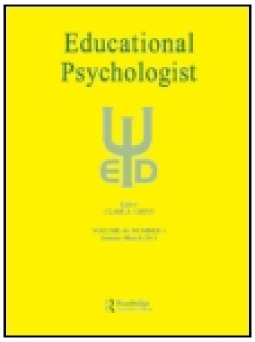

Routledge
Taylor & Francis Group

Educational Psychologist

ISSN: 0046-1520 (Print) 1532-6985 (Online) Journal homepage: http://www.tandfonline.com/loi/hedp20

Planned Missing Data Designs in Educational Psychology Research

Mijke Rhemtulla & Gregory R. Hancock

EDUCATIONAL PSYCHOLOGIST, *51*(3–4), 305–316, 2016
Copyright © Division 15, American Psychological Association
ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2016.1208094

Routledge
Taylor & Francis Group

Planned Missing Data Designs in Educational Psychology Research

Mijke Rhemtulla[1,2] and Gregory R. Hancock[3]
[1]*Department of Psychology, University of Amsterdam, The Netherlands*
[2]*Department of Psychology, University of California, Davis*
[3]*Department of Human Development and Quantitative Methodology, University of Maryland*

Although missing data are often viewed as a challenge for applied researchers, in fact missing data can be highly beneficial. Specifically, when the amount of missing data on specific variables is carefully controlled, a balance can be struck between statistical power and research costs. This article presents the issue of planned missing data by discussing specific designs (i.e., multiform designs, longitudinal wave-missing designs, and 2-method measurement designs), introducing the power and cost benefits of such scenarios to applied education and educational psychology researchers.

Within educational and psychological research, missing data seem to come with the territory. The study of learning within school settings, for example, a common focus of educational psychology research, might involve tightly designed randomized controlled trials to compare programs to facilitate learning. Also common are longitudinal investigations (possibly even within randomized controlled trials), designed to gauge whether, at what rate, and under what circumstances learning is occurring. In addition to their inferential threats arising from cohort/classroom effects and changes in context over time (e.g., students transitioning to new classrooms and/or to new school settings, such as from elementary to middle school), longitudinal designs present the very practical challenges that following individuals over time is costly and typically results in missing data. These missing data could arise for fairly simple reasons, such as school absences, but are more often part of a pattern of attrition resulting from families moving or perhaps from individual students requiring alternative educational settings more tailored to their specific learning needs. Suffice it to say, dealing with missing data, whether in cross-sectional or longitudinal designs, is a very real part of the research process, as can be the dread of having to do so.

Fortunately, with the advent of modern missing data handling methods, and the implementation of these modern methods in most statistical software, missing data are becoming far less troublesome. In particular, so long as the missingness patterns in data are not related to the missing values themselves, analyses can proceed without much trouble or fear of bias. In addition, although some patterns of missing data can result in dramatically reduced power, other patterns can largely avoid it. For example, if two variables are highly correlated, such as perceptions of academic efficacy and academic self-worth, and one is missing many observations, not much power will be lost for testing parameters associated with that variable; in contrast, if the variable with a high rate of missing data (e.g., teacher stress) is unrelated to other variables (e.g., peer norms for cooperation), much more power will be lost for testing its parameters (e.g., means, variances, regression slopes).

Planned missing data designs take advantage of these facts and impose random missingness in such a way that it does not introduce bias and it minimizes power loss. The result can be a dramatic reduction in cost, and even an increase in validity due to reducing participant burden. In this article, after a brief review of some foundations related to missing data, we present three planned missing data designs and how they could be implemented in educational research. The first type of design (multiform design) is presented in the greatest detail, laying the groundwork for the remaining two: accelerated longitudinal designs and two-method measurement designs.

MISSING DATA MECHANISMS AND METHODS

The reasons that missing data occur, *missingness mechanisms*, are typically described as falling into three categories (see,

Correspondence should be addressed to Mijke Rhemtulla, Department of Psychology, University of California, Davis, One Shields Avenue, Davis, CA 95616. E-mail: mrhemtulla@ucdavis.edu

e.g., Little & Rubin, 2002). Put simply, the mechanism is (a) completely unrelated to any observed or missing variables (missing completely at random [MCAR]), (b) related to the observed variables but not to the missing values themselves (missing at random [MAR]), or (c) related to the values that are missing (missing not at random). The first category, MCAR missingness, represents an ideal because procedures for accommodating the missing data, modern (e.g., full information maximum likelihood or multiple imputation) or traditional (e.g., listwise or pairwise deletion), will provide unbiased estimates of the parameters of interest. Unfortunately, this ideal seldom occurs naturally. The second category, MAR missingness, occurs when missingness is predictable from other variables in the observed data (e.g., if lower socioeconomic status [SES] respondents tended to omit a particular item on a motivation survey whereas higher SES respondents did not, then the missingness would be predicted by SES). When missingness is MAR, modern approaches use information from all correlated variables to approximate the missingness mechanism, thereby helping to ensure that no bias arises and power is maximized. Finally, when missing data are missing not at random, the missingness is predicted by the values on the missing variables themselves (e.g., if lower SES students declined to provide information about their parents' level of education), and there are no other measures directly related to the missingness mechanism; as such, the missing values cannot be informed by available data, leading to bias under any method. Thus, to the extent possible, it is critical for researchers to gather data on variables that are anticipated to be as highly predictive of the missingness as possible (e.g., if an SES measure is of interest, data could be collected on students' free/reduced lunch status). Indeed, the success of modern missing data approaches depends on it.

The two modern missing data methods most highly recommended (see, e.g., Enders, 2010, 2013) are full information maximum likelihood (FIML) and multiple imputation (MI). As the name implies, FIML uses all of the information in the observed data to derive a single set of parameter estimates (e.g., means, variances, regression slopes) that maximize the likelihood of the observed data having come from the population(s) implied by those estimates. In MI, on the other hand, data are imputed multiple times to create *m* complete data sets ($m > 20$ typically), each of which is in turn analyzed and the results of which are aggregated to yield parameter estimates and standard errors. Although results from MI are generally as accurate as those produced by FIML and can be employed in a wide range of simple and complex analyses, FIML is available in most structural equation modeling software packages and is assumed in the current article.

INFORMATION, EFFICIENCY, AND POWER

Returning to the seemingