvantages. MI requires imputing data and performing any statistical analyses multiple times and pooling results. Importantly, it had been suggested that a large number of imputed datasets (*m*) is needed to ensure sufficient power (Graham et al., 2007), making analyses computationally demanding. Further, to a non-technical reader, the process of imputing data can seem dubious. On the other hand, ML consists of complex computational steps that may not be built into existing statistical software. The choice between these two techniques should depend on the resources available and the statistical analyses that researchers wish to perform. For the quantitative components of this dissertation, MI is used to treat missingness incurred by PM.

Techniques for Treating Non-Ignorable Missingness

Handling non-ignorable missingness or MNAR generally requires the simultaneous modeling of both the observed data as well as the missingness processes. For example, Tsonaka et al. (2009) demonstrate the use of a semi-parametric shared parameter model that analyzes longtiduinal responses with nonrandom missingness. It has no distributional requirements for the random effects and therefore does not require sensitivity analyses. When the joint distribution of response and missingness processes needs to be specified, sensitivity analyses are needed. Some recent developments for specifying a distribution for unknown parameters of missingness include selection models (e.g., Schafer & Graham, 2002; Kenward, 1998; Diggle & Kenward, 1994; Little, 1995; Little & Schenker, 1995) and pattern-mixture models (e.g., Andridge & Little, 2011; Hedeker & Gibbons, 1997; Little, 1993; Schafer & Graham, 2002; Thijs, Molenberghs, & Verbeke, 2000).

Planned Missingness

Much of the missingness in collected data occurs unintentionally (e.g., attrition, inattentive responding, software malfunction) and is regarded as at least an inconvenience and sometimes a major challenge. As detailed before, across the different types of missingness mechanism, ignorable missingness can be properly handled with multiple imputation or full-information maximum likelihood. Therefore, it is not so much of a problem but actually poses an opportunity. Planned missingness (PM), also known as designed missingness (e.g., Harel et al., 2015), partial or split questionnaire design (e.g., Houseman & Milton, 2006; Raghunathan & Grizzle, 1995; Wacholder et al., 1994), or matrix sampling (e.g., Munger & Loyd, 1988; Thomas et al., 2006), is a survey design technique that deliberately collects incomplete data with an ignorable missingness pattern. In a PM design, only a proportion of the study items are administered to each respondent such that respondents receive different subsets of items and each item may be answered by a different subset of respondents. In the earliest and simplest PM design, all items in a study are divided into a number of subsets and each respondent is assigned a random subset. These subsets are derived from sampling items without replacement such that each item belongs in one subset and one subset only (Munger & Loyd, 1988). The randomness with which items are assigned to each participant allows for an even and systematic pattern of missingness throughout the survey and across respondents.

A more recent and more popular variation PM design is the three-form design (Graham et al., 1996), which is an example of a multi-form design in which each respondent is randomly assigned one of the several different forms of a survey (Table 1; e.g., Arminger & Sobel, 1990; Raghunathan & Grizzle, 1995; Wacholder, Carroll, Pee, & Gail, 1994). Specifically, all items in the survey are divided into four sets (A, B, C, and X), with a common set (X) that often contains the items most central to the study (Graham et al., 2006). The rest of the items that are not contained in X are equally divided into three subsets A, B, and C. All respondents answer items in X and answer two other randomly selected subsets of items. Thus, each respondent's missing data are items in A, B, or C. There is also the option to include a subset of respondents with no missing data. The 3-form design has been proposed as advantageous to simpler PM designs because each pair of items are completed together by at least one-third of the respondents, thus allowing better estimation of covariances of the data (Graham et al., 1996). However, because the common set of items is recommended to be chosen based on their importance to the study, having them be answered more frequently than the other item sets poses a potential threat to whether the missingness can be considered truly random. Other than the inclusion of a common set of items that are administered to all respondents, the practice of dividing items into a fixed number of sets to make up several different forms appears to be for pragmatic reasons. When a questionnaire could only be administered on paper-and-pencil, it is necessary to have just a few alternate forms of the questionnaire. However, with technological development and increased use of online survey platforms, an alternative PM design is possible.

Each respondent can be assigned a certain percentage of the items, with possible overlap in items between different respondents' items (Table 2). Thus, while each respondent exhibits the same level of data missingness, item overlap between two different respondents can potentially range from being assigned none of the same items to being assigned all of the same items. This will be hereafter referred to as the random percentage (RP) design. With proper randomization, a sufficient sample size, and a reasonably level of missingness, an RP design allows some data to be obtained about each pair of items, just like the 3-form design. In the long run however, no one item would be expected to be administered more often than others. On web-based survey platforms such as Qualtrics, this can be simply implemented by adding a randomizer in the survey flow. Zhang and Yu (In press) simulated planned missingness using the 3-form and random percentage designs, and found that the two designs yielded almost identical estimates when treated with FIML across a range of sample sizes and amount of missingness. Therefore, researchers can make the decision of whether or not to have a common set of items with no intentional missingness based on the central research objectives. Then whether to use a finite number of forms or an RP design in the rest of the items measured can depend on practical factors (e.g., administration mode) (Little & Rhemtulla, 2013).

The Present Research: Planned Missingness as an Approach to Reduce Survey Length

In organizational research, a common research strategy is to administer a survey containing a number of scales measuring a substantial number of variables of interest. Depending on the scope of the study and the variables of concern, survey length can become daunting. In this dissertation I compare two different approaches to shorten survey length with regards to their technical effectiveness and researchers' perceptions of them: the commonplace use of short forms and the much less used planned missingness design.

Given the popularity and convenience of survey research, survey data quality is an important issue. Longer surveys have lower completion rates and higher levels of unplanned missing data (Bowling et al., 2021; Deutskens et al., 2004; Galesic & Bosnjak, 2009; Liu & Wronski, 2018). Thus, shortening the survey can be a proactive way of reducing unplanned missingness. Further, questions near the end of long surveys tend to be answered more quickly, simply, and uniformly (Galesic & Bosnjak, 2009). There have been proposals of ideal and maximum survey lengths based on individuals' average attention span (Revilla & Ochoa, 2017). Although there are a number of post-hoc methods to identify and exclude inattentive responses (e.g., Berry et al., 2019; Meade & Craig, 2012), it would be more efficient to design the survey in ways that are less conducive to careless responding.

In addition to concerns about data quality, logistical constraints are common in survey research. An organization may grant a researcher access to employees, with a caveat that the survey not take more than a certain number of minutes. The use of subject pools at universities is often similarly constrained: researchers are allocated certain number of participant time blocks of a fixed number of minutes. Thus, it is often in researchers' interest to shorten the length of their surveys.

When facing the challenge of reducing survey length and participant burden, one unattractive option is to reduce the number of constructs assessed (e.g., scale back and measure five constructs, rather than the intended eight). Another commonly used approach is to reducing survey length without dropping constructs is the use of short forms of measures. For example, the 50-item International Personality Item Pool (IPIP) Big 5 measure has a shorter version that measures each construct with four items and is 20 items in total (Donnellan et al., 2006; Goldberg, 1999). In cases in which there are no existing short forms, it is common practice to conduct a preliminary study to develop them before administering the survey of interest, retaining items with the highest factor loadings or those with the strongest relationships with criteria of interest, sometimes also with content sampling constraints. Thus, in many scenarios where no short forms exist and one has to be developed, a simple survey study can quickly turn into a much more time- and resource-consuming project.

A potential alternative is to use a PM design, in which all items are retained, with a randomly selected percentage of items are administered to each respondent (Enders, 2010; Graham et al., 2006). The resulting dataset has missingness completely at random (MCAR), which can be subsequently treated with maximum likelihood or multiple imputation (Newman, 2014; Rubin, 1976).

Planned missingness designs have grown in popularity in some social science fields (e.g., Rhemtulla & Hancock, 2016; Rhemtulla & Little, 2012). In developmental and educational research, studies have capitalized on multiform PM designs to reduce assessment time. For example, Foorman et al. (2015) examined factors regarding children's language and reading and how they related to comprehension. They were able to reduce average test time by constructing and administering multiple forms of the variables measured. Similarly, Smits and Vors (2007) demonstrated that study skills and motivation measured using three-form and six-form PM designs closely reproduced full measures. Conrad-Hiebner et al. (2015), in their development and validation of a scale that measures protective factors against child abuse, used the three-form design to reduce the burden of the surveys placed on participating caregivers.

In addition, because of the inherently longitudinal nature of their work, developmental and educational psychologists have expanded planned missingness within questionnaire to across questionnaires (e.g., Little & Rhemtulla, 2013; Mistler & Enders, 2012). Longitudinal wave missing or distributed-lag PM designs omit certain repeated measurement occasions for respondents or cohorts (e.g., Barbot, 2019; Hogue et al., 2013; Little et al., 2017; Little & Rhemtulla, 2013). For example, Lin et al. (2014) studied the relationship between prenatal stress of Mexican American mothers and the infants' subsequent self-regulatory capacity. They collected data at the prenatal, 6-week, and 12-month timepoints from all participants, and administered a wave missing design for the follow-ups at 12-, 18-, and 24-weeks. This type of wave missingness can even be combined with item-level PM within each measurement occasion, although efficiency in estimating rates of change parameter can be reduced (e.g., Rhemtulla et al., 2014; Wood et al., 2019). Because change parameters are usually key effects examined in longitudinal studies, Wu et al. (2016) developed an algorithm that identifies efficient PM designs for longitudinal data collection given the multiple parameters of interest in linear and quadrative growth models.

Another PM design is the two-method design (Graham et al., 2006; Hogue et al., 2013). Developmental researchers are sometimes faced with two methods for measuring the same construct, one considered as the gold standard in terms of measurement quality but more resource-consuming to administer than the other. For example, to examine research questions related to children's attentiveness, the method of independent observation can be less prone to bias and can yield a rich list of children's behaviors. However, it is very expensive to conduct on the entire sample. The alternative is a teacher- or parental-report that is convenient and inexpensive to administer but faces measurement issues such as susceptibility to bias. The two-method PM design provides a solution by administering the less expensive method (e.g., other-report questionnaire) to all respondents but the more expensive method (e.g., independent observer) to only a randomly selected subset of respondents (Rhemtulla & Hancock, 2016; Rhemtulla & Little, 2012).

Interestingly, PM designs have not been solely used for the objective of reducing survey length and improving the cost-effectiveness of studies. In randomized trial studies of behavioral interventions, PM has presented itself as a solution to the assessment reactivity problem, which is when the research procedures other than the specific intervention studied impact the group designated as control. The classic example is alcohol consumption studies finding a substantial reduction not only in the intervention group, but also in the control group (Jenkins et al., 2009). While a number of factors may explain this phenomenon such as legitimate changes in consumption behaviors in the general population, there is worry that the control group is reacting to the baseline or pretest assessment, obscuring the actual effect of the interventions studied. Harel et al. (2011, 2015) tested the effectiveness of a suicide prevention intervention using a randomized pretest-posttest design, with PM in the pretest questionnaire. Participants assigned to the control group were given either the full pretest questionnaire or one of three truncated forms to reduce the amount of information collected from the control group, thereby reducing their exposure to any possible assessment reactivity. They found that the mere exposure to an item at pretest can have an effect on its response at posttest, regardless of the interventions. Using a PM design at pretest allowed them to better tease out the efficacy of the intervention from assessment reactivity.

More recently, social and personality psychology researchers have started capitalizing planned missingness designs for large-scale data collection. For example, Revelle et al. (2020) collected personality data with what the authors call "massively missing at random (MMCAR)," administering a large number of items in total but each item to a small proportion of the total sample, and analyzing the resulting variance-covariance matrix. However, the use of PM designs in substantive research has been largely unexplored in the field of I/O psychology. The only exceptions to my knowledge are Marcus-Blank et al. (2015) and Yamada (2020), both unpublished. The former examined the predictive ability of rationality for various life outcomes in addition to cognitive ability, personality, and decision-making style, and implemented 25% PM in all control variables in order to reduce study length. The latter investigated perceptions of sleep climate and used a PM design to reduce the respondent burden placed on busy resident physicians. In I/O survey research, the need to reduce survey length is routinely addressed with either compromising the number of constructs measured or using a short version of a full-length measure. Planned missingness, although rarely used, presents itself as a promising alternative.

Study 1: Planned Missingness vs. Short Forms in Existing Data

One previous study conducted an initial comparison of short form and planned missingness as strategies for reducing survey length (Yoon & Sackett, 2016). The authors used a dataset containing self-reports of personality and workplace behaviors (Sackett et al., 2006). For each of the measures, the authors conducted exploratory factor analyses and created half-length short forms based on the highest loading items and computed correlation estimates based on short forms. A 50% PM design was implemented by randomly removing half of the datapoints per measure for each respondent. Multiple imputation was then used to treat the resulting PM dataset. The average absolute difference in correlations between those computed in the full dataset (i.e., "truth") and those computed with short forms was .034, and that between truth and PM was .019. Results demonstrated that estimates of scale intercorrelations based on the planned missingness design more closely approximated those of the full dataset than did estimates based on short forms.

To examine whether their finding was a function of the specific characteristics idiosyncratic to the data used, the current study replicates Yoon and Sackett (2016) in two other publicly available datasets. These two datasets were chosen based on several criteria. First, a sufficient sample with complete responses was needed. Second, constructs need to be measured with multi-item scales. Third, the dataset needs to contain multiple constructs. Last, constructs collected should be broadly representative of those typically studied in survey research (i.e., self-reports and not ability constructs). In addition, I reanalyzed the data used in Yoon and Sackett (2016) with a minor change to the original procedure.

Method

Datasets

Datasets A and B are public datasets collected by the Open Source Psychometrics Project (2018). The quality of these open source datasets had previously been demonstrated to be comparable to or better than those collected from Amazon Mechanical Turk (Open Source Psychometrics Project, 2018).

Dataset A. Dataset A measured the respondents' dark triad of personality: machiavellianism, narcissism, and psychopathy (Paulhus & Jones, 2011). A total of 27 items measured each of these three constructs with nine items, with values ranging from 1 (disagree) to 5 (agree). The raw dataset contained 18,192 complete cases with no missing data. Reverse items were recoded accordingly.

Dataset B. Dataset B measured individuals' vocational interest and personality during 2015 to 2018. It includes a 48-item Holland's RIASEC vocational interest measure, with each of the domains measured by eight items (Armstrong et al., 2008). Items specified various vocational activities and asked respondents to indicate how much they would like to perform that task on a five-point scale ranging from 1 (dislike) to 5 (enjoy). Also included was the ten-item personality inventory (TIPI; Gosling et al., 2003), a vocabulary test, and demographic questions, which were not used in the current study. The raw dataset contained 145,828, and complete cases were retained, resulting in a final population sample of 90,513.

Dataset C. Dataset C was the same dataset used in Yoon and Sackett (2016). Sackett et al. (2006) used a Web-based survey via email to recruit participants from the 4,218 full-time employees of a large university in the Midwest. 965 individuals started the study. 65 individuals were deleted from the analysis because of missing or duplicate data, resulting in a final sample of 900. Sackett et al. (2006) used a 50-item Big Five Personality measure from the International Personality Item Pool (Goldberg, 2001), a 15-item organizational citizenship behavior measure developed by Laczo (2002), an 18-item CWB measure drawn from Bennett and Robinson (2000), and a 17-item CWB developed by Laczo (2002). One item from the original 19-item Bennett and Robinson (2000) measure of workplace deviance “cursed at someone at work” was excluded because it is similar to an item included in another part of their survey. For Big Five measures, they used a Likert-type response format: 1 indicates 'strongly disagree', whereas 5 indicates 'strongly agree'. For OCB and CWB measures, they used a 4-point scale (1 = Never to 4 = Frequently).

Procedure

In Yoon and Sackett (2016), the entire dataset was used to create short forms. However, in a realistic research setting, a study most likely administers a short form to measure a construct of interest that was developed based on a separate sample. Therefore, in the current study, I first randomly divided each of the datasets into two equal halves: a development dataset and a test dataset. Pairwise correlations computed using the full test dataset without any missing data are used as true correlation values against which short form and planned missingness correlational estimates are compared.

Two versions of half-length short forms were created for the Big Five scales in Dataset A, the RIASEC scales in Dataset B, and the personality, OCB, and CWB scales in Dataset C. First, following the procedure of Yoon and Sackett (2016), an exploratory factor analysis was conducted on each scale in each of the development datasets using the

R package *fungible* (Waller, 2020). Half of the items with the highest loadings for each scale were used to create loadings-based short forms. Although this approach is simple to implement and commonly used in practice, Cortina et al. (2020) prescribe a combination of psychometric criteria that should be considered simultaneously for shortening scales, including internal consistency reliability, part-total correlations, and construct validity and content coverage of the items when applicable. They created an R Shiny app called the Optimization App for Selecting Item Subsets (OASIS) to generate short forms based on these considerations. By inputting an item-level dataset of the original full-length scale as well as any convergent and divergent validation scales if available, the OASIS iterates through all possible combinations of items given the desired length of the short form and computes Cronbach's alpha (α), Guttman's lambda-2 (λ_2), omega-hierarchical (ω_h), part-whole correlation between each item subset and the full scale, and correlations with any validation scales. An optimal short scale can then be chosen by simultaneously considering the recommended thresholds of $\alpha > .75$, $\lambda_2 > .75$, $\omega_h > .80$, correlation with convergent validity scale > .70, and correlation with divergent validity scale < .35. Thus, a second version of the short forms were created using OASIS. The respective short form items in the test datasets were then used to obtain both loadings-based short form estimates and OASIS-based short form estimates.

To simulate a planned missingness design in the datasets, half of the items within each measurement scale were randomly selected to be missing for each case in each of the test datasets. Multiple imputation was then performed on the resulting MCAR datasets to generate 40 imputed datasets using the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2010).

Analyses

For each of the three datasets, coefficient alphas of all measures and correlation matrices among constructs were computed based on the full data (“truth”), loadings-based short forms, OASIS-based short forms, and planned missing and imputed data. Then, I calculated the absolute values of the differences between the correlations based on the "limited" data sets (short forms and imputed dataset) and the correlation based on the full data in order to investigate which limited version is closer to the full data. For better comparison, I computed the mean and standard deviation of these absolute values. Because I was interested in investigating the differences between the correlations based on the short forms and on the imputed data set, I conducted paired-samples t-tests to compare the absolute values of the differences between the correlations based on the short forms and the correlations based on the full data, and the absolute values of the differences between the correlations based on the imputed data and the correlations based on the full data. All analyses were conducted in R.

Results

Although the three datasets differed in their sample sizes and the constructs they measured, results were consistent across samples. With regards to scale internal consistencies, data with planned missingness and treated with multiple imputation (Tables 3~5) closely replicated the scale coefficient alphas of the full data in all three datasets. On the other hand, internal consistencies of short forms created using either factor loadings only or multiple criteria via OASIS tended to fall short of those of the full measures despite retaining the empirically best-performing items. This is to be expected as the number of items for each scale was reduced by half.

Related to the short forms' decreased reliabilities, scale intercorrelations computed using short forms tended to be attenuated compared to the "true" intercorrelations in the full datasets, whereas imputed data more closely estimated the intercorrelations (Tables 6~8). Across all three datasets, both the absolute differences between true intercorrelations and loadings-based short form intercorrelations (M = .05, SD = .03; M = .07, SD = .04; and M = .06, SD = .05 for datasets A, B, and C respectively) and those between true intercorrelations and OASIS-based short form intercorrelations (M = .02, SD = .01; M = .07, SD = .04; and M = .06, SD = .04 for datasets A, B, and C respectively) were significantly larger and varied more than the absolute differences between true intercorrelations and imputed data intercorrelations (M = .01, SD = .01; M = .01, SD = .01; and M = .02, SD = .02 for datasets A, B, and C respectively). In other words, when short forms were created based strictly on strengths of factor loadings, imputed data produced intercorrelations that were significantly closer to "true" intercorrelations than short forms by .047 in dataset A ($t(2)$=3.5, p = .036, 95% CI = [.015, .079]), .065 in dataset B ($t(13)$=6.03, p < .001, 95% CI = [.042, .088]), and .037 in dataset C ($t(27)$=3.57, p < .001, 95% CI = [.015, .058]). Similarly, when short forms were created based on multiple psychometric criteria using the OASIS, intercorrelations estimates were statistically significantly farther from "true" intercorrelations than those computed with imputed data by .061 in dataset B ($t(13)$=5.49, p < .001, 95% CI = [.039, .084]) and .033 in dataset C ($t(27)$=3.69, p = .001, 95% CI = [.015, .051]). Differences in the accuracy of intercorrelation estimates between OASIS-based short forms and imputed data were not significantly difference in dataset A. The two versions of the short forms did not yield significant different results in all three datasets.

The difference in the results obtained in the current study with loadings-based short forms in dataset C and those found by Yoon and Sackett (2016) was subtle but expected. The original findings were based on short forms that were created and tested on the same, full dataset. In the current study, the full sample was divided into a development sample, on which exploratory factor analyses were conducted and short forms were created, and a test sample, on which reliability and correlation estimates were obtained. Therefore, the mean absolute difference between true and short form intercorrelations found here (.06) was larger than that found originally (.03).

Discussion

Researchers often face a time or length constraint when designing surveys. Assuming that the number of constructs or scales to be measured cannot be reduced, a common practice is to use short forms. However, this approach inevitably excludes some items from being administered altogether and requires much time and resources to first develop the short forms when they do not already exist. A much less explored alternative is to implement a planned missingness design such that data is collected with missingness randomly.

Yoon and Sackett (2016) conducted the first empirical comparison between these two approaches for reducing survey length. Their results demonstrated that correlations based on short forms somewhat closely reproduced the correlations found in the full dataset. However, implementing the alternative approach (planned missingness design paired with multiple imputation) yielded closer estimates.

To further explore the generalizability of their finding, I replicated this initial study on two more large-sample public datasets. I also slightly modified the original procedure such that short forms were created on a sample separate from the one on which estimates were based. This is a more realistic representation of standard research practice and reduces the risk of overfitting. In addition to creating short forms strictly based on factor loadings, I also used a second approach recommended by Cortina et al. (2020) that considers multiple indices of internal consistency, part-whole correlation, and convergent and divergent validity by using the R Shiny app OASIS. Overall, I reached similar conclusions based on all three datasets, namely that a planned missingness design can outperform using short measures for reproducing true intercorrelation estimates. Intercorrelations calculated using both versions of the short forms face the attenuation issue due to decreased reliability associated with the short measure length. Practically, a planned missingness design can be easily implemented using the many internet-based survey platforms whereas short forms may need to be first developed and validated in prior pilot studies. Thus, the planned missingness approach for reducing survey length may be advantageous both statistically and practically.

This study is not without limitations. Although it sought to explore the generalizability of the comparison results by using multiple datasets, more systematic examination can beneficial. It is unclear whether results were associated with characteristics specific to the datasets used. Relatedly, half-length short forms were created, constraining the amount of missingness at 50%. It will be valuable to systematically vary data characteristics and study procedures to expand the short form versus planned missingness comparison across different conditions and explore potential boundary conditions of this apparent advantage of planned missingness. A simulation is much better suited for these study objectives and is conducted in the following study.

Study 2: Systematic Simulation of Short Form vs. Planned Missingness

With the use of computer-based survey tools, planned missingness designs can be much easier to implement than having to develop short forms, and prior results suggest they can be advantageous to producing accurate estimates as well. The present study aims to expand upon the findings of Yoon and Sackett (2016) and Study 1 in this dissertation. With the goal of reproducing "true" scale intercorrelations in the population, I conducted a Monte Carlo simulation to explore if and how the short form–planned missingness comparison is affected by various data characteristics, including length of short form or amount of missingness, sample size, number of scales, and true intercorrelation among the scales.

Method

To provide a clear description of the simulation procedures, the term "variable" is used throughout to refer to the population and sample characteristics manipulated in the simulation. The term "item" is used to refer to the individual items simulated in the data. And the term "scale" is used to refer to the items that measure a single construct.

First, several population parameters were manipulated. The number of scales (s) ranged from two to five, each indicated by ten items (with a total of 20 to 50 items). Factor loadings of the ten items on each scale were specified to vary from .30 to .75 at .05 increments. Further manipulated were the scale intercorrelations. For the two-scale conditions, the correlation between the two constructs varied from 0 to .70 at .10 increments. For each of the three-scale (30-item), four-scale (40-item), and five-scale (50-item) conditions, the mean intercorrelation among constructs (M_r) varied from 0 to .70 at .10 increments, and the standard deviation of construct intercorrelations (SD_r)

varied from 0 to .30 at .05 increments, resulting in a total of 176 unique combinations of population parameters.

For each of the unique population parameter combinations, common factor correlation matrices are first generated based on normal distributions using M_r and SD_r. For example, for the population condition of four-scale, intercorrelation mean (M_r) = .50, and intercorrelation standard deviation (SD_r) = .20, six correlation values are randomly generated from $N(.50, .20^2)$ to fill in the lower triangle of the square matrix then mirrored onto the upper triangle. If a number larger than 1 was sampled from the normal distribution and therefore impossible for a correlation matrix, it was manually adjusted to be .99. The resulting matrices were checked to ensure its positive definiteness.

Next, the item-level correlation matrices were created based on the factor correlation matrix for each condition and the fixed factor loadings matrix using the standard common factor model for a population correlation matrix:

$$P = \Lambda\Phi\Lambda' + \Omega$$

Where P is the item-level $k \times k$ correlation matrix if the total number of items is k, Λ is the $k \times r$ factor pattern loading matrix for r number of constructs or scales, Φ is the $r \times r$ factor correlation matrix, and Ω is the $k \times k$ diagonal matrix of unique variances, defined here as $I - diag(\Lambda\Phi\Lambda')$ (Gnambs & Staufenbiel, 2016; Hong, 1999).

For each population (defined by its item-level correlation matrix), independent samples were drawn using the mvrnorm function in the R package MASS (Ripley et al., 2021) with sizes (n) ranging from 100 to 1,000 at 100 increments and varying missingness levels (m) per scale from 10% to 80% at 10% increments, consisting of 80 unique conditions. Each condition was performed for 100 iterations. See Table 9 for all variables that were manipulated and their ranges.

For each unique condition, four different scenarios for reducing survey length were examined, including three scenarios of using short forms and the planned missingness approach with the goal of comparing the capacity to which short form and planned missingness estimates reproduce true population estimates (Figure 1). Short Form A and B both use short forms developed empirically based on exploratory factor analysis results, choosing items with the highest loadings. The loadings-based method of developing short forms was used because the two methods of empirically creating short forms included in Study 1 (i.e., loadings-based and OASIS-based) yielded similar results and using factor loadings is much less computationally expensive than using the iterative algorithm of OASIS. Short Form A simulates the scenario in which such short measures already exist, or factor loadings needed to readily create such short measures have already been obtained in a previous study or exist in the literature. In this situation, no part of the current sample needs to be spent on first developing short forms. Short Form B simulates the scenario in which no established short forms or psychometric information needed to readily create short forms exist, and part of the sample needs to first be expended on developing them. Short Form C simulates the scenario that items are chosen for short measures based on theoretical justifications or item content coverage reasons and are therefore not necessarily items with the highest loadings. Lastly, the Planned Missingness scenario implements a planned missingness design treated with multiple imputation. Specific simulation procedures for each of these four scenarios are elaborated below.

Short Form A: short forms based on pre-existing factor analysis results

To simulate the situation that short forms have already been developed prior to the study of interest, a separate development sample with sufficient and typical sample size (n = 500) was first drawn to conduct exploratory factor analyses for each scale to identify the best-performing items per scale; these best-performing items constituted the short forms of the scales. The number of items that made up the short form was dependent on the variable amount of missingness (m). Then a second test sample was drawn (n depending on the sample size specified in the condition), in which only the short form items were used to compute the Short Form A estimates.

As an example, for the two-scale, M_r = .2, n = 100, and m = 50% condition, based on the specified population item-level correlation matrix, a sample of 500 cases was first drawn to be the development dataset (sample A) on which exploratory factor analyses were conducted. For each scale, the five items (50% of ten items) with highest loadings on the first factor were retained as the short form. In a different 100-case sample (sample T) drawn from the population, the five items per scale previously identified for the short form were aggregated to compute the scale scores, and the correlation between the two scales was computed as the Short Form A estimate.

Short Form B: short forms to be developed with part of the sample

Scenario Short Form A makes the assumption that prior short forms of the measures already exist and the entirety of the second, current sample is used. However, another common survey study scenario is when short forms of the measures do not yet exist, and the sample that is to be recruited needs to be divided into a development sample to first create short forms and a second sample to administer the short form and obtain estimates.

Therefore, for Short Form B, I assigned half of the sample to develop the short forms. A further complication is the constraint on survey length. Assuming the missingness level (m%) in each condition puts a constraint on the maximum number of items that can be attentively answered by respondents. The same constraint should be applied to the developmental sample and therefore only a portion of it (1-m%) can be used to develop each short form.

Take again the example of the two-scale, $M_r = .2$, n = 100, and m = 50% condition. The simulated sample T of 100 was first divided into sample T1 for developing short forms ($n = 50$) and sample T2 for testing ($n = 50$). Because the maximum number of items to be administered to each respondent was specified by the missingness level of 50% to be 10 (2 scales × 10 items per scale × (1-50% missingness)), sample T1 was further divided into two samples of 25, each in charge of developing the short form of one scale. The items selected were then applied to sample T2 to obtain short form estimates.

Short Form C: short forms developed based on item content

Short Form A and B develop short forms empirically based on factor analysis results. However, in situations where researchers have reason to retain items based on factors other than empirical loadings (e.g., for construct content coverage, balance between positively and negatively coded items, etc.), items retained in the short form do not necessarily have the highest loading. I simulated this scenario by choosing items that compose the short forms randomly, then obtaining Short Form C estimates from the sample drawn.

Continuing with the example of the two-scale, $M_r = .2$, $n = 100$, and $m = 50\%$ condition, five items were randomly identified to make up the short form for each scale. The correlation between the two scales was computed in the simulated sample T as the Short Form C estimate.

Planned Missingness

In the Planned Missingness scenario, a number of randomly selected items were set to be missing for *each* case (i.e., respondent), simulating missingness completely at random (MCAR). In the example of the two-scale, $M_r = .2$, $n = 100$, and $m = 50\%$ condition, each case in the simulated sample T has data for a different subset of five items for each scale. Multiple imputation was implemented to deal with the resulting data using the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2010). The correlation between the two scales were computed for each of the 20 imputed datasets and pooled as the Planned Missingness estimate.

Thus, for each condition, I computed the true scale intercorrelations and four sets of estimates: Short Form A, B, and C scale intercorrelations, and the Planned Missingness scale intercorrelations. Two sets of criteria were examined: 1) the absolute differences between truth and the four sets of estimates; and 2) the differences between deviations of each of the Short Form estimates from truth and deviations of Planned Missingness estimates from truth (with a positive value indicating Planned Missingness estimates were closer to truth than Short Form estimates). Because the criteria were calculated for each pair of scales, the pairwise values were aggregated prior to further analyses. To do so, within each condition with more than two scales (i.e., three-, four-, and five-scale conditions), the mean across all scale pairs was computed for each criterion.

Working with a large number of different population parameter combinations, simulating the four different length-reduction scenarios with unique combinations of amount of missingness and sample size for each of the defined populations, and conducting each simulation condition for 100 iterations required tremendous processing power. Further, the simulation procedure (including sampling, removing data to simulate missingness, and imputation) was very time-consuming given the sheer number of conditions. Therefore, computational resources at the Minnesota Supercomputing Institute were used. By implementing parallel processing within condition and dividing conditions into four scripts that ran simultaneously, the simulation lasted approximately 114 hours.

Analyses

To systematically compare the performance of different scenarios of using Short Forms and the Planned Missingness design under different conditions, the criteria are used as dependent variables in separate regression models, with independent variables including amount of missingness (m), sample size (n), number of scales (s), intercorrelation mean (M_r), and intercorrelation standard deviation (SD_r) in the population. Step 2 models included all pairwise interaction terms. To aid interpretation of intercepts and coefficients, all predictors were centered, and sample sizes were scaled to be in units of 100. In addition, results were plotted to allow visual inspections of interaction effects.

Results

Across all conditions, the average absolute deviation of estimates from truth was .039 (*SD* = .024) for short form A, .056 (*SD* = .036) for short form B, .072 (*SD* = .055) for short form C, and .050 (*SD* = .042) for planned missingness.

Regression models were first performed to predict the absolute deviation from "truth" of each of short form A, short form B, short form C, and planned missingness with the various data characteristics that were manipulated (Table 10). When short forms were empirically derived based on factor loadings and regardless of whether they already exist or had to be developed using part of the sample, there was a greater absolute difference between short form estimates of scale intercorrelations and the true correlations in the population (DV1 and DV2) when short forms were shorter (higher missingness levels), when samples were smaller, and when the mean of true intercorrelations was smaller. The number of scales or the average magnitude of true scale intercorrelations did not significantly affect the accuracy of short form estimates.

Interestingly, short forms behaved similarly to the planned missingness design when the items retained in the short forms were chosen at random, such that the deviation from true intercorrelations became larger when the scales were shorter, sample size was smaller, and when true intercorrelations were stronger overall. Furthermore, when using a planned missingness design, deviation from true parameters became larger when the number of scales increased and intercorrelation SD became larger.

Step 2 regression models using this same set of DVs additionally included pairwise interaction terms of the five predictors. The sensitivity of *p* values to sample size makes statistical significance difficult to interpret given the large number of conditions and iterations included in the regression. Thus, visualization of the interactions can be more helpful in determining their practical significance (Figures 2~11). Interactions that notably exacerbated deviation from true population intercorrelations included higher missingness level × smaller sample size (Figure 6), especially for the planned missingness method, higher missingness level × larger mean scale intercorrelation (Figure 7), higher missingness level × larger scale intercorrelation standard deviation (Figure 8), and higher mean scale intercorrelation × smaller sample size (Figure 9), especially when short form items were chosen randomly.

The second half of the regression results subtracted the planned missingness deviations (DV4) from each of the short form deviations (DVs 1, 2, and 3). Thus, positive values indicated that planned missingness was more accurate in reproducing true intercorrelations while negative values indicated that the short forms were more accurate. Results from Table 11 demonstrated that at average levels of all predictors (missingness level = 45%, sample size = 550, number of scales = 3.5, population intercorrelation mean = .30, population intercorrelation standard deviation = .11), the intercepts of the models were –.014, –.002, and .018, respectively. This means that overall, across all conditions, empirically derived short form estimates slightly outperformed planned missingness estimates in approximating truth when short forms or data needed to readily create short forms already existed. When a proportion of the sample needed to be spent on empirically developing short forms, they performed essentially equally with implementing planned missingness on average. However, when short forms were created not solely empirically, planned missingness tended to reproduce true intercorrelations more accurately. Two-way interactions were added at the second steps of the regression models and plotted in Figures 12~21. Visually, several interaction effects favored using short forms over planned missingness in reproducing true population intercorrelations: larger number of factors × higher missingness level for empirically derived short forms A and B (Figure 12), smaller sample size × higher missingness level for all short form types (Figure 16), higher missingness level × larger mean scale intercorrelation (Figure 17) and smaller sample size × larger mean scale intercorrelation (Figure 19) for empirically derived short forms. However, across many other conditions, deviations from truth when using planned missingness remained around the same magnitude or smaller than that when using short forms.

Discussion

The present study systematically varies a number of data characteristics to compare the performance of using short form scales and implementing a planned missingness design with multiple imputation in order to reduce survey length for producing accurate estimates.

Findings from the comparison between short form A and planned missingness demonstrate that when short scales already exist, using them may produce slightly more accurate correlation estimates than implementing planned missingness. This would also include scenarios in which no short forms are explicitly available, but there are published factor analysis results that enable researchers to readily create short forms with the highest-loading items without collecting additional data. Established short forms likely retain items that are the best indicators of the constructs, and no subsequent effort is needed to impute data as is in the case of planned missingness.

However, when short forms do not yet exist and part of the sample needs to be expended for scale development purposes (short form B), the benefit of planned missingness is more pronounced. Across different conditions, the two approaches perform equivalently in yielding accurate estimates. Furthermore, the additional processing power needed to impute data after data collection is negligible compared to the time and effort needed to develop and validate short forms.

It is worth pointing out that the short form A condition assumes that short forms exist for *all* scales measured, and the short form B condition assumes the opposite for *all* scales. It is more likely that there are existing short scales for some of the constructs measured but not others. This scenario would fall in the middle of and be bounded by the results of A and B, with short forms producing slightly more accurate estimates than planned missingness on average.

While short forms A and B compose short forms strictly based on empirical factor loadings, this is not the only way to create short scales. An item with a weaker loading may be preferred to one with a higher loading to better represent the construct domain, if it is more central to the construct theoretically, if it is superior on other psychometric properties, if it is less transparent or has higher face validity, if subject matter experts feel that it is more relevant or critical to measure, etc. Although an imperfect simulation, short forms C tests this by choosing short form items randomly for each iteration. Thus, across different iterations, short forms can consist of the highest-loading items, the lowest-loading items, and everything in between. Findings show that when a non-empirical approach is taken to select items for the short forms, planned missingness actually outperform short forms in reproducing population estimates on average.

In real research studies, a hybrid approach to short form construction may be undertaken where the highest-loading items are retained with some consideration of content coverage, theoretical relevance, etc. There was no reliable way to simulate the human judgment component of this process, but any hybrid short forms should fall somewhere in between short forms B and C.

It is important to not overinterpret the differences between the performance of the three short forms and planned missingness designs in approximating true estimates. Although subtle variations exist, general deviations from population estimates remain small, and difference between short form and planned missingness deviations are even smaller.

This study is a comparison of using short forms and implementing a planned missingness design across various different research conditions with regards to strictly their technical capacity to produce accurate population estimates. While findings show that both approaches have conditions in which they slightly outperform the other, technical accuracy is often not the sole consideration when designing a research study. When well-developed, empirically based short forms already exist in the literature or from prior studies for all scales of a study, administering them will both obtain accurate estimates and be minimally effortful. However, when empirically derived short forms do not already exist for at least some of the scales intended to be collected, conducting scale development studies will likely be much more costly and effortful, whether by collecting another independent sample for the empirical approach, recruiting subject matter experts for the rational approach, or both for a hybrid approach. The findings of the current simulation show that the extra effort of developing short scales does not pay off in more accurate parameter estimates as compared with the simple procedure of implementing planned missingness.

As with any simulation, the present study is not without limitations. I simulated population data with up to five constructs as more number of constructs becomes a strain on processing power and computer memory. However, I do not expect results to differ significantly with more constructs.

Further, I simulated short form and planned missing data by drawing full, complete samples, then removing data based on each specified condition. Thus, the external generalizability of findings relies on the assumption that individuals respond in the same manner to a full measure, a shortened measure, and a subset of items that are randomly selected from the full measure. Although an important assumption to point out, it is likely a valid one. From any individual respondent's perspective, being given the short form or the planned missingness design will be indistinguishable, as the number of items administered will be equal, and any differences in specific items will only be apparent if compared side-by-side to those given to another respondent. The full measure will differ from both short forms and PM designs in length. However, unless the total length of the survey is made very salient, it is unlikely that participants would be aware of how many items they are answering and respond in a systematically different manner as a result.

It is also worth noting that all three types of short forms simulated in this study begin with a full measure of a construct and retain a subset of the items within the full measure. As is the case with many psychological constructs, often many different instruments have been developed to measure the same construct with varying number of items. Identifying a separate, shorter measure that contain completely different items is another approach to reduce study length not examined in the present study. The effectiveness of such an approach will depend solely on the psychometric properties of the shorter measure used.

Additionally, there is an important caveat to the use of planned missingness designs in certain situations. In this particular study, the specific multiple imputation procedure implemented to treat missing data was predictive mean matching (PMM). It is the default technique in many statistical software and is a hot deck procedure that imputes by identifying and selecting suitable donor cases. Prior research has found that this may become problematic in extreme conditions where donor cases are rare, due to a small sample size and large amounts of missingness (Kleinke, 2018). In the present study, I found that the particular combination of a high level of missingness percentage and a small sample size (70% missingness with n = 100 and 80% missingness with n = 100, 200, or 300) led to high risks of imputation failures (Figure 22). It is possible that in such extreme cases, using short form is the only option. No imputation failures occurred for missingness levels lower than 70% or sample sizes of 500 or more. The same pattern of findings was reported by Zhang and Yu (In press), who simulated planned missingness designs and treated the missing data with FIML estimation. In conditions that suffered both from a small sample size and a large amount of missingness, FIML estimations became increasingly inaccurate and even faced convergence failure issues. Future research should systematically test these thresholds under different conditions as well as explore the effectiveness of other treatment techniques for such extreme conditions.

Although results in the present study point to 70% missingness as a threshold for inaccurate estimates and imputation failure when using planned missingness, I caution against using this percentage as a definitive guide for implementing a planned missingness design. First, the impact of the proportion of missing data will likely depend on the total number of items in the initial, full measure. In this simulation, full measure length was fixed at ten items. In a ten-item measure, 70% missingness leaves three items per scale to be observed. In a longer measure, a 50-item instrument for example, 70% missingness means that 15 items are still observed, providing more information needed for imputation or maximum likelihood estimation. Another factor to consider is that no unplanned missingness was simulated in this study. Any researcher designing a survey might anticipate a degree of naturally occurring missingness, which would increase the overall missingness level beyond the amount of missingness planned. Thus, I urge researchers to err on the side of caution when determining the level of planned missingness.

In summary, this present study compares the performance of two general strategies for shortening survey length while still producing accurate estimates: short form and planned missingness. Findings are in favor of implementing a planned missingness design when empirically derived short forms are to be developed, as it requires much less resources and is able to approximate population estimates well. Planned missingness is also recommended if short form items are not selected empirically.

Study 3: Knowledge and Perceptions of Planned Missingness

Studies 1 and 2 demonstrate the technical equivalence at the least and advantage in some conditions of the planned missingness approach for reducing survey length to using short forms. This study takes a qualitative approach to examine the feasibility of planned missingness in practice. Fortunately, planned missingness generally has high implementation feasibility with the assistance of many internet-based survey platforms. For example, on Qualtrics, planned missingness can be achieved with the simple addition of a randomizer within a scale block, and the number of items randomly selected for each respondent can be specified with an entry of a number. This current study is interested in feasibility with regards to researcher knowledge, sentiment, and receptiveness.

The overall goal of this study is to understand why planned missingness designs are not used more in the social sciences and at all in industrial/organizational (I/O) psychology. I have had several anecdotal experiences that led me to suspect it is primarily because of unfamiliarity. Oftentimes when I mentioned the concept of planned missingness to a professional I/O colleague, they did not know what I was referring to. There also seems to be an inherent distrust of missing data—the sentiment that missing data is always a bad thing so we should never deliberately collect missing data. This distrust seems to only get worse with the use of planned missingness with multiple imputation. As if purposefully collecting incomplete data wasn't bad enough, now we are making up data?

I believe that concerns such as these are not unreasonable given the traditional statistical training we have received, but can be alleviated with more knowledge and understanding of planned missingness. An understanding of the different types of missingness (e.g., ignorable vs. non-ignorable) and their implications will be crucial. Planned missingness is a useful tool precisely because the missingness is deliberate, and therefore the researcher is in control of its nature and can statistically deal with it accordingly. This is fundamentally different from missingness that occurs naturally throughout survey implementation for a variety of reasons. The key is to appreciate that missing data are not necessarily a bad thing. It might also help to point out that using short forms of measures is also collecting missing data, and the researcher needs to make the decision beforehand about which items to exclude altogether. Similarly, the interpretation of multiple imputation as “making up data” is an understandable but superficial one. The imputation of data is merely a process by which population parameters can be estimated. There are of course caveats to directly using item-level imputed data, but as an intermediate step to obtaining more accurate estimates, it is no different from the process of the well-accepted procedure of correcting observed meta-analytic correlations for measurement unreliability. Akin to unreliability, missing data stands in the way of obtaining “true” estimates, and imputing data a number of time and pooling the results is just a means to an end.

At the same time, it is without question that planned missingness is not appropriate across all data collection settings. I talked to an I/O research scientist about planned missingness whose work consists of mostly selection testing, and their first reaction was the legal implications and risks, and with good reason. It enters dangerous territory if applicants are selected based on a test on which they are all given different questions, and their item-level imputed responses are used for high-stakes decisions, and this is certainly not what I am advocating. However, I also encountered a dataset collected for a concurrent validation study that measured fewer constructs with fewer items per construct than the researcher would have wanted because of time constraint. When I asked whether using planned missingness approach was considered, the pushback was less founded and simply "we can't do that for a validation study." Thus, this study was devised to explore the research settings or scenarios that people deem appropriate to implement planned missingness, and in those that they deem not appropriate, why not.

Method

Participants

Given the goal of assessing familiarity of and sentiment toward planned missingness among working professionals in the I/O psychology or related fields, several channels were used for recruitment, including contacting I/O organizations and I/O academics, disseminating the study advertisement on social media, and reaching out to personal networks.

Procedure

Upon giving consent, respondents were given a 10-minute survey (Appendix A). The first part of the survey mainly sought to understand researchers' prior familiarity with planned missingness designs and their past experiences with reducing survey length. They were first asked about the frequency at which they collect self-report data and need to reduce the length of the study. They were then provided with the short definition "A planned missingness (PM) design can be implemented in survey studies, in which a randomly selected percentage of items are administered to each respondent. By using a PM design, the length of a survey can be reduced," and asked if they were familiar with this concept and if they have implemented it in their past work. Respondents who reported that they have used it in their work were then asked what kind of study they were conducting and what missingness treatment techniques they used. Respondents who reported that they have not used it were asked why. Lastly, all respondents were inquired about the approach(es) that they have taken to reduce survey length.

The second part of the survey provided respondents with more detailed information about both the short form and the planned missingness approaches, including a more elaborate definition of planned missingness, an example of what the two approaches would look like, and briefly summarized the simulation findings from study 2, namely that the two approaches are technically equivalent across many scenarios. Respondents were then provided four different test scenarios and asked whether they would prefer one approach to the other for any contextual or practical reasons in each situation. The four test situations included a personality research study on Amazon Mechanical Turk (MTurk), a selection battery, an engagement survey, and a concurrent validation study.

Lastly, respondents were asked to provide information regarding their educational and work background.

Results

A total of 88 research scientists took part in this survey. Of those who provided their educational and work information, two obtained Bachelor's degrees, 16 Master's, and 59 Doctoral. A majority of participants held degrees in I/O psychology (n = 57), while 14 had business degrees (including Organizational Behavior, Human Resources, etc.). Four respondents held degrees in other non-I/O areas of psychology, and one in political science.

With regards to work background, 32 were in academia, 13 in the government sector, 25 in private industry, 5 in external consulting, and 1 other. Years of work experience ranged from a minimum of .5 to 45 years (M = 9.57, SD = 9.60).

All respondents indicated that they have designed or contributed to designing at least one self-report data collection effort in their work, with 49 (56%) indicating that they have designed more than 20 (Table 12). Further, while four respondents (4.5%) indicated that they have never had the need to reduce the length of a self-report study, everyone else indicated that they have needed to reduce study length in their work (Table 12).

Among those who have had the need to reduce study length, 70 researchers (83.3%) reported using short forms, 75 (89.3%) reported cutting down the number of scales collected, and 8 (9.5%) reported using planned missingness.

38 researchers indicated that they were familiar with the concept of planned missingness prior to participating in the study while 50 indicated that they were not. Of individuals who were familiar with the concept, only ten have actually implemented a PM design in their work. Among individuals who have used PM, five have used FIML and two have used multiple imputation. In these individuals' past studies with a PM design, seven were conducting survey research studies (e.g., academic research), two were conducting concurrent validation studies, none applied it with predictive validation studies or missingness-related methodological studies.

The 28 individuals who were aware of what PM is, have had a need to reduce study length, but have not implemented PM, were asked to provide a free response of reason(s) for not having used it. I coded these reasons into several broad categories: lack of understanding, worry about others' perception, preference for alternative approaches, and limitation of PM. Several responses provided reasons across multiple categories.

Overall, 13 responses reflected a lack of understanding such as general unawareness (e.g., "not comfortable using it;" "it creates problems of its own;" "had honestly not seen it used in applied settings"), lack of knowledge in subsequent analysis (e.g., "unfamiliar with how to analyze"), and a misunderstanding about its effects (e.g., "I worry it will reduce the sample size too much"). Ten responses mentioned concerns regarding how the methodology would be perceived by others including reviewers (e.g., "It's not clear to me that reviewers will 'trust' it", "hesitant that it will get pushback in the review process as a 'fatal flaw'), management and leadership (e.g., "tough sell to management", "face validity concerns from stakeholders"), and colleagues (e.g., "other team members not on board"). Three responses stated that there are alternative methods to shorten surveys without pointing to a specific approach. Lastly, three responses suggest limitations of PM under certain setting, including "not giving all participants the same set of questions."

In the second part of the survey, respondents were given a more detailed description of planned missingness and a short summary of results from Study 2, namely that the technical effectiveness of planned missingness for reproducing population parameters is largely equivalent to that of using short forms across many scenarios (Appendix A). Under this technical equivalence, respondents were given four study scenarios and asked if they prefer one approach to the other for any contextual reasons. For each scenario, if respondents preferred either using short forms or implementing a planned missingness design to the alternative, they were invited to provide a free response of the rationale for their preference. The breakdown of respondents' preferences by scenario can be found in Table 13.

For scenario 1 (a personality research study using a sample of Amazon Mechanical Turk workers), 14 researchers (18.4%) preferred using short forms whereas 41 (53.9%) preferred implementing a planned missingness design. The rest had no preference. Among rationales provided for preferring short forms, eight individuals (57.1%) indicated a general lack of understanding or misunderstanding of planned missingness, including not having the software or analytical capacity to implement it, and that using short forms is generally more intuitive. Two responses (14.3%) stated preferring using a short form because it has already been validated. Three responses (21.4%) expressed concern about the acceptance of the planned missingness procedure by reviewers, clients, external stakeholders, etc. Two responses (21.4%) preferred having the same set of items administered to all respondents. Among 33 researchers who provided a rationale for preferring implementing planned missingness, 13 (39.4%) suggested that having some data on all items is preferrable than having data on only a subset of items. Five individuals (15.2%) mentioned that planned missingness can result in better construct content coverage and prevent construct deficiency. One response (3.0%) mentioned higher reliability as a result of more items as an advantage of planned missingness. A few rationales centered specifically around the research scenario. Three researchers (15.2%) stated that planned missingness is appropriate for answering construct-level research questions such as in a convergent/discriminant validation study. One response (3.0%) highlights the MTurk sample pool as it affords a large sample size for a planned missingness design. Interestingly, two responses introduced methods of reducing survey length alternative to short forms and planned missingness. One researcher preferred planned missingness to short form but preferred dropping the number of constructs measures to planned missingness if given the choice. Another preferred scale-level rather than item-level planned missingness.

For scenario 2 (a test battery used to make selection decisions), 54 researchers (71.1%) preferred short forms and 10 (13.2%) preferred planned missingness. Of the 49 rationales given for the preference of using short forms, six (12.2%) mentioned a general lack of understanding of planned missingness, and 13 (26.5%) mentioned concern about others' negative reaction or unacceptance of planned missingness. However, the majority of the responses pointed to inappropriateness of planned missingness for this scenario specifically. 32 (65.3%) regarded administering the exact same items to all respondents particularly important in this scenario. 8 (16.3%) emphasized the high-stakes nature of this scenario, 18 (36.7%) specifically mentioned concerns about legal defensibility, and 6 (12.2%) mentioned issues of fairness. One response (2.0%) suggested that administering the same items to all candidates will allow investigation of item-level properties and the comparison of candidates at the item-level. Among the five rationales provided for preferring planned missingness in this scenario, two mentioned that using short forms may result in a lower reliability and subsequently a lower observed predictive validity. One researcher did not view the specific testing scenario as important for the decision of whether to use short forms or implement planned missingness. One response suggested that exposing each candidate to only a subset of all items can protect test security. One response viewed collecting some data on all items as important.

For scenario 3 (an engagement survey administered to employees), 26 (34%) preferred short forms and another 26 preferred planned missingness. For the 18 who provided a rationale for preferring short forms, six (33.3%) reflected a lack of understanding regarding planned missingness, and three (16.7%) expressed concern about others' acceptance of planned missingness. Eight (44.4%) responses pointed to administering different items to employees as potentially problematic for making item-level comparisons across employees or leading to negative reaction. Three responses (16.7%) referred to the nature of engagement surveys specifically, including their historically low response rates which when combined with planned missingness might lead to an overly high proportion of missing data. On the other hand, the 19 rationales provided for preferring planned missingness were spread evenly across the importance of having data on all items (21.1%), better construct content coverage (26.3%), suitability for construct-level comparisons across employees (26.3%), and the low-stakes measurement context specific to this scenario (26.3%).

Lastly, for scenario 4 (a concurrent validation study), 17 (22.7%) preferred short forms and 35 (46.7%) preferred planned missingness. Of the 15 rationales given for preferring short forms, six responses (40%) showed a lack of understanding in planned missingness and two (13.3%) were worried about others' acceptance of it. Five individuals (33.3%) viewed administering a different set of items to each respondent in a validation study as potentially indefensible legally. One response mentioned that given the nature of some common predictors that are validated, planned missing will be unsuitable for tests of maximal performance. One researcher expected that historically low response rate among incumbents means that planned missingness is not suitable with such a sample. Among the 20 rationales provided in support of the use of planned missingness, the majority viewed collecting some data on all items (60.0%) and having better construct content coverage (25.0%) as advantages, with the rest mentioning its appropriateness for examining construct-level research questions (5.0%), higher reliability (5.0%), and that it is methodologically more interesting (5.0%). One individual expressed that the legal implications with a concurrent validation study may be less severe than in a high-stakes selection setting, making planned missingness more appropriate.

Discussion

The present study explores researchers' general knowledge and perception of planned missingness. As results of studies 1 and 2 show that using short forms and implementing a planned missingness design for reducing study length are equivalently effective on average, the choice may depend on the research context or other factors. Thus, I wanted to qualitatively capture how researchers conceptualize the advantages and disadvantages of these two approaches.

Findings show that while researchers often have the need to reduce the length of a study, the most common approach to do so is to cut down the number of constructs measured, followed closely by using short forms of scales. Only a very small proportion of individuals have implemented planned missingness. This lack of usage is at least in part due to a lack of knowledge and familiarity, as fewer than half of the researchers indicated being familiar with the concept of planned missingness. However, even among those who were familiar, only a small proportion reported having implemented it, mostly in survey research studies as opposed to in operational studies. One respondent stated that they “don’t like purposely having missing data in [their] research studies,” which not only reflects a lack of knowledge in planned missingness, but a general misunderstanding of the broader missing data problem.

Other than a general lack of knowledge about the usefulness of planned missingness, researchers converge on several reasons for choosing not to implement a PM design, including 1) not having the statistical or software capacity to analyze data with a PM design, 2) concern about the methodological complication being confusing to upper management or slowing down the publication process, and 3) giving respondents different subsets of items being problematic in certain study setting.

The admitted inability to implement PM designs or analyzing data with PM is arguably the most easily solvable challenge. Most survey platforms have a point-and-click option to randomly present each respondent with a subset of items. For example, on Qualtrics, this involves adding Randomizer in survey flow (Qualtrics, 2021). With regards to analytical complications, there exist a number of published papers (e.g., Arbuckle & Marcoulides, 1996; Enders & Bandalos, 2001; Graham et al., 2007; Newman, 2003; Vink & van Buuren, 2014; Yuan, 2000) and tutorial resources (van Buuren, 2018; Vink & van Buuren, 2011) that detail common analyses for data with PM. Further, various statistical packages in common software have been developed to make FIML estimation and multiple imputation procedures accessible. For example, in R, *lavaan* makes available the FIML estimator for structural equation modeling-type analyses (Rosseel et al., 2012). The R packages mice (van Buuren & Groothuis-Oudshoorn, 2010) and hmi (Speidel et al., 2020) provide a diverse and flexible range of multiple imputation capabilities.

Secondly, the concerns regarding how data collected with a PM design would be perceived by management, clients, or throughout the publication process are understandable. After all, using a shorter version of a scale is much more intuitive and familiar than implementing any PM design and the analytical procedures that accompany it. However, I believe that any statistical or analytical method now considered standard and commonly accepted by management or clients have become so from being unfamiliar at first. With thorough research supporting its effectiveness and its efficiency, PM should be introduced to and used by stakeholder who can stand to benefit from it.

In response to worry about how PM would affect the publication process, I turned to scientists in other disciplines who have published substantive research with PM designs. I conducted a search on Google Scholar using the key phrase "planned missingness," and emailed 42 corresponding authors who have published with a PM design in the past five years. These researchers spanned a variety of expertise areas, from substance use and addictions, tourism management, clinical and developmental psychology, to education. I briefly introduced my studies, described major findings, and asked for any tips or advice, or simply their experience in understanding, applying, and/or publishing with PM as "lessons learned" that may be able to guide other researchers. I received nine very helpful replies. Overall, researchers have not had much pushback from reviewers regarding their use of PM, particularly at quantitative and methodologically advanced journals. They recommended alleviating any skepticism due to unfamiliarity by providing more detailed explanation of the methodology (i.e., devoting extra space to explain the underworking of the method and why it is useful, including a supplemental appendix describing the approach, and citing established research on the topic).

Last but not least, the respondents of the current study pointed out that PM may not be appropriate in all data collection settings, given that different items are given to each individual. The second part of this study explores this point precisely. Armed with at least a basic understanding of the concept and the assumption that it is technically comparable with using short forms, researchers were asked to choose between the two for each of four data collection scenarios. Findings show that preference between using short forms and planned missingness does vary depending on the type of data collection being conducted. Overall, for the two research scenarios—both an MTurk study investigating a new personality scale as well as a concurrent validation study using incumbents—approximately twice as many researchers preferred planned missingness to short forms than vice versa. For an internal engagement survey, an operational study using incumbents, preferences were split evenly between short forms and PM. Lastly, the majority of researchers preferred using short forms to PM in an operational selection test battery using applicants.

Many rationales provided for preferring short forms to planned missingness in all four scenarios reiterated a lack of knowledge in implementation and analyses of PM designs and worry over how it might be perceived by others. However, some context-specific rationales highlighted the strengths and limitations of PM in different settings. For example, while the importance of standardization was mentioned as an advantage of using short forms over PM for all scenarios, it was mentioned at the highest frequency by far for the selection battery context. Researchers elaborated the importance of this with issues of fairness, legal defensibility of selection decisions based on different test items, and the need for comparability across individuals' results. The same issue in the engagement survey context surrounded the potential of negative employee reaction if they were presented engagement results on items that were never given to them or discover that they received different items than their coworkers, as well as the frequent need to report and present item-level analyses.

On the other hand, having some data on all items and not limited to items included in the short form was mentioned as a reason for preferring planned missingness across all scenarios, particularly for the two research scenarios. Researchers mentioned that although short forms are usually validated, they do not always preserve the full construct coverage and may suffer from construct deficiency. PM would ensure full construct coverage reflected in the full measures.

Interestingly, researchers mentioned that the two scenarios that target incumbent samples (engagement survey and concurrent validation study) face historically low response rate, which might make implementing PM designs more difficult. On the other hand, the large pool of MTurk workers was mentioned as a factor that could enable proper implementation of planned missingness.

Overall, although there are research scenarios for which PM designs are not suitable for good reasons, the majority of I/O researchers surveyed have not deliberately chosen to use an approach other than PM for reducing survey length. Rather, its rare use in I/O stems primarily from a lack of familiarity with the technique or an inaccurate understanding of it.

General Discussion

In a field where surveys are one of the most common study methodologies, I/O researchers often face the problem of overly lengthy surveys. Standard solutions include either dropping some of the measurement scales that are less central to the questions of interest or using short forms rather than full versions of instruments. While the former compromises the number of constructs measured and is an unideal option, the latter does measure each construct with some items. However, the approach of using short forms relies on the existence or development of short forms. In the event of working with measures with no short versions, much time and resources needed to be dedicated to developing short forms first, including collecting additional samples, evaluating items against psychometric criteria, and validating the chosen items. I conducted a series of studies to examine an alternative and much more convenient approach to reduce study length, namely planned missingness designs.

Study 1 replicated Yoon and Sackett (2016) and compared using short forms and implementing planned missingness in two additional public datasets. Findings were consistent across datasets such that planned missingness yielded smaller deviations from true population estimates than short forms, regardless of whether short forms were created based on factor loadings alone or based on multiple psychometric criteria using OASIS (Cortina et al., 2020). These results suggested initial promise for the effectiveness of planned missingness for shortening study length.

Study 2 then systematically simulated data with varying characteristics and compared short forms and planned missingness under different research conditions. Overall, the two approaches performed similarly and resulted in estimates with small deviations from population truths. However, each showed slight advantages over the other in different conditions. When empirically based short forms already exist for use or information needed to readily compile the short forms can be found in prior studies, short forms yielded slightly more accurate estimates than planned missingness on average of all the conditions simulated. When there are no existing short forms or psychometric information to create short forms, and part of the sample needs to be used to first develop short forms, the two approaches performed equally. Lastly, when short forms are created not strictly based on factor loadings, planned missingness performed slightly better than short forms on average. Across all conditions examined, planned missingness performed poorly when a high missingness level was combined with a small sample size, even experiencing imputation failures at extreme conditions. This is consistent with Zhang and Yu (In press), who reported similar convergence failure issues when treating planned missingness data with full-information maximum likelihood estimation. However, barring extreme conditions, results present planned missingness as a viable alternative to using short forms when the length of a survey needs to be reduced.

Across the different conditions simulated in Study 2, short form A was most similar to the short forms developed in Study 1, such that short form items are selected strictly based on factor loadings, and that short form and planned missingness estimates were computed based on the same sample. Whereas Study 1 found that PM produced slightly more accurate intercorrelation estimates than short forms, Study 2 found that a small advantage favoring short forms when they have already been developed. Although a small discrepancy, it highlights the relative benefits and drawbacks of both using real data and conducting a simulation. In Study 1, the three large public datasets do exhibit some characteristics that may be more realistic compared to data simulated in Study 2 in that the number of constructs varied more (included beyond five scales that were simulated in Study 2), the number of items varied across constructs (rather than being fixed at 10 in Study 2), and the factor loadings of the items onto each construct varied across scales (rather than being fixed at .30~.75 in Study 2). At the same time, datasets with very large samples were chosen to minimize sampling error. Having a large sample size is particularly important for the effectiveness of planned missingness, potentially explaining the findings of Study 1. However, it is important to keep in mind that although not impossible, collecting a sample of 100,000 or larger for a single survey is rare. Compared with only three datasets included in Study 1 which represented a very limited and possibly idiosyncratic combination of data characteristics, Study 2 captured a broader universe of data that could be observed. Thus, results from Studies 1 and 2 should be interpreted alongside each other. Although either using short forms or implementing a planned missingness design may prevail slightly in any one study, differences are likely to be small on average and decisions on which method to choose may ultimately depend on practical and contextual considerations rather than statistical.

In the effort to examine some of these practical considerations, Study 3 captured common reasons why planned missingness has not been used more in I/O and related fields. Findings show that the lack of use has mainly resulted from a lack of awareness and knowledge of what it is, how to implement it, and how to analyze the data. In addition, researchers worry about how studies with a PM design would be received by others, including reviewers, management, clients, etc. Lastly, results highlight contexts in which planned missingness would not be suitable, such as when the purpose of data collection is for making high stakes, operational decisions, when a higher level of unplanned missingness is expected, and when negative reactions among respondents is particularly worrisome.

Limitations of planned missingness

While planned missingness can be a very useful tool in some scenarios, it is important to keep in mind its limitations.

First, as data gathered with a PM design need to be subsequently treated with either multiple imputation or maximum likelihood estimation, the constraints of such statistical procedures apply to the survey design itself. Therefore, in cases where imputation or FIML estimation fails or yield inaccurate estimates, PM should not be used. This could occur when there is not enough information in the observed data from which to impute or produce a covariance structure for proper estimation, due to the combination of a large amount of data being missing and an inadequate sample size. Luckily, these two factors are determined by the judgment of the researcher for the most part. When planning to implement a PM design, the researcher should be prepared to gather a reasonably large dataset and be cognizant of the level of missingness designed. The specific sample size and missingness level will vary depending on the length of original scales, target response time, expectation of completion rate and any unplanned missingness, and the type of statistical analyses planned. In the event where certain factors have hard constraints (e.g., each respondent only has time for i items; it is only possible to have access to n respondents), these constraints should be taken into consideration when determining the other components of the study design. When in doubt, it might be helpful to conduct a pilot study to estimate average response time and determine the amount of missingness and sample size that are suitable accordingly.

Second, as some respondents in Study 3 expressed when asked why they have not implemented planned missingness, it can complicate analyses. Even researchers in other fields who have published with PM designs warned about this challenge, particularly when used in a multi-level study (Folberg, 2020; Lüdtke et al., 2016; Westfall et al., 2015; Wood et al., 2019). Research has detailed the effectiveness of different PM designs and approaches to impute data in cases of multiple measurement, but researchers should be prepared for the added complexities.

Third, as planned missingness assigns a different subset of items to each respondent by design, it is not suitable for all data collection purposes. As the respondents in Study 3 pointed out, in high stakes testing situations, this may cause issues for fairness and legal defensibility and result in negative applicant reaction. Further, in research that collects data from incumbents which typically suffer from low response and completion rates, further implemented planned missingness might lead to overly sparse data.

Overall, planned missingness is not argued to be a substitute for using short forms to reduce survey length. It is a valuable and convenient alternative in some contexts, but not in situations outlined above. When planned missingness is not a viable option or well developed and validated short forms are readily available for use, the standard practice of using short version is still recommended. Short forms were created solely based on factor loadings in the simulation study to reduce processing time and power, but a number of psychometric criteria can also be used to select items for the short form. For researchers who need to develop their own short forms, I recommend following the prescriptions provided by Cortina et al. (2020) and using their R Shiny app, OASIS, to ensure simultaneous consideration of reliability, content validity, and construct validity.

Practical Recommendations for Using Planned Missingness

Although not without limitations, implementing planned missingness designs can pose a number of advantages over using short forms, which can guide researchers' decisions about when and how to use PM.

I summarize four conditions that when fulfilled, would characterize a good opportunity to implement PM.

1. **There is a need to reduce study length.** PM can be very effective for producing accurate estimates. However, if there is no length concern on the number of items included in the study and participant response burden is estimated to be low, there is no need for any method of length reduction.

2. **The study is designed for low stakes, research purposes.** Administering a different subset of items in a high-stakes setting for purposes of selection or promotion decisions can lead to issues of fairness across individuals, legal defensibility, and negative applicant reactions. On the other hand, for a research study whose purpose is not to make any individual decisions but to advance understanding of certain psychological constructs, PM designs are a useful tool for reducing study length, thus minimizing participant burden and improving measurement efficiency and data quality.

3. **Short forms of measures have not been previously developed and validated.** If empirically based short versions of the measures that a researcher hopes to use have previously been developed, then no additional researcher effort or participant numbers

are needed for scale development. In such cases, using short forms is a perfectly fine approach to reduce study length. However, when short forms have yet to be developed or are developed at least in part based on human judgment, implementing PM is much more inexpensive and convenient and can produce more accurate estimates.

4. **An adequate sample size is expected OR there is unlikely to be a high level of unplanned missingness (ideally both).** Overall, the effectiveness of PM designs depends on the observed data providing enough information that can allow subsequent imputation or maximum likelihood estimation. Either a limited availability of recruitment pool or a large amount of unplanned missing data will both hurt the effectiveness of PM designs. Thus, when facing a large sample (e.g., MTurk workers or incumbents of a high volume position), or when there is expectation of reasonable response and completion rate, planned missingness can be a great option.

5. **You have the methodological and analytical expertise (or are willing to learn).** PM may be new to many but for many purposes, learning to use PM does not require a big time investment. However, it is worth noting that for some complex analyses, more expertise is needed. As some of the researchers who have published substantive research with a PM design expressed, PM does have the potential to severely complicate analyses, particularly when testing multi-level research questions or structural equation models that are complex to begin with (Folberg, 2020; Lüdtke et al., 2016; Westfall et al., 2015; Wood et al., 2019). Research has detailed the effectiveness of different PM designs and approaches to impute data in cases of multiple measurement, but users should be prepared for the added complexities.

Conclusion

The technical effectiveness of planned missingness designs found in the current dissertation along with the practical convenience of their implementation have implications that stem beyond the field of I/O and are relevant to all survey research that assesses latent constructs. I hope that this series of studies can help researchers better understand planned missingness designs, better distinguish planned missingness from traditional types of missing data that may be more problematic, and move from reacting to to anticipating missing data. I also hope that by demonstrating the conditions under which planned missingness is useful and appropriate and those under which it is not, it can become simply another methodological tool in our belt.

Table 1. Three-Form PM Design at 30% Missingness

		Item Set									
		X	A			B			C		
	Respondent	1	2	3	4	5	6	7	8	9	10
Form 1	1	1	1	1	1	1	1	1	0	0	0
	2	1	1	1	1	1	1	1	0	0	0
	3	1	1	1	1	1	1	1	0	0	0
Form 2	4	1	1	1	1	0	0	0	1	1	1
	5	1	1	1	1	0	0	0	1	1	1
	6	1	1	1	1	0	0	0	1	1	1
Form 3	7	1	0	0	0	1	1	1	1	1	1
	8	1	0	0	0	1	1	1	1	1	1
	9	1	0	0	0	1	1	1	1	1	1

Note. 1 = item administered. 0 = item not administered.

Table 2. Random Percentage PM Design at 30% Missingness

	Items									
Respondent	1	2	3	4	5	6	7	8	9	10
1	0	1	0	1	1	0	1	1	1	1
2	1	0	1	1	1	1	0	1	1	0
3	1	1	1	0	0	1	1	1	0	1
4	0	1	1	1	1	1	1	0	0	1
5	1	1	1	0	0	0	1	1	1	1
6	1	1	0	1	1	1	0	0	1	1
7	0	0	1	1	1	1	1	1	1	0
8	1	1	1	1	1	0	1	0	1	0
9	0	1	1	1	1	1	0	1	0	1

Note. 1 = item administered. 0 = item not administered.

Table 3. Scale Internal Consistency in Study 1 Dataset A

Variable	Full	Loadings-based SF	OASIS-based SF	PM
Machiavellianism	.86	.84	.84	.86
Narcissism	.80	.74	.75	.80
Psychopathy	.80	.79	.76	.79

Notes. SF = short form. OASIS = Optimization App for Selecting Item Subsets. PM = planned missingness.

Table 4. Scale Internal Consistency in Study 1 Dataset B

Scale	Full	Loadings-based SF	OASIS-based SF	PM
Realistic	.88	.84	.84	.88
Investigative	.89	.88	.86	.90
Artistic	.86	.75	.80	.87
Social	.85	.81	.81	.85
Enterprising	.83	.75	.75	.84
Conventional	.90	.84	.84	.91

Notes. SF = short form. OASIS = Optimization App for Selecting Item Subsets,.PM = planned missingness.

Table 5. Scale Internal Consistency in Study 1 Dataset C

Scale	Full	Loadings-based SF	OASIS-based SF	PM
Extraversion	.88	.79	.79	.88
Agreeableness	.81	.75	.75	.83
Conscientiousness	.78	.68	.70	.79
Emotional Stability	.87	.84	.82	.83
Openness to Experience	.80	.74	.74	.75
CWB (Laczo)	.75	.67	.72	.74
CWB (B&R)	.79	.71	.72	.78
OCB	.82	.75	.74	.80

Notes. SF = short form. OASIS = Optimization App for Selecting Item Subsets. PM = planned missingness. CWB = counterproductive work behaviors. OCB = organizational citizenship behaviors. Laczo = Laczo (2002). B&R = Bennett & Robinson (2000).

Table 6. Scale Intercorrelation in Study 1 Dataset A

					Absolute Difference from Full		
Scale Pair	Full	Loadings-based SF	OASIS-based SF	PM	Loadings-based SF	OASIS-based SF	PM
Machiavellianism–Narcissism	.47	.50	.50	.47	.02	.02	.00
Machiavellianism–Psychopathy	.72	.79	.72	.71	.07	.01	.01
Narcissism–Psychopathy	.45	.51	.48	.45	.06	.03	.00
Mean absolute difference					.05	.02	.01
SD absolute difference					.03	.01	.01

Notes. SF = short form. OASIS = Optimization App for Selecting Item Subsets. PM = planned missingness. SD = standard deviation

Table 7. Scale Intercorrelation in Study 1 Dataset B

					Absolute Difference from Full		
Scale Pair	Full	Loadings-based SF	OASIS-based SF	PM	Loadings-based SF	OASIS-based SF	PM
Realistic–Investigative	.30	.25	.23	.30	.05	.07	.00
Realistic–Artistic	.19	.18	.18	.19	.01	.01	.00
Realistic–Social	.09	.06	.09	.10	.02	.00	.01
Realistic–Enterprising	.32	.22	.22	.31	.10	.10	.01
Realistic–Conventional	.48	.34	.34	.48	.13	.14	.00
Investigative–Artistic	.32	.27	.25	.32	.06	.07	.00
Investigative–Social	.22	.15	.15	.22	.06	.07	.00
Investigative–Enterprising	.07	-.01	.00	.07	.08	.07	.00
Investigative–Conventional	.11	.10	.13	.12	.02	.02	.00
Artistic–Social	.34	.27	.30	.33	.06	.04	.01
Artistic–Enterprising	.31	.21	.18	.30	.10	.13	.01
Artistic–Conventional	.04	.02	.04	.04	.02	.00	.00
Social–Enterprising	.40	.27	.31	.39	.13	.09	.01
Social–Conventional	.20	.14	.16	.19	.06	.04	.00
Enterprising–Conventional	.51	.39	.39	.50	.12	.12	.01
Mean absolute difference					.07	.07	.01
SD absolute difference					.04	.04	.01

Notes. SF = short form. OASIS = Optimization App for Selecting Item Subsets. PM = planned missingness. SD = standard deviation

Table 8. Scale Intercorrelation in Study 1 Dataset C

					Absolute Difference from Full		
Scale Pair	Full	Loadings-based SF	OASIS-based SF	PM	Loadings-based SF	OASIS-based SF	PM
EXT–AGR	.30	.13	.13	.27	.17	.15	.03
EXT–CON	.04	.02	.02	.05	.02	.04	.01
EXT–ES	.22	.09	.09	.22	.13	.08	.00
EXT–OPN	.25	.23	.23	.24	.02	.02	.01
EXT–CWB (Laczo)	-.06	.06	.06	-.09	.12	.05	.03
EXT–CWB (B&R)	-.11	-.05	-.05	-.10	.06	.03	.01
EXT–OCB	.29	.24	.24	.28	.05	.13	.01
AGR–CON	.27	.20	.20	.25	.07	.03	.02
AGR–ES	.19	.10	.10	.29	.09	.09	.10
AGR–OPN	.18	.18	.18	.18	.00	.01	.00
AGR–CWB (Laczo)	-.26	-.13	-.13	-.28	.13	.11	.02
AGR–CWB (B&R)	-.30	-.20	-.20	-.29	.10	.14	.01
AGR–OCB	.38	.40	.40	.36	.02	.01	.02
CON–ES	.24	.24	.24	.30	.00	.04	.06
CON–OPN	.04	.08	.08	.05	.04	.08	.01
CON–CWB (Laczo)	-.34	-.31	-.31	-.39	.03	.05	.05
CON–CWB (B&R)	-.41	-.29	-.29	-.40	.12	.07	.01
CON–OCB	.29	.28	.28	.28	.01	.03	.01
ES–OPN	.02	.01	.01	.02	.01	.03	.00
ES–CWB (Laczo)	-.28	-.25	-.25	-.33	.03	.01	.05
ES-CWB (B&R)	-.31	-.31	-.31	-.33	.00	.01	.02
ES–OCB	.20	.22	.22	.25	.02	.00	.05

OPN–CWB (Laczo)	-.04	-.01	-.01	-.04	.03	.03	.00
OPN–CWB (B&R)	-.06	-.04	-.04	-.08	.02	.01	.02
OPN–OCB	.31	.27	.27	.30	.04	.04	.01
CWB (Laczo)–CWB (B&R)	.80	.66	.66	.79	.14	.05	.01
CWB (Laczo)–OCB	-.26	-.13	-.13	-.31	.13	.12	.05
CWB (B&R)–OCB	-.31	-.24	-.24	-.33	.07	.11	.02
Mean absolute difference					.06	.06	.02
SD absolute difference					.05	.04	.02

Notes. SF = short form. OASIS = Optimization App for Selecting Item Subsets. PM = planned missingness. CWB = counterproductive work behaviors. OCB = organizational citizenship behaviors. Laczo = Laczo (2002). B&R = Bennett & Robinson (2000). SD = standard deviation.

Table 9. Variables manipulated in Study 2

Variables manipulated	Range
Population characteristics	
Number of scales (s)	2-5
Intercorrelation mean (M_r)	0-.70 at .10 increments
Intercorrelation standard deviation (SD_r)	0-.30 at .05 increments (only for 3-5 scales)
Sample characteristics	
Sample size (n)	100 to 1,000 at 100 increments
Amount of missingness (m)	10% to 80% at 10% increments

Table 10. Study 2 Simulation Regressions for Absolute Deviations

	DV1: Abs(truth-SFA)		DV2: Abs(truth-SFB)		DV3: Abs(truth-SFC)		DV4: Abs(truth-PM)	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
(Intercept)	.039** (.00)	.039** (.00)	.056** (.00)	.056** (.00)	.072** (.00)	.072** (.00)	.053** (.00)	.053** (.00)
m	.023** (.00)	.023** (.00)	.036** (.00)	.036** (.00)	.144** (.00)	.144** (.00)	.074** (.00)	.079** (.00)
n	-.004** (.00)	-.004** (.00)	-.007** (.00)	-.007** (.00)	-.003** (.00)	-.003** (.00)	-.007** (.00)	-.007** (.00)
s	.000 (.00)	.000 (.00)	.000 (.00)	.000 (.00)	.000 (.00)	.000 (.00)	.004** (.00)	.005** (.00)
M_r	-.005** (.00)	-.006** (.00)	-.004** (.00)	-.004** (.00)	.135** (.00)	.136** (.00)	.023** (.00)	.030** (.00)
SD_r	.000 (.00)	.000 (.00)	-.000 (.00)	.000 (.00)	.000 (.00)	.001 (.00)	.005** (.00)	.007** (.00)
$m \times n$		.001** (.00)		-.002** (.00)		.004** (.00)		-.010** (.00)
$m \times s$		.000* (.00)		.000 (.00)		.000 (.00)		.022** (.00)
$m \times M_r$		.114** (.00)		.178** (.00)		.600** (.00)		.248** (.00)
$m \times SD_r$		.008** (.00)		.011** (.00)		-.004** (.00)		.035** (.00)
$n \times s$		.000 (.00)		.000 (.00)		.000 (.00)		-.001** (.00)
$n \times M_r$		.004** (.00)		.002** (.00)		.007** (.00)		-.012** (.00)
$n \times SD_r$		.000** (.00)		.000 (.00)		.000* (.00)		.000** (.00)
$s \times M_r$		.000 (.00)		.000 (.00)		.000 (.00)		.024** (.00)
$s \times SD_r$		.000 (.00)		.000 (.00)		.000 (.00)		.009** (.00)
$M_r \times SD_r$		-.004** (.00)		-.001 (.00)		.060** (.00)		.036** (.00)
R^2	.306	.347	.337	.373	.537	.705	.349	.449

Notes. SF = short form, PM = planned missingness, DV = dependent variable, *m* = amount of missingness, *n* = sample size, *s* = number of scales, M_r = true intercorrelation mean, SD_r = true intercorrelation standard deviation

Table 11. Study 2 Simulation Regressions for Differences in Absolute Deviations

	DV5: DV1-DV4		DV6: DV2-DV4		DV7: DV3-DV4	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
(Intercept)	-.014** (.00)	-.015** (.00)	-.002** (.00)	.001** (.00)	.018** (.00)	.017** (.00)
m	-.055** (.00)	-.061** (.00)	-.048** (.00)	-.054** (.00)	.065** (.00)	.056** (.00)
n	.003** (.00)	.004** (.00)	.001** (.00)	.001** (.00)	.004** (.00)	.005** (.00)
s	-.004** (.00)	-.005** (.00)	-.004** (.00)	-.005** (.00)	-.004** (.00)	-.005** (.00)
M_r	-.030** (.00)	-.037** (.00)	-.033** (.00)	-.039** (.00)	.012** (.00)	.103** (.00)
SD_r	-.005** (.00)	-.007** (.00)	-.006** (.00)	-.007** (.00)	-.004** (.00)	-.006** (.00)
$m \times n$		.014** (.00)		.013** (.00)		.018** (.00)
$m \times s$		-.022** (.00)		-.023** (.00)		-.022** (.00)
$m \times M_r$		-.140** (.00)		-.101** (.00)		.336** (.00)
$m \times SD_r$		-.027** (.00)		-.025** (.00)		-.039** (.00)
$n \times s$		.001** (.00)		.001** (.00)		.001** (.00)
$n \times M_r$		.016** (.00)		.016** (.00)		.019** (.00)
$n \times SD_r$		.000 (.00)		.000 (.00)		-.001** (.00)
$s \times M_r$		-.024** (.00)		-.025** (.00)		-.025** (.00)
$s \times SD_r$		-.008** (.00)		-.008** (.00)		-.008** (.00)
$M_r \times SD_r$		-.044** (.00)		-.044** (.00)		.021** (.00)
R^2	.202	.357	.116	.238	.317	.509

Notes. SF = short form, PM = planned missingness, DV = dependent variable, m = amount of missingness, n = sample size, s = number of scales, M_r = true intercorrelation mean, SD_r = true intercorrelation standard deviation

Table 12. Frequency of Survey Responses in Study 3

Response	In your work, approximately how many self-report data collections (e.g., surveys, test batteries) have you designed or contributed to designing?	Approximately how many of these self-report data collections (e.g., surveys, test batteries) have you had the need to reduce their length?
0	0	4
1-5	9	31
6-10	15	26
11-15	8	5
15-20	7	0
>20	49	16

Table 13. Preference between SF and PM in Study 3

Scenario	I have no preference	I would prefer using short form	I would prefer implementing planned missingness
Scenario 1: An Amazon Mechanical Turk research study examining the convergent and discriminant validity of a new personality measure by administering the new measure along with a number of other personality scales	21	14	41
Scenario 2: A battery of tests is administered to applicants of an entry-level job and used to make selection decisions	12	54	10
Scenario 3: An engagement survey is being designed to evaluate job attitudes and perceptions of organizational norms and culture internally	24	26	26
Scenario 4: A number of new selection tools are being developed and validated, and they are administered to incumbents for validation research purposes only	23	17	35

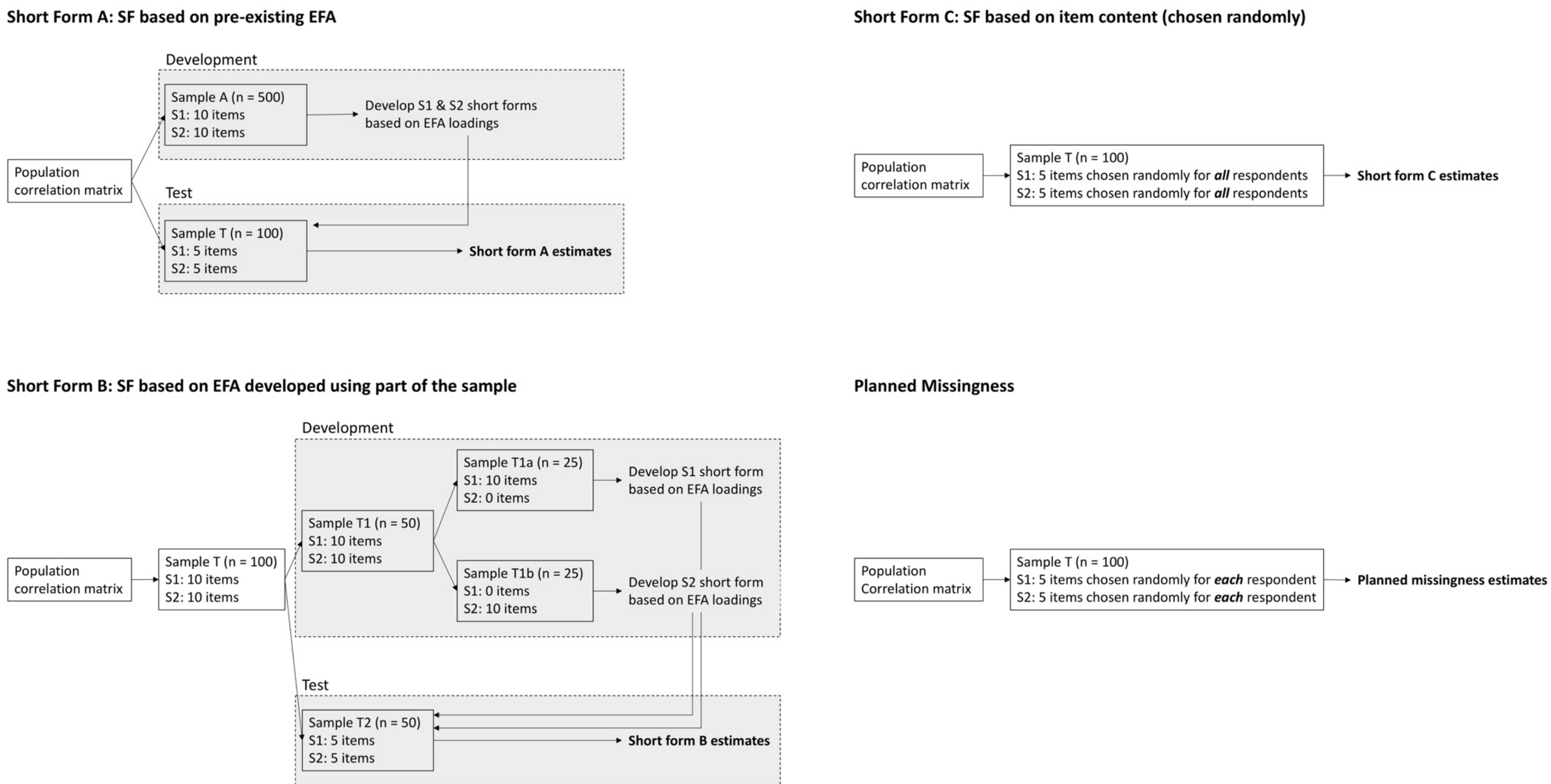

Figure 1. Simulation procedure in Study 2.

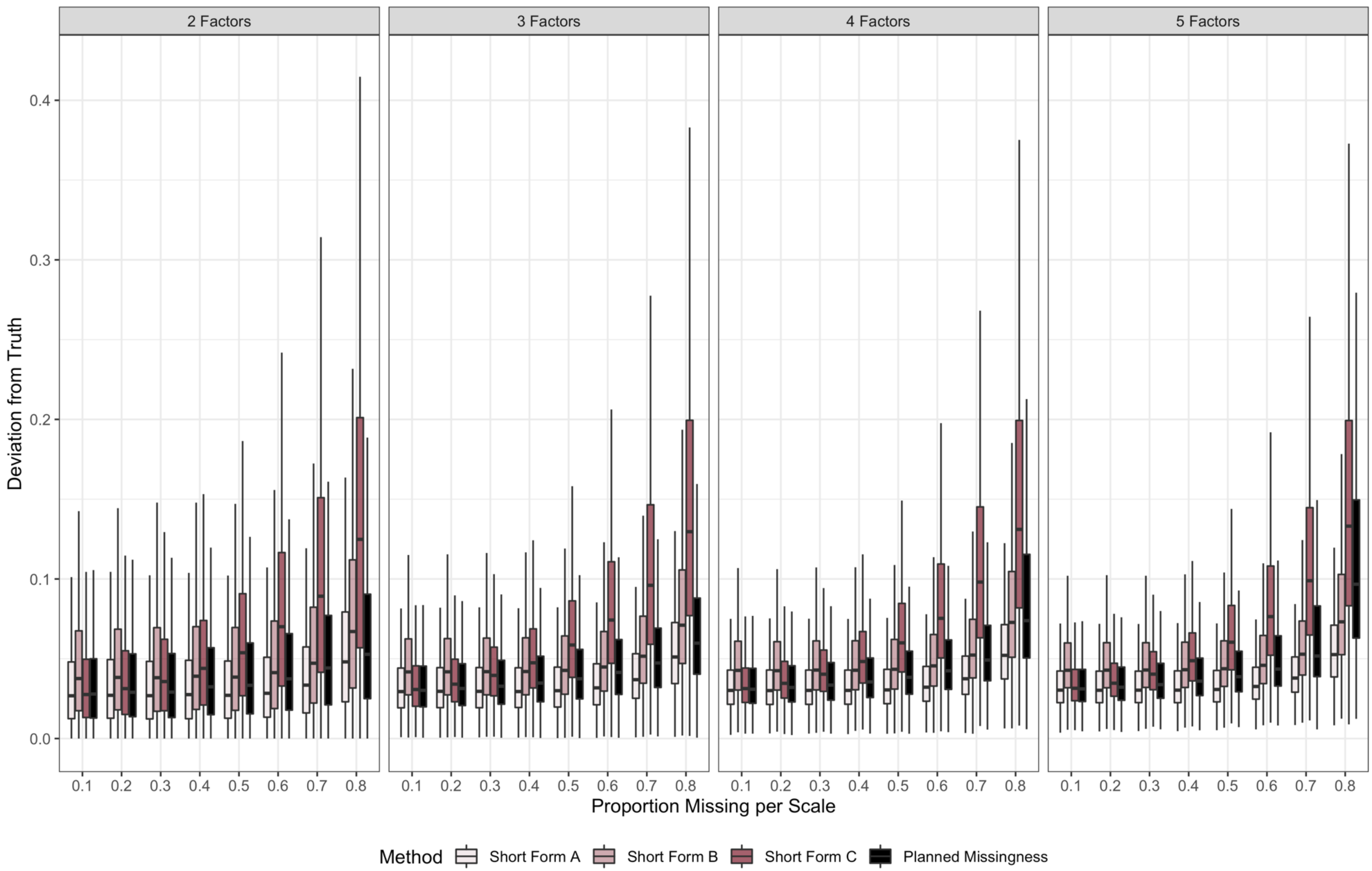

Figure 2. Deviations from True Population Intercorrelation by Method, Amount of Missingness, and Number of Factors

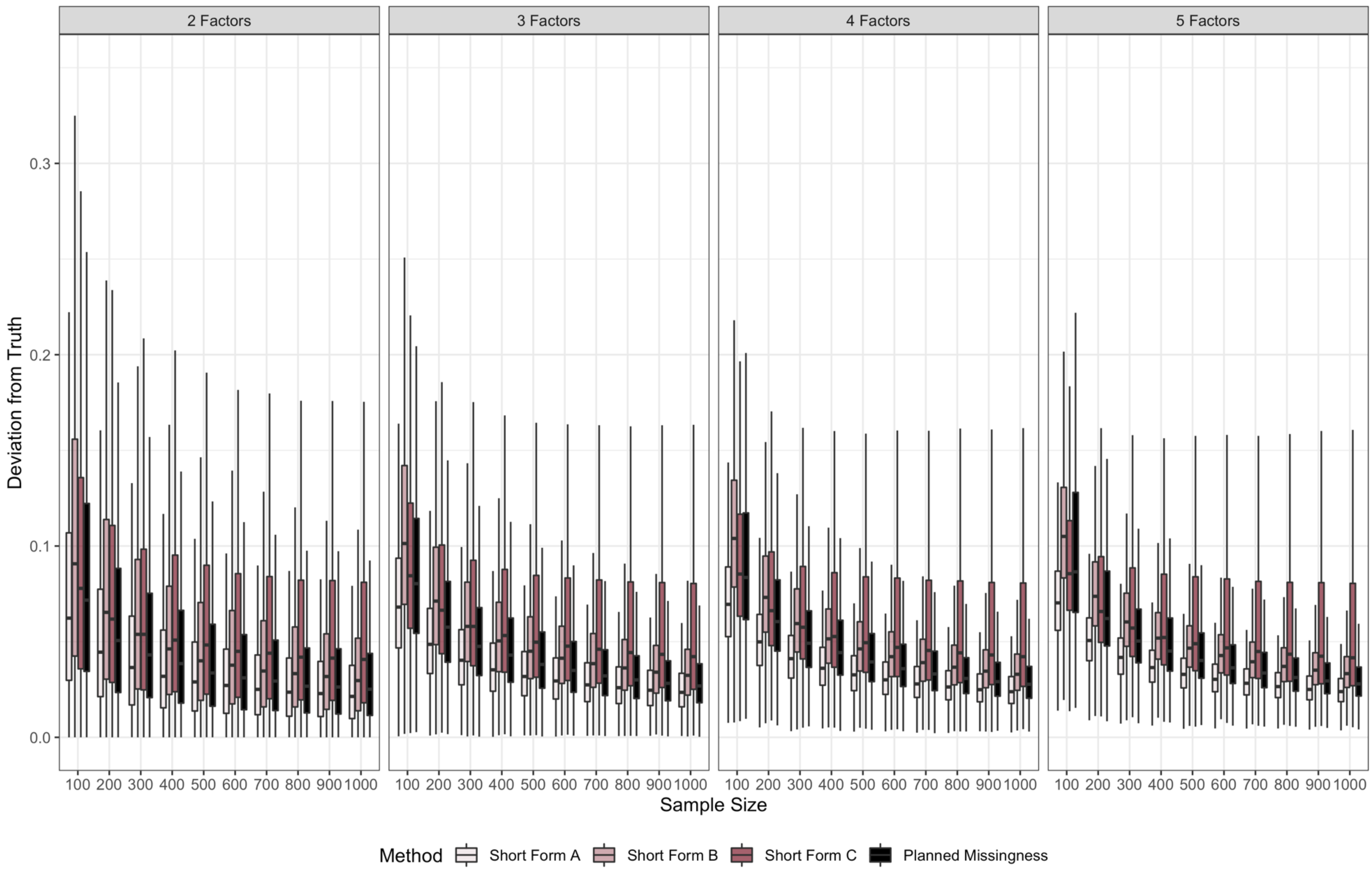

Figure 3. Deviations from True Population Intercorrelation by Method, Sample Size, and Number of Factors

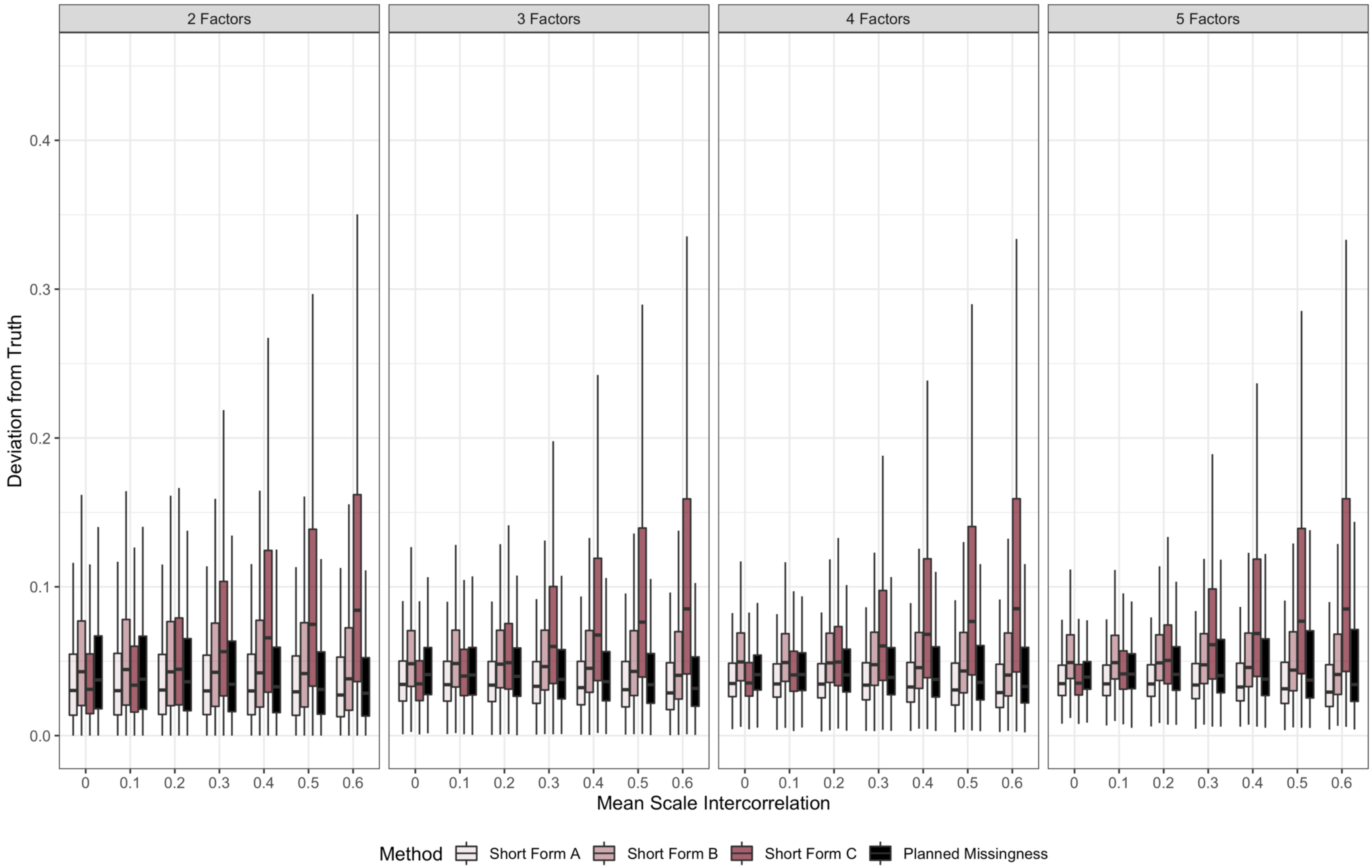

Figure 4. Deviations from True Population Intercorrelation by Method, Mean Intercorrelation, and Number of Factors

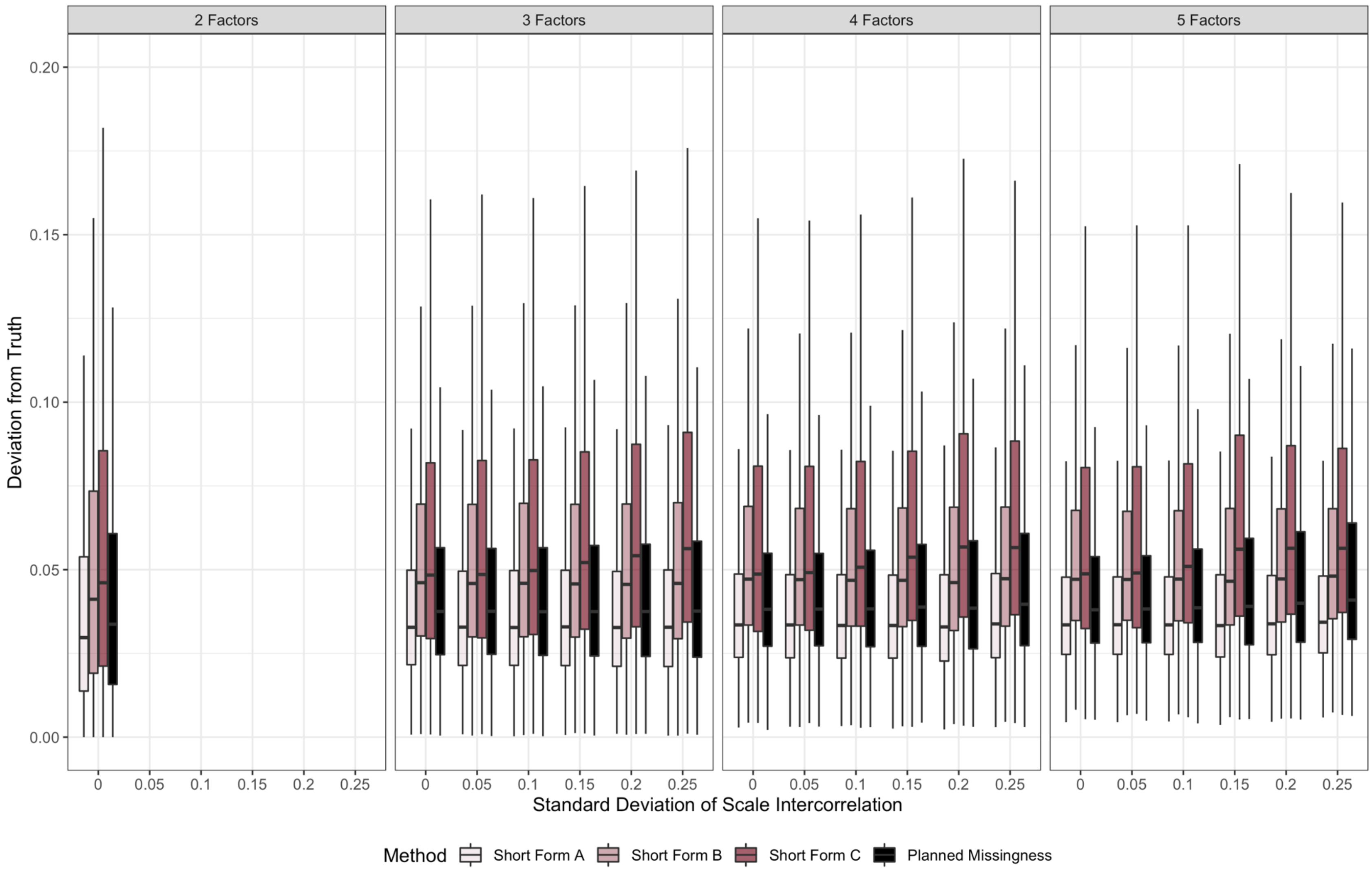

Figure 5. Deviations from True Population Intercorrelation by Method, Intercorrelation Standard Deviation, and Number of Factors

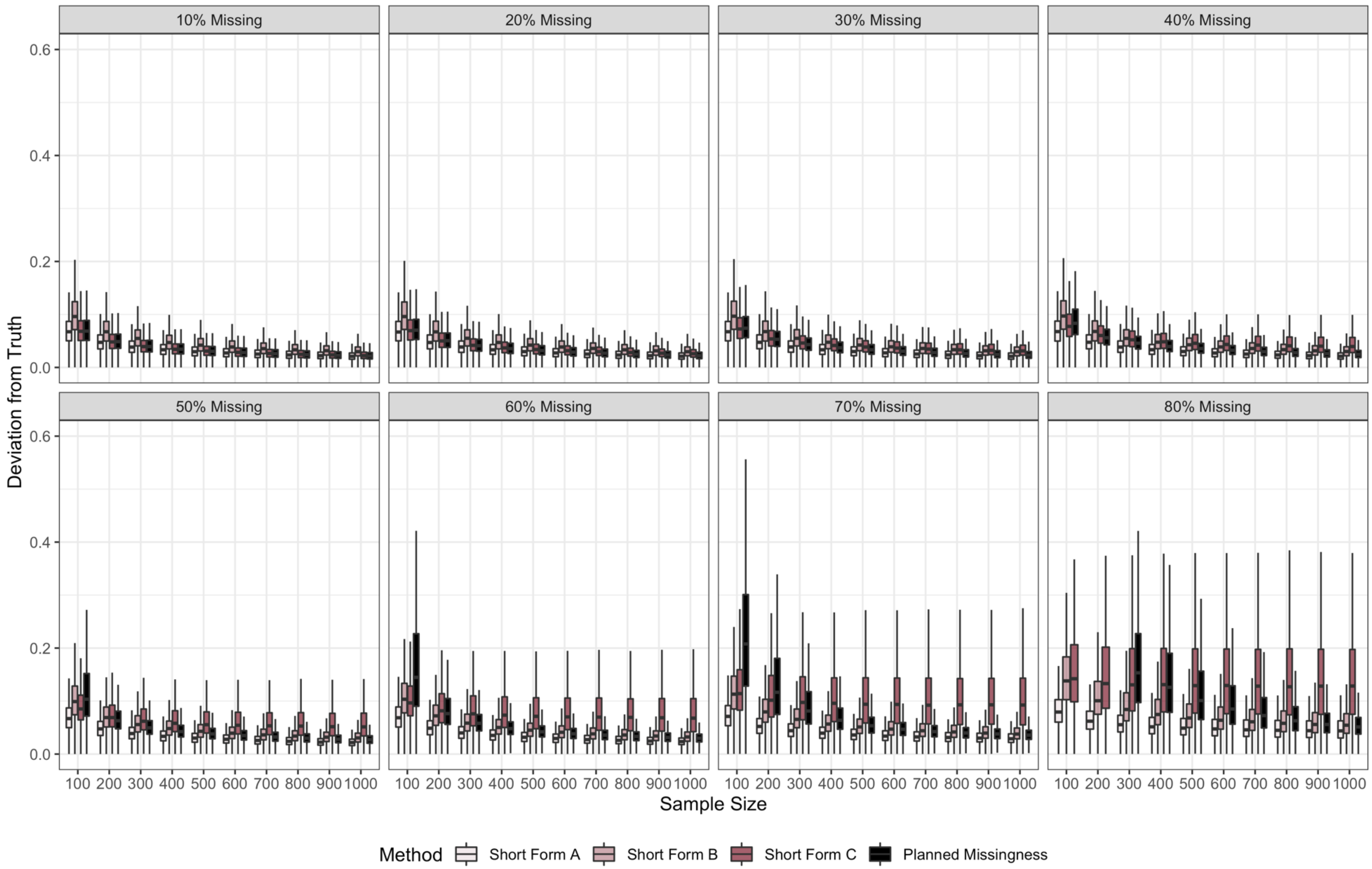

Figure 6. Deviations from True Population Intercorrelation by Method, Sample Size, and Amount of Missingness

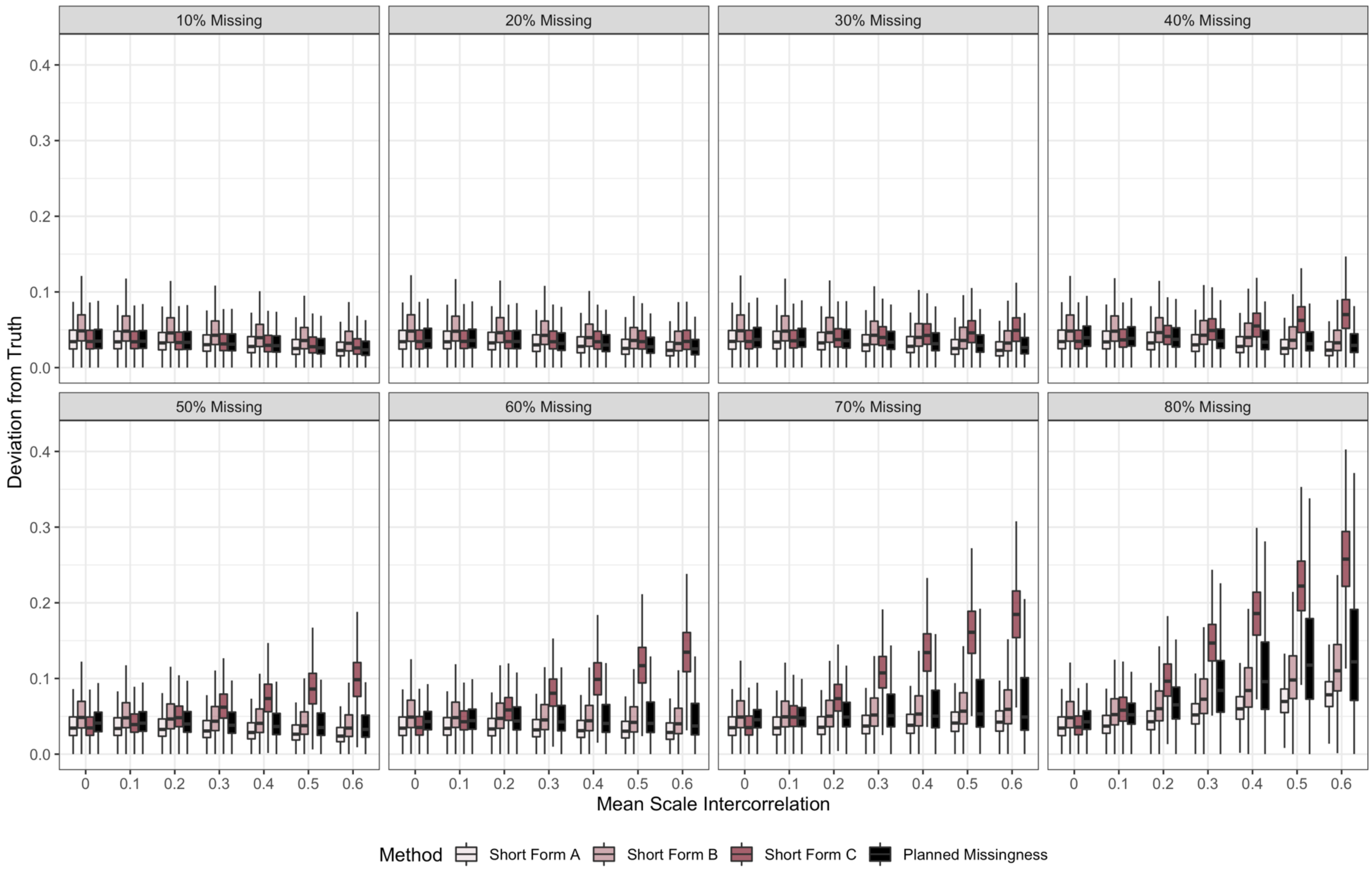

Figure 7. Deviations from True Population Intercorrelation by Method, Mean Intercorrelation, and Amount of Missingness

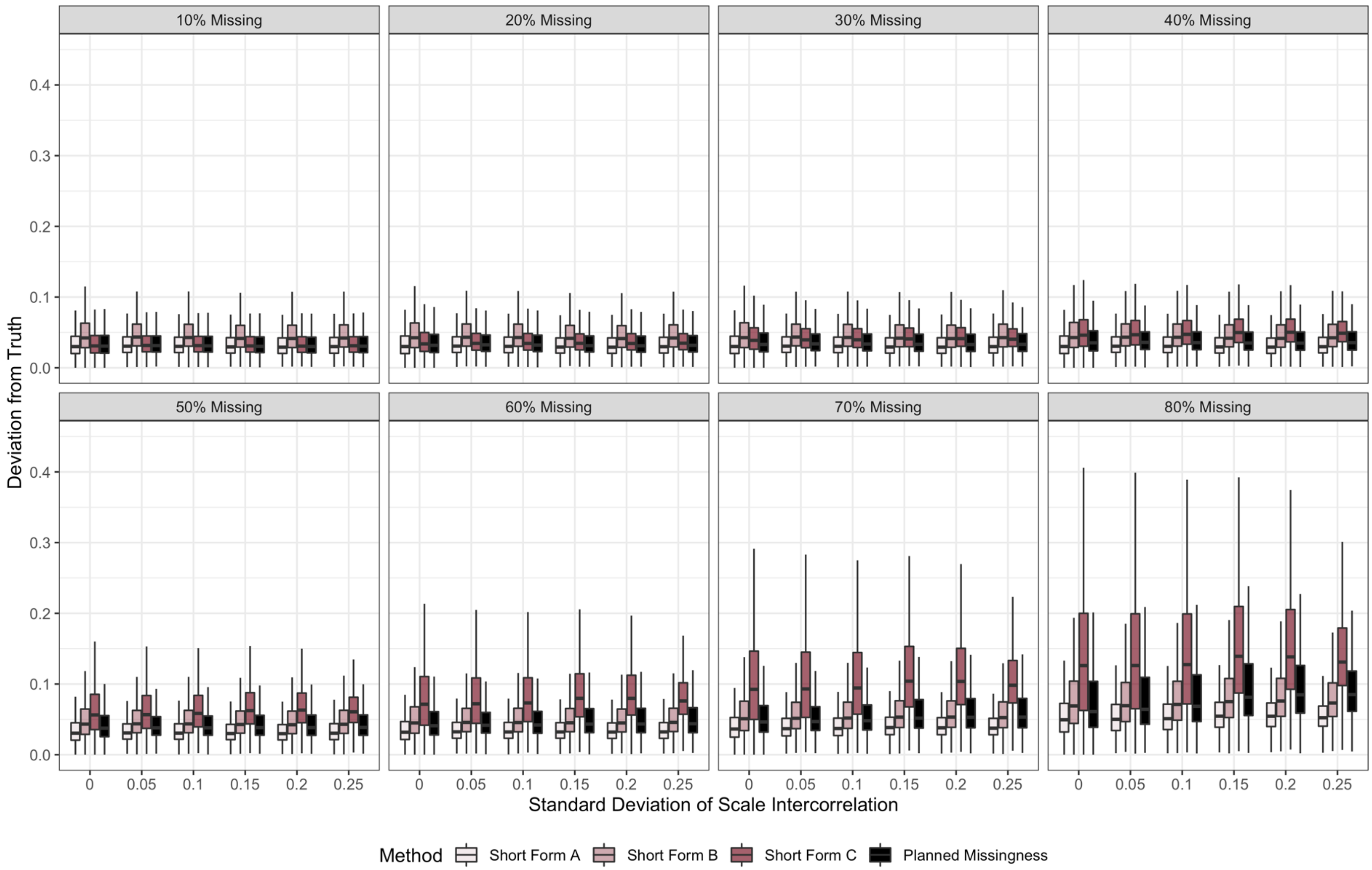

Figure 8. Deviations from True Population Intercorrelation by Method, Intercorrelation Standard Deviation, and Amount of Missingness

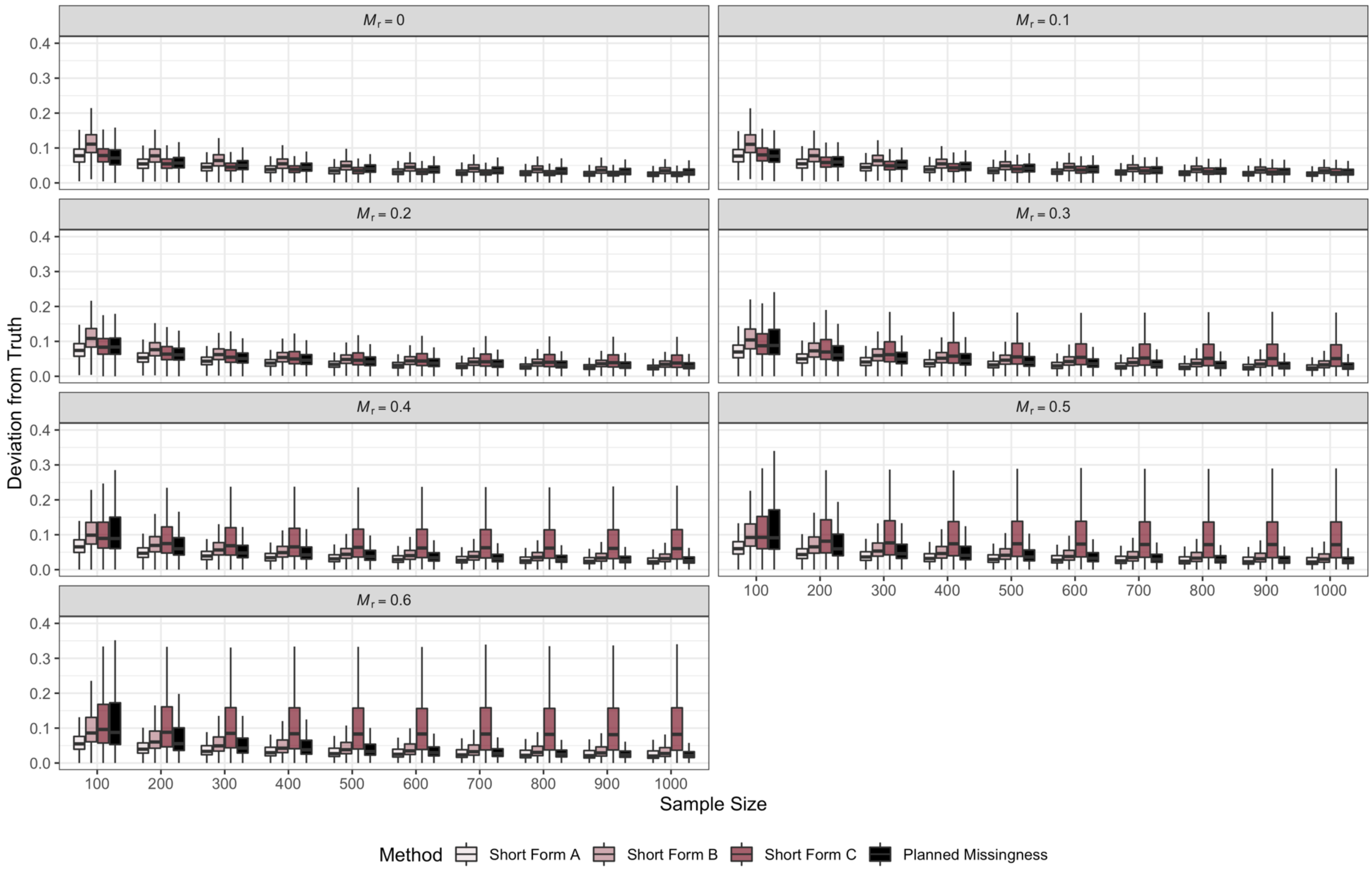

Figure 9. Deviations from True Population Intercorrelation by Method, Sample Size, and Mean Intercorrelation

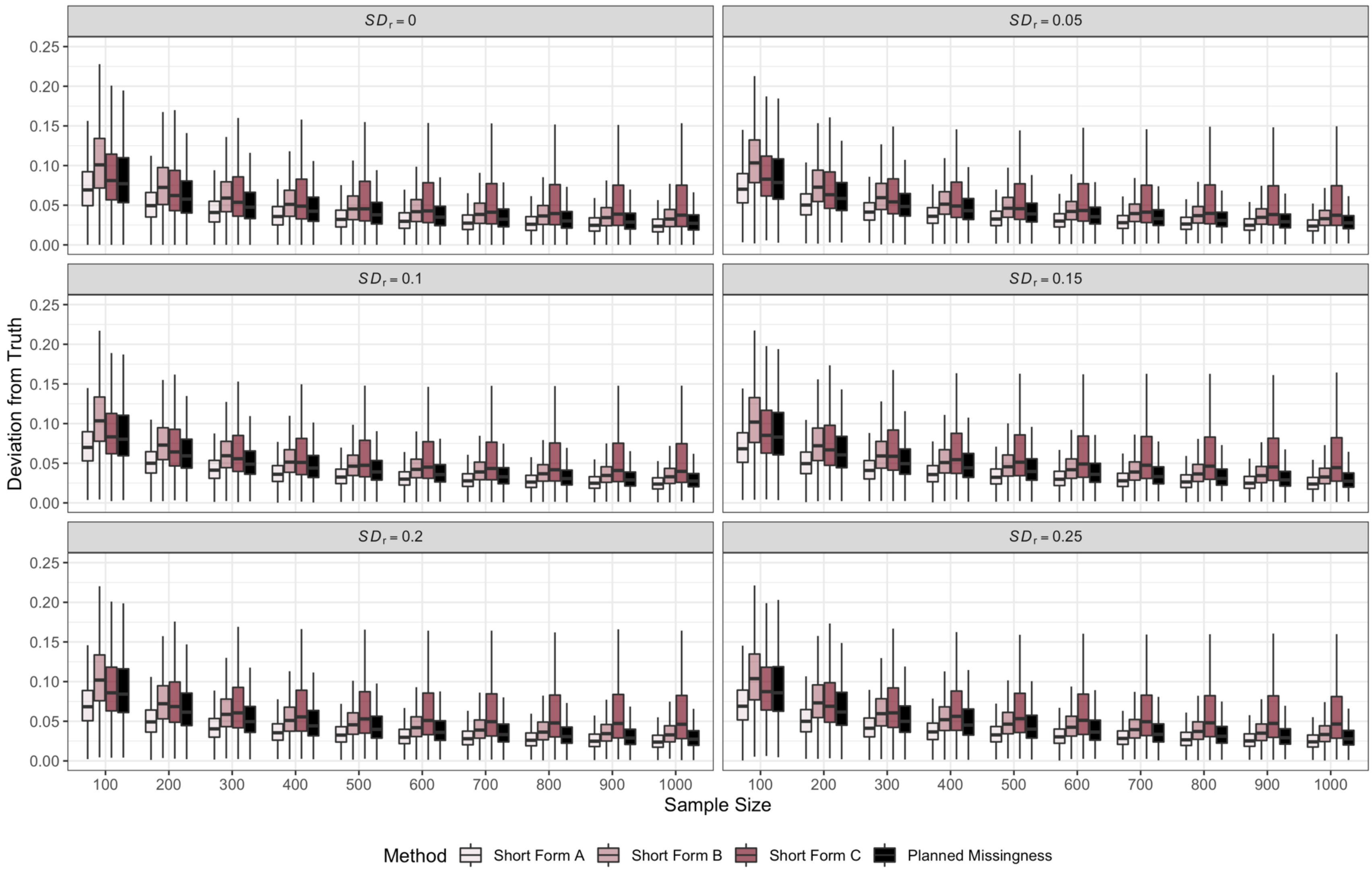

Figure 10. Deviations from True Population Intercorrelation by Method, Sample Size, and Intercorrelation Standard Deviation

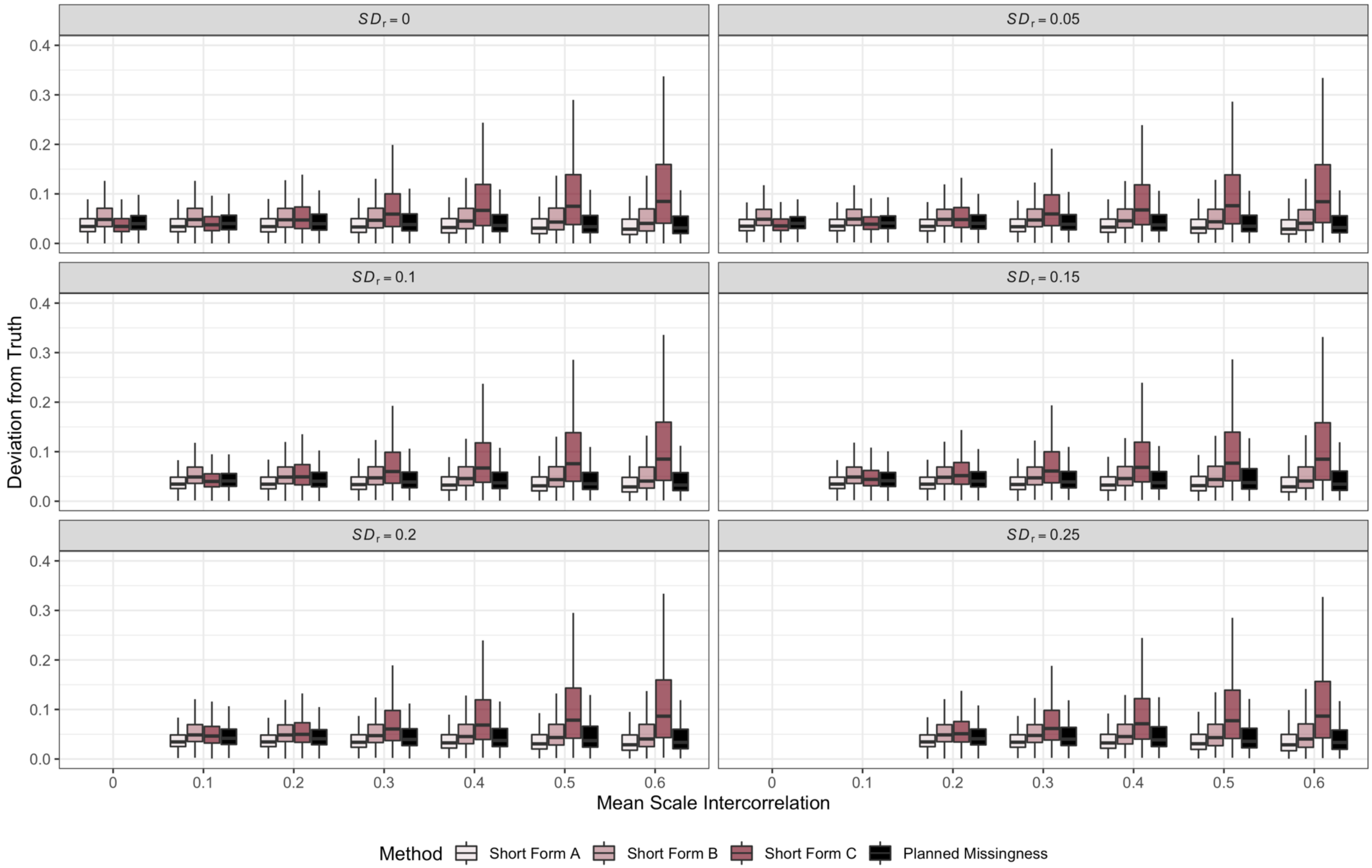

Figure 11. Deviations from True Population Intercorrelation by Method, Mean Intercorrelation, and Intercorrelation Standard Deviation

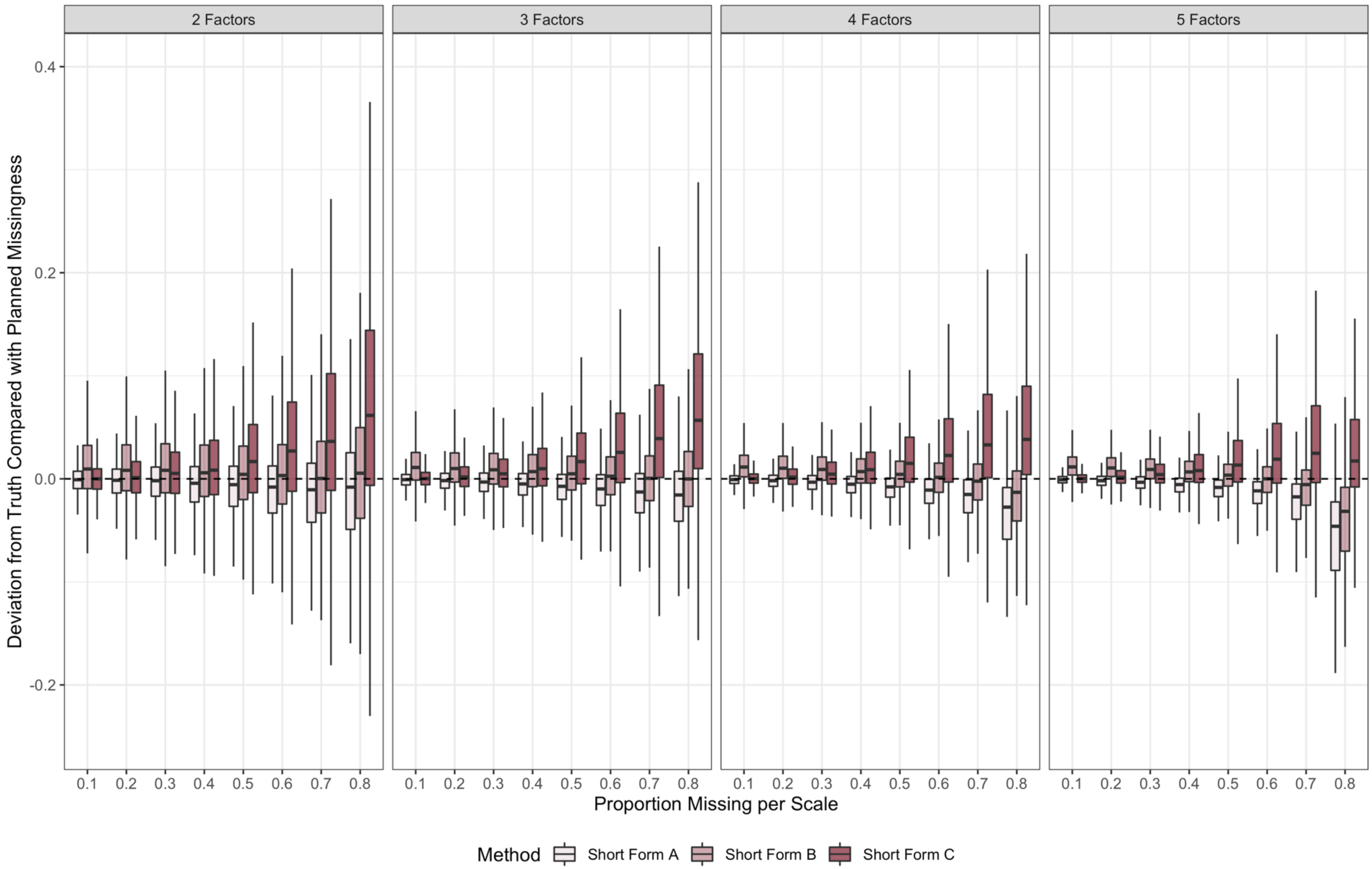

Figure 12. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Amount of Missingness, and Number of Factors

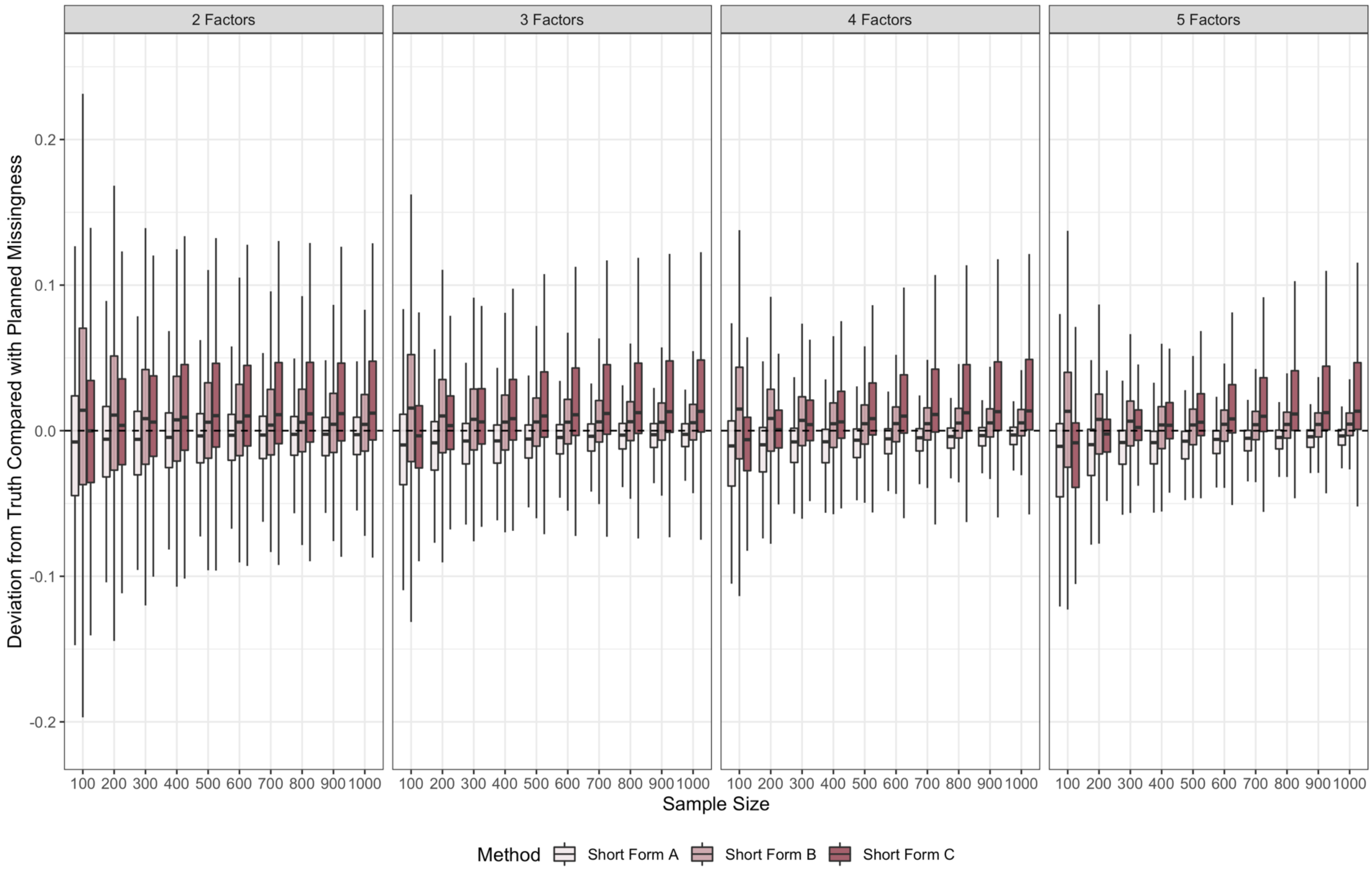

Figure 13. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Sample Size, and the Number of Factors

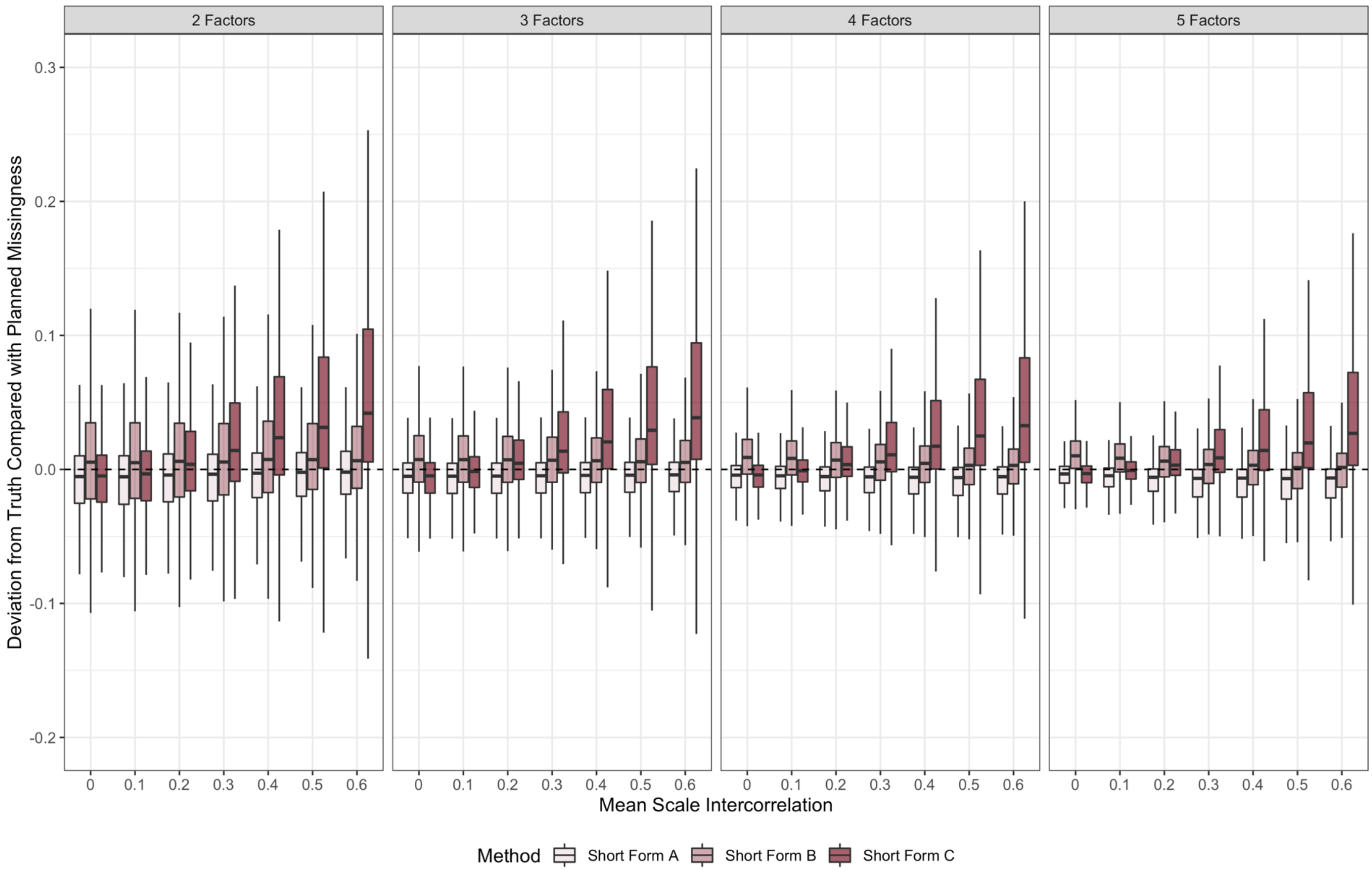

Figure 14. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Mean Intercorrelation, and the Number of Factors

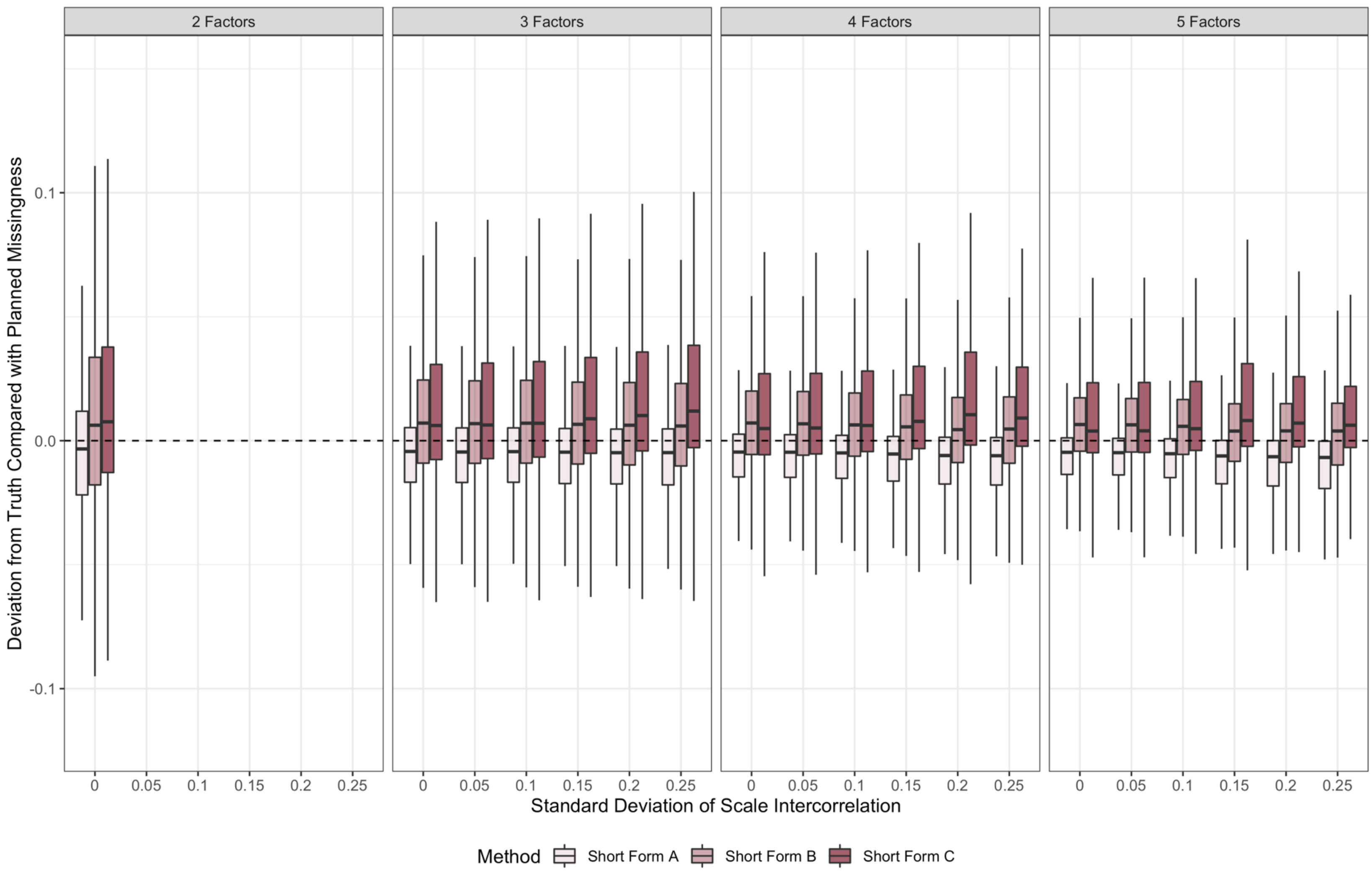

Figure 15. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Intercorrelation Standard Deviation, and the Number of Factors

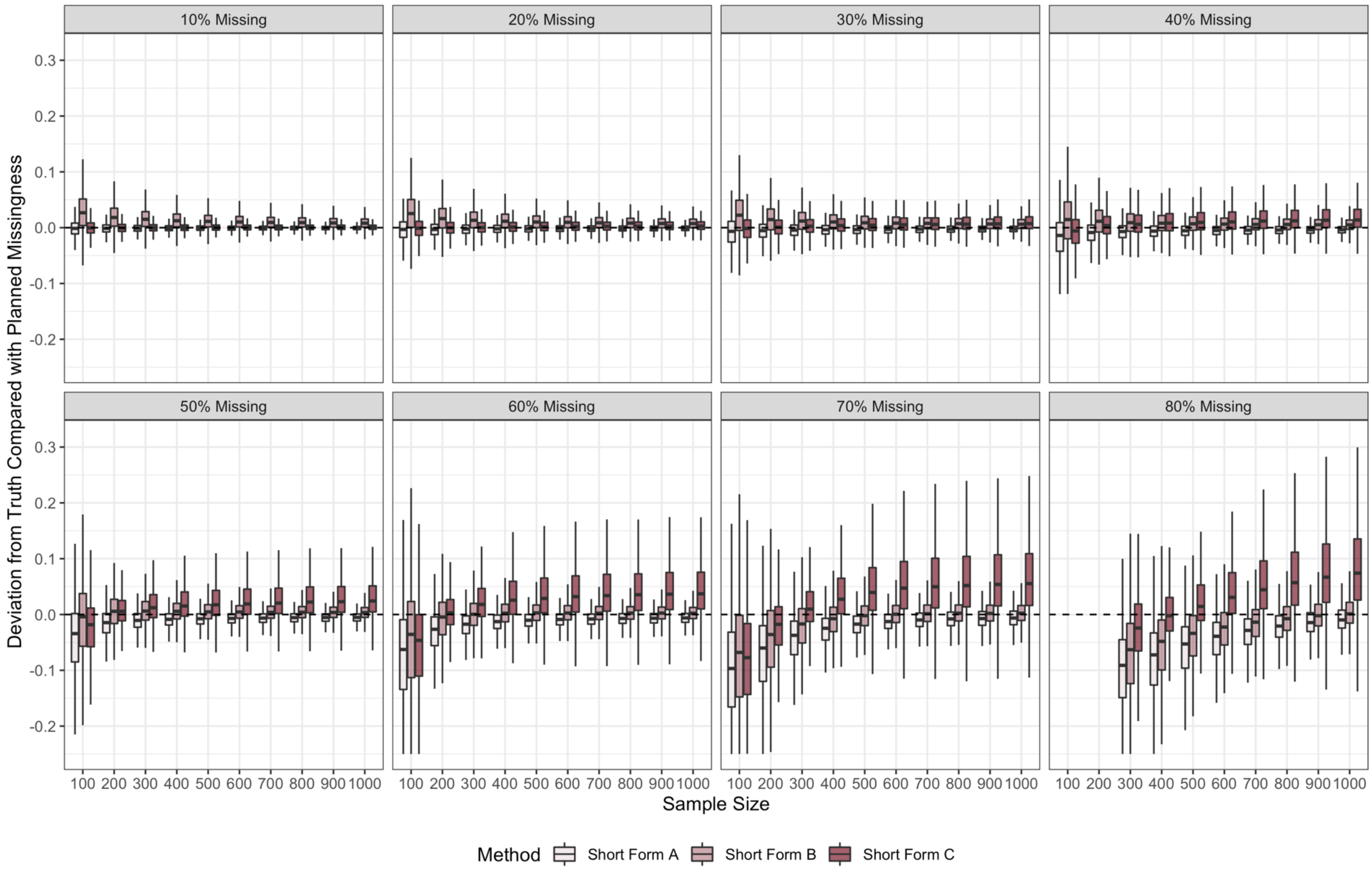

Figure 16. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Sample Size, and the Amount of Missingness

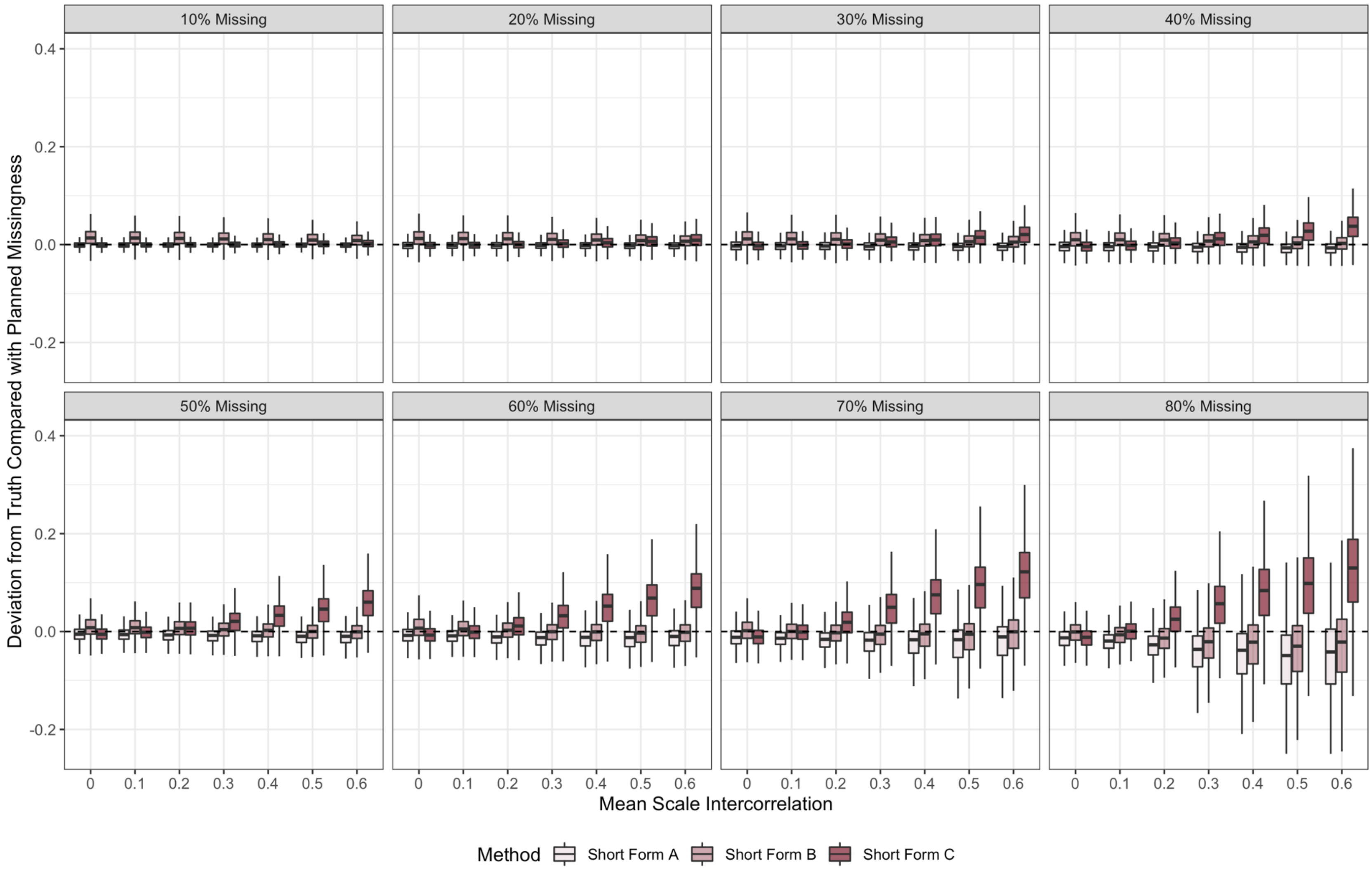

Figure 17. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Mean Intercorrelation, and the Amount of Missingness

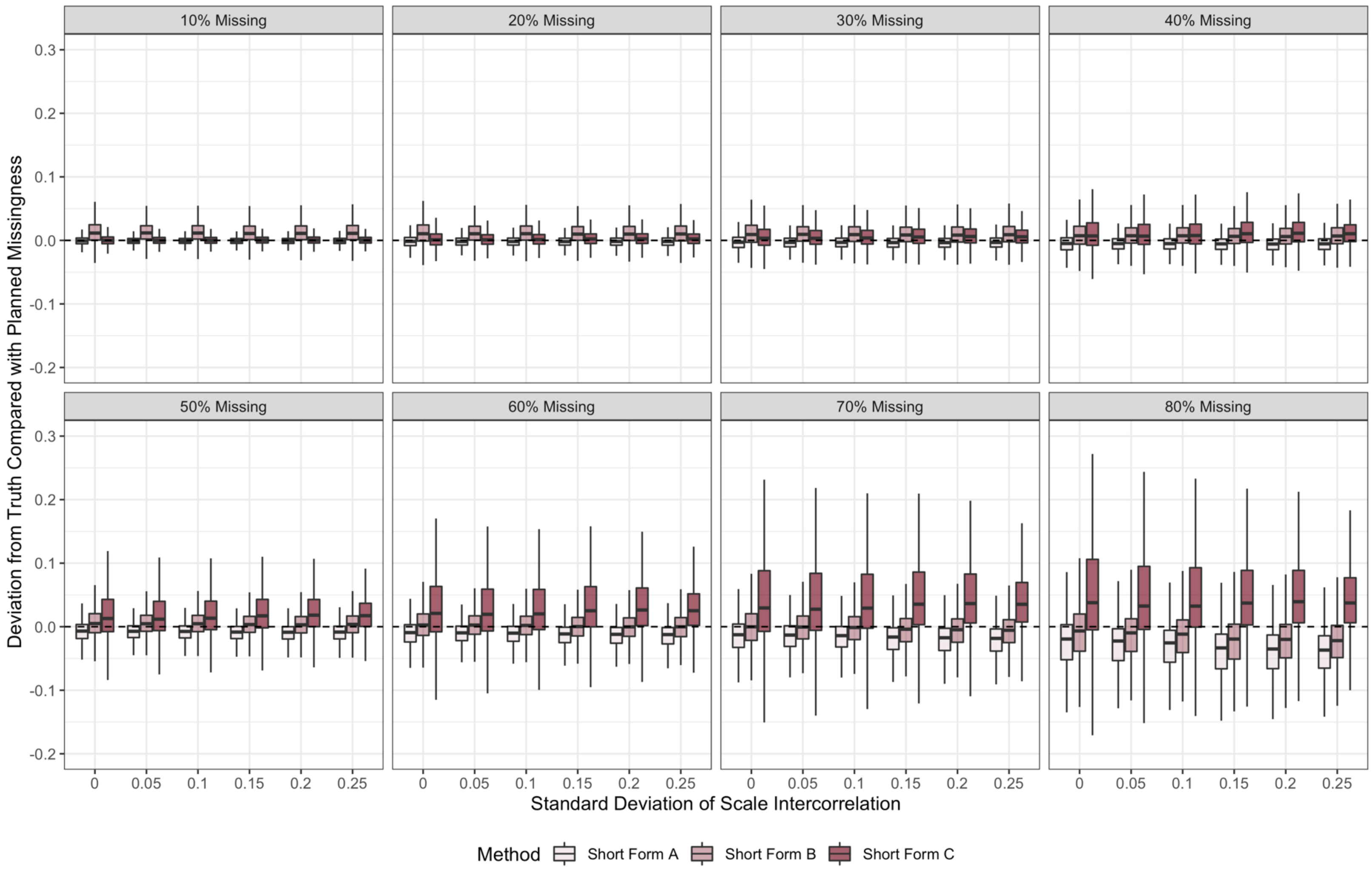

Figure 18. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Intercorrelation Standard Deviation, and the Amount of Missingness

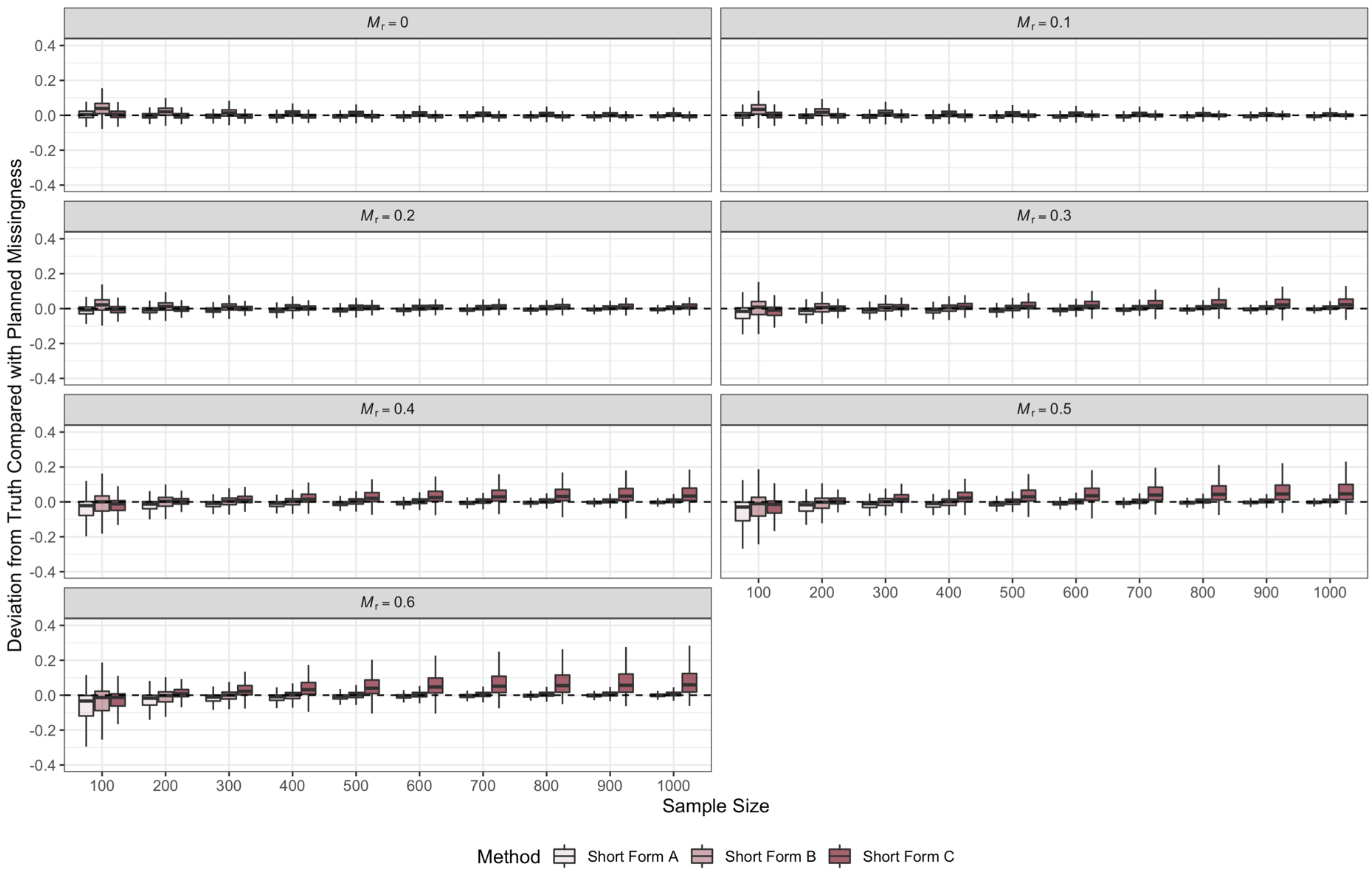

Figure 19. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Sample Size, and Intercorrelation Mean

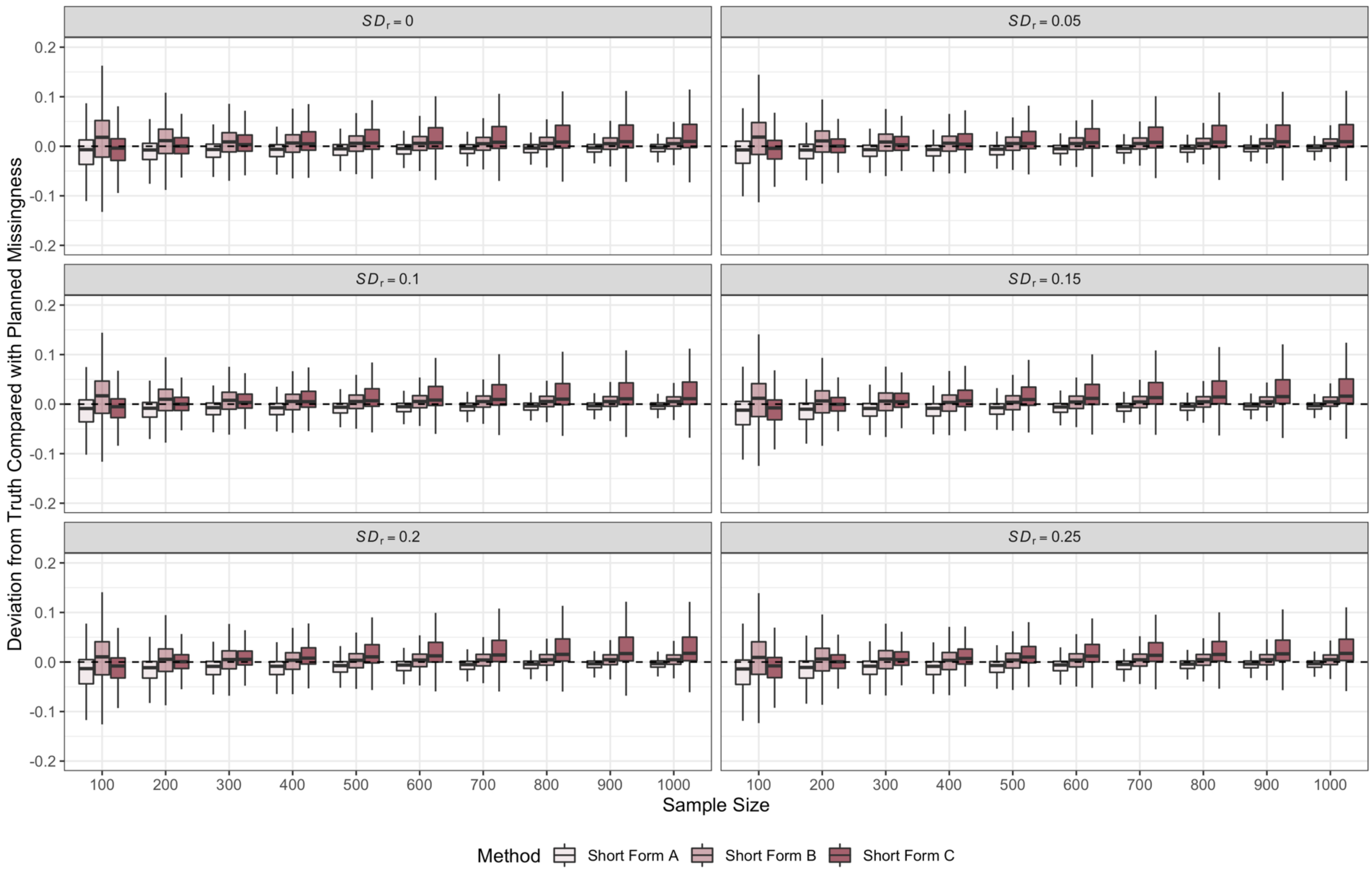

Figure 20. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Sample Size, and Intercorrelation Standard Deviation

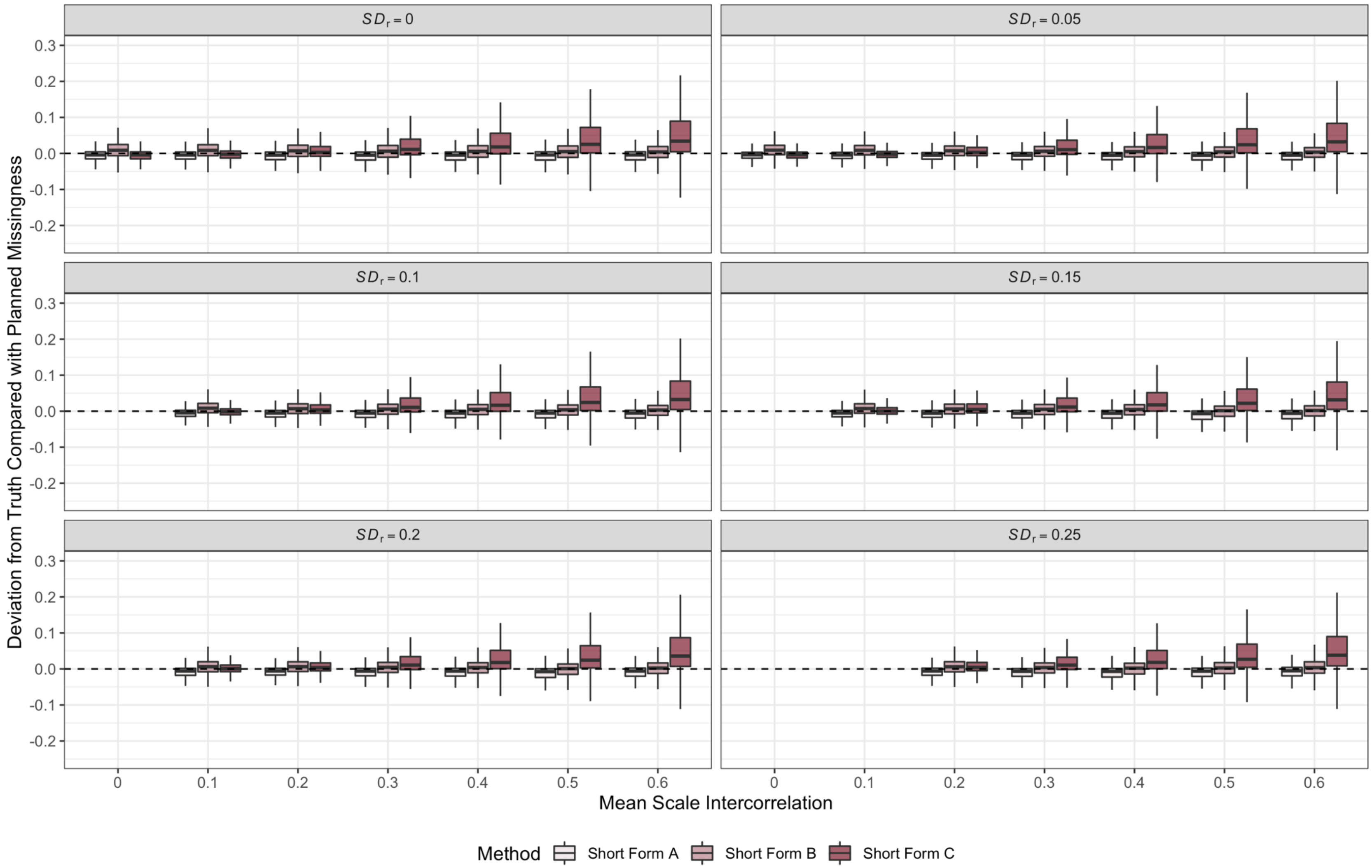

Figure 21. Deviation from Truth of Short Forms Compared with Planned Missingness by Method, Mean Intercorrelation, and Intercorrelation Standard Deviation

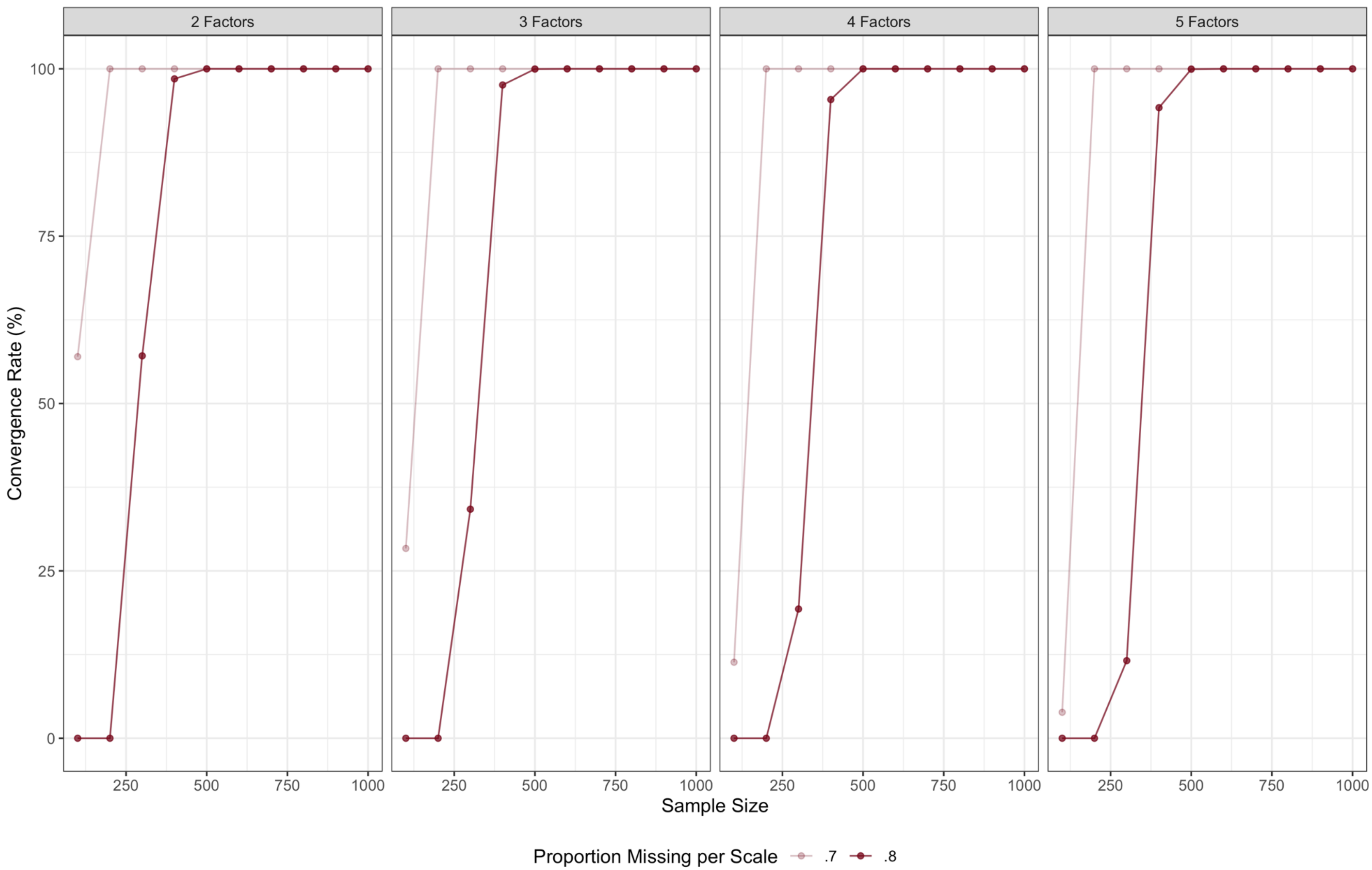

Figure 22. Percentage of Iterations that Imputed Successfully by Sample Size and the Amount of Missingness

References

Andridge, R. R., & Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, *27*(2), 153.

Arbuckle, J. L. (1995). AMOS for Windows analysis of moment structures. Version 3.5. *Chicago, IL: SmallWaters*.

Arbuckle, J. L., & Marcoulides, G. A. (1996). Full information estimation in the presence of incomplete data. In *Advanced structural equation modeling* (G. A. Marcoulides&R. E. Schumacker (Eds.), pp. 243–277). Psychology Press. https://www.jstor.org/stable/2289545

Arminger, G., & Sobel, M. E. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, *85*(409), 195–203.

Armstrong, P. I., Allison, W., & Rounds, J. (2008). Development and initial validation of brief public domain RIASEC marker scales. *Journal of Vocational Behavior*, *73*(2), 287–299. https://doi.org/10.1016/j.jvb.2008.06.003

Barbot, B. (2019). Measuring creativity change and development. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 203. https://doi.org/10.1037/aca0000232

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, *85*(3), 349–360. https://doi.org/doi:10.1037/0021-9010.85.3.349

Berry, K., Rana, R., Lockwood, A., Fletcher, L., & Pratt, D. (2019). Factors associated with inattentive responding in online survey research. *Personality and Individual Differences*, *149*, 157–159. https://doi.org/10.1016/j.paid.2019.05.043

Bowling, N. A., Gibson, A. M., & DeSimone, J. (2021). Questionnaire Length and Scale Validity. *Paper Presented at the 36th Annual Conference of Society for Industrial and Organizational Psychology Conference*.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, *22*(2), 302–306.

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

Conrad-Hiebner, A., Schoemann, A. M., Counts, J. M., & Chang, K. (2015). The development and validation of the Spanish adaptation of the Protective Factors Survey. *Children and Youth Services Review*, *52*, 45–53. https://doi.org/10.1016/j.childyouth.2015.03.006

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology*, *105*(12), 1351–1381. https://doi.org/10.1037/apl0000815

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, *15*(1), 21–36. https://doi.org/10.1023/B:MARK.0000021968.86465.00

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 49–73.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological Methods*, *8*(3), 322–337. https://doi.org/10.1037/1082-989X.8.3.322

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

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.

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

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*(4), 409–420.

Folberg, A. (2020). Advancing stereotyping research: How and why to use linear mixed-effects models in gender stereotyping research. *Testing, Psychometrics, Methodology in Applied Psychology*, *27*(3), 407–431. https://doi.org/10.4473/TPM27.3.6

Foorman, B. R., Herrera, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). The structure of oral language and reading and their relation to comprehension in Kindergarten through Grade 2. *Reading and Writing*, *28*(5), 655–681. https://doi.org/10.1007/s11145-015-9544-5

Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete Data in Sample Surveys*, *2*(Part IV), 185–207.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gnambs, T., & Staufenbiel, T. (2016). Parameter accuracy in meta-analyses of factor structures. *Research Synthesis Methods*, *7*(2), 168–186. https://doi.org/10.1002/jrsm.1190

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7–28.

Goldberg, L. R. (2001). *Big-Five Factor Markers*. International Personality Item Pool. https://ipip.ori.org/newBigFive5broadKey.htm

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

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

Graham, J. W., Cumsille, P. E., & Shevock, A. E. (2012). Methods for handling missing data. *Handbook of Psychology, Second Edition*, *2*.

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., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

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. https://doi.org/10.1037/1082-989X.11.4.323

Harel, O., Stratton, J., & Aseltine, R. (2011). Designed missingness to better estimate efficacy of behavioral studies. *Paper Presented at the 10th Conference on Health Survey Research Methods*. National Center for Health Statistics. http://www.srl.uic.edu/Publist/HSRM10_proceedings.pdf#page=154

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. https://doi.org/10.7243/2053-7662-3-2

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*(1), 64–78. https://doi.org/10.1037/1082-989X.2.1.64

Hogue, C. M., Pornprasertmanit, S., Fry, M. D., Rhemtulla, M., & Little, T. D. (2013). Planned missing data designs for spline growth models in salivary cortisol research. *Measurement in Physical Education and Exercise Science*, *17*(4), 310–325. https://doi.org/10.1080/1091367X.2013.831766

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 727–730. https://doi.org/10.3758/BF03200754

Houseman, E. A., & Milton, D. K. (2006). Partial questionnaire designs, questionnaire non-response, and attributable fraction: Applications to adult onset asthma. *Statistics in Medicine*, *25*(9), 1499–1519. https://doi.org/10.1002/sim.2265

Jenkins, R. J., McAlaney, J., & McCambridge, J. (2009). Change over time in alcohol consumption in control groups in brief intervention studies: Systematic review and meta-regression study. *Drug and Alcohol Dependence*, *100*(1), 107–114. https://doi.org/10.1016/j.drugalcdep.2008.09.016

Joreskog, K. G., & Sorbom, D. (1983). Lisrel VI: Supplement to the Lisrel V Manual. *Uppsala, Sweden: University of Uppsala, Department of Statistics*.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, *17*(23), 2723–2732. https://doi.org/10.1002/(SICI)1097-0258(19981215)17:23<2723::AID-SIM38>3.0.CO;2-5

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*(4), 609–623.

Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. *Methodology*, *14*(1), 3–15. https://doi.org/10.1027/1614-2241/a000141

Laczo, R. M. (2002). *An examination of the dimensionality of non-task performance* [Unpublished Doctoral Dissertation]. University of Minnesota.

Lin, B., Crnic, K. A., Luecken, L. J., & Gonzales, N. A. (2014). Maternal prenatal stress and infant regulatory capacity in Mexican Americans. *Infant Behavior and Development*, *37*(4), 571–582. https://doi.org/10.1016/j.infbeh.2014.07.001

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.

Little, R. J. (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. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies: Journal of the American Statistical Association: Vol 90, No 431. *Journal of the American Statistical Association*, *90*(431), 1112–1121. https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476615

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). Wiley & Sones.

Little, R. J., & Schenker, N. (1995). Missing data. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). Springer.

Little, T. D., Gorrall, B. K., Panko, P., & Curtis, J. D. (2017). Modern practices to improve human development research. *Research in Human Development*, *14*(4), 338–349. https://doi.org/10.1080/15427609.2017.1370967

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199–204. https://doi.org/10.1111/cdep.12043

Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, *36*(1), 116–124. https://doi.org/10.1177/0894439317695581

Lüdtke, O., Robitzsch, A., & Grund, S. (2016). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. https://doi.org/10.1037/met0000096

Marcus-Blank, B., Kuncel, N. R., & Sackett, P. R. (2015). Does rationality predict performance in major life domains? *Paper Presented at the 30th Annual Conference of Society for Industrial and Organizational Psychology Conference*.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://psycnet.apa.org/buy/2012-10015-001

Mendoza, J. L. (1993). Fisher transformations for correlations corrected for selection and missing data. *Psychometrika*, *58*(4), 601–615. https://doi.org/0033-3123/93/1200-92020500.75/0

Mistler, S. A., & Enders, C. K. (2012). Planned missing data designs for developmental research. *Handbook of Developmental Research Methods*, 742–754.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 371–388. https://doi.org/10.1111/j.1467-9868.2007.00640.x

Molenberghs, G., & Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. Springer Science & Business Media. https://books.google.com/books?hl=en&lr=&id=LjfyKpw36S8C&oi=fnd&pg=PA3&dq=molenberghs+verbeke+2005+models+for+discrete+&ots=kA89Nyog-c&sig=VzKRTQgXTbctgf86LNJucme_LU0#v=onepage&q=molenberghs%20verbeke%202005%20models%20for%20discrete&f=false

Munger, G. F., & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187–191. https://doi.org/10.1080/00220973.1988.10806486

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431–462. https://doi.org/0033-3123/87/0900-8055500.75/0

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

Newman, D. A. (2009). Missing data techniques and low response rates. In *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (Charles E. Lance&Robert J. Vandenberg (Eds.), Vol. 7). Routledge.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Newman, D. A., & Cottrell, J. M. (2015). Missing data bias: Exactly how bad is pairwise deletion? In *More statistical and methodological myths and urban legends* (pp. 133–161). Routledge/Taylor & Francis Group.

Open Source Psychometrics Project. (2018). *Open Source Psychometrics Project*. https://openpsychometrics.org/about/

Paulhus, D. L., & Jones, D. N. (2011). *Introducing a short measure of the Dark Triad*. Poster presented at the meeting of the Society for Personality and Social Psychology, San Antonio.

Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study. *PLOS ONE*, *11*(3), e0152330. https://doi.org/10.1371/journal.pone.0152330

Qualtrics. (2021). *Qualtrics Support-Randomizer*. https://www.qualtrics.com/support/survey-platform/survey-module/survey-flow/standard-elements/randomizer/

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*(429), 54–63.

Revelle, W., Dworak, E. M., & Condon, D. M. (2020). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, 109905. https://doi.org/10.1016/j.paid.2020.109905

Revilla, M., & Ochoa, C. (2017). Ideal and maximum length for a web survey. *International Journal of Market Research*, *59*(5), 557–565. https://doi.org/10.2501/IJMR-2017-039

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

Rhemtulla, M., 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. https://doi.org/10.1177/0165025413514324

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

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2021). *MASS: Support Functions and Datasets for Venables and Ripley's MASS* (7.3-53.1) [Computer software]. https://CRAN.R-project.org/package=MASS

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, *5*(3), 127–134. https://doi.org/10.1111/j.1467-9280.1994.tb00646.x

Rosseel, Y., Oberski, D., Vanbrabant, L., Vanbrabant, L., Savalei, V., & Merkle, E. (2012). *Package 'lavaan'. R package version 0.6-3*.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227–241. https://doi.org/10.1177/1536867X0400400301

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=bQBtw6rx_mUC&oi=fnd&pg=PR24&dq=Multiple+imputation+for+non-response+in+surveys&ots=8OsK8M1_gQ&sig=P-5VrHI4g_RMyFVPgjMjLN2AkzQ

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Sackett, P. R., Berry, C. M., Wiemann, S. A., & Laczo, R. M. (2006). Citizenship and Counterproductive Behavior: Clarifying Relations Between the Two Domains. *Human Performance*, *19*(4), 441–464. https://doi.org/10.1207/s15327043hup1904_7

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

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

Smits, N., & Vorst, H. C. M. (2007). Reducing the length of questionnaires through structurally incomplete designs: An illustration. *Learning and Individual Differences*, *17*(1), 25–34. https://doi.org/10.1016/j.lindif.2006.12.005

Speidel, M., Drechsler, J., & Jolani, S. (2020). *hmi: Hierarchical Multiple Imputation* (1.0.0) [R]. https://CRAN.R-project.org/package=hmi

Thijs, H., Molenberghs, G., & Verbeke, G. (2000). The milk protein trial: Influence analysis of the dropout process. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *42*(5), 617–646. https://doi.org/10.1002/1521-4036(200009)42:5<617::AID-BIMJ617>3.0.CO;2-N

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods. *Survey Methodology*, *32*(2), 217–231. https://scholar.google.com/scholar_lookup?title=An%20evaluation%20of%20matrix%20sampling%20methods%20using%20data%20from%20the%20national%20health%20and%20nutrition%20examination%20survey&journal=Survey%20Methodol.&volume=32&issue=2&publication_year=2006&author=Thomas%2CN&author=Raghunathan%2CTE&author=Schenker%2CN&author=Katzoff%2CMJ&author=Johnson%2CCL

Thorndike, R. L. (1949). *Personnel selection; test and measurement techniques.* Wiley.

Troxel, A. B., Ma, G., & Heitjan, D. F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica*, 1221–1237.

Tsonaka, R., Verbeke, G., & Lesaffre, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, *65*(1), 81–87. https://doi.org/10.1111/j.1541-0420.2008.01021.x

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Chapman & Hall/CRC. https://stefvanbuuren.name/fimd/

van Buuren, S., & Groothuis-Oudshoorn, K. (2010). *MICE: Multivariate Imputation by Chained Equations in R* [Article]. Journal of Statistical Software. http://dspace.library.uu.nl/handle/1874/44635

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., & Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics*, *57*(1), 7–14. https://doi.org/10.1111/j.0006-341X.2001.00007.x

Vink, G., & van Buuren, S. (2011). *Ad hoc methods and mice*. https://www.gerkovink.com/miceVignettes/Ad_hoc_and_mice/Ad_hoc_methods.html

Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *ArXiv:1409.8542 [Math, Stat]*. http://arxiv.org/abs/1409.8542

Wacholder, S., Carroll, R. J., Pee, D., & Gail, M. H. (1994). The partial questionnaire design for case-control studies. *Statistics in Medicine*, *13*(5–7), 623–634. https://doi.org/10.1002/sim.4780130523

Waller, N. G. (2020). *fungible: Psychometric Functions from the Waller Lab* (1.95.4.8) [R].

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, *10*(3), 390–399. https://doi.org/10.1177/1745691614564879

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation*, *14*(5), 1–9. https://doi.org/10.7275/AS0K-TM88

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, *3*(3), 163–195.

Wood, J., Matthews, G. J., Pellowski, J., & Harel, O. (2019). Comparing Different Planned Missingness Designs in Longitudinal Studies. *Sankhya B*, *81*(2), 226–250. https://doi.org/10.1007/s13571-018-0170-5

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (T.D. Little, K. U. Schnabel, J. Baumert (Eds.), pp. 219–240, 269–281). Laurence Erlbaum Associates Publishers.

Wu, W., Jia, F., Rhemtulla, M., & Little, T. D. (2016). Search for efficient complete and planned missing data designs for analysis of change. *Behavior Research Methods*, *48*(3), 1047–1061. https://doi.org/10.3758/s13428-015-0629-5

Yamada, T. (2020). *Organizational and work correlates of sleep* [Dissertation, University of Minnesota]. http://conservancy.umn.edu/handle/11299/216378

Yoon, H. R., & Sackett, P. R. (2016). Addressing time constraints in surveys: Planned missingness vs. Short forms. *Paper Presented at the 31st Annual Conferences of the Society for Industrial and Organizational Psychology*.

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, *267*.

Zhang, C., & Yu, M. C. (In press). Planned missingness: How to and how much? *Organizational Research Methods*.

Appendices

Appendix A. Planned Missingness Perceptions Study (Study 3) Survey

Part I

1. In your work, approximately how many self-reported data collections (e.g., surveys, test batteries) have you designed or contributed to designing?
    a. 0
    b. 1-5
    c. 6-10
    d. 11-15
    e. 16-20
    f. >20
2. Approximately how many of these self-report data collections (e.g., surveys, test batteries) have you had the need to reduce their length?
    a. 0
    b. 1-5
    c. 6-10
    d. 11-15
    e. 16-20
    f. >20
3. A planned missingness (PM) design can be implemented in survey studies, in which a randomly selected percentage of items are administered to each respondent. By using a PM design, the length of a survey can be reduced. Were you familiar with this concept prior to participating in this study?
    a. Yes
    b. No
4. Have you ever implemented a planned missingness deign in your research/work?
    a. Yes
    b. No
5. If yes to #2, how did you treat the missingness?
    a. Full information maximum likelihood
    b. Multiple imputation
    c. Other:
6. If yes to #2, what kind of study were you conducting?
    a. Concurrent validation study
    b. Predictive validation study
    c. Survey research (e.g., engagement survey, academic research)
    d. Methodological study concerning data missingness
    e. Other:
7. If no to #2, why not?
    a. Never heard of it/do not know what it is
    b. Not appliable to my work because I have not had a need to reduce survey length
    c. Others:

8. In your work, what approaches have you taken to make sure a survey is not too lengthy? (check all that apply)
    a. Short forms
    b. Planned missingness
    c. Cut down number of constructs/scales
    d. N/A

Part II

Oftentimes, there is motivation to shorten a survey. Sometimes there is constraint on the amount of time respondents have to spend. It can also be beneficial for researchers to reduce the length of a survey to prevent respondent fatigue and careless responding.

Two approaches can be used to reduce survey length without cutting down the number of constructs measured. First, short forms of scales administered rather than full measures. Sometimes, such short forms already exist in the literature; sometimes they have to be developed. Second, a planned missingness (PM) design can be implemented, in which a randomly selected percentage of items are administered to each respondent. Thus, all respondents receive a survey of the same length, but the subset of items given to each individual is different and every item is answered by a subset of the sample. Research shows that the true covariance structure can be closely replicated with two techniques. When all that is needed for subsequent analyses is the covariance matrix, full-information maximum likelihood estimation can be used. Alternatively, multiple imputation can be used to fill in each of the missing datapoints, and the completed datasets can then be analyzed.

Below is a simple example of what the two approaches would look like if the full survey consists of 10 items, but respondents only have time to complete seven items. If using a short form, seven items would be chosen to be administered to all respondents. If implementing a planned missingness design, each respondent would be given a randomly selected seven items.

	Short form										vs.	Planned Missingness									
	Items											Items									
Respondent	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	0	0	0		1	0	0	1	1	0	1	1	1	1
2	1	1	1	1	1	1	1	0	0	0		0	1	1	0	1	1	0	1	1	1
3	1	1	1	1	1	1	1	0	0	0		1	1	1	1	0	1	1	0	1	0
4	1	1	1	1	1	1	1	0	0	0		1	1	1	0	1	0	1	1	0	1

5	1	1	1	1	1	1	1	0	0	0	1	0	1	1	0	1	0	1	1	1
6	1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	0	1	0
7	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	0	1	1	0	1
8	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1	1	1	0	1	0
9	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	0	1	0	1

Note. 1 = item administered. 0 = item not administered.

In a series of simulation studies, I compared the performance of these two approaches for reproducing population estimates. Across a wide range of sample sizes, number of variables, mean variable intercorrelations, and missingness levels, using short forms and implementing planned missingness performed comparably. Please consider the following scenarios in which your goal is to reduce study length. Given the technical equivalence of using short forms and planned missingness, please indicate whether you would prefer one approach to the other for any contextual reasons.

9. An MTurk research study examining the convergent and discriminant validity of a new personality measure by administering the new measure along with a number of other personality scales.
    a. I would prefer using short forms.
    b. I would prefer implementing planned missingness.
    c. I have no preference.
    d. If A or B, why:
10. A battery of tests is administered to applicants of an entry-level job and used to make selection decision.
    a. I would prefer using short forms.
    b. I would prefer implementing planned missingness.
    c. I have no preference.
    d. If A or B, why:
11. An engagement survey is being designed to evaluate job attitudes and perceptions of organizational norms and culture internally.
    a. I would prefer using short forms.
    b. I would prefer implementing planned missingness.
    c. I have no preference.
    d. If A or B, why:
12. A number of new selection tools are being developed and validated. They are administered to incumbents for validation research purposes only.
    a. I would prefer using short forms.
    b. I would prefer implementing planned missingness.

c. I have no preference.
d. If A or B, why:

Background Info

13. Highest Degree Earned
    a. Bachelor's
    b. Master's
    c. Doctoral
    d. Other:
14. What is your degree in?
    a. Industrial/Organizational Psychology
    b. Other areas of psychology
    c. Business (e.g., Organizational Behavior, Human Resources)
    d. Education
    e. Other:
15. Number of years since graduating:
16. Do you primarily work in
    a. Academia
    b. External consulting
    c. Private industry
    d. Government
    e. Other: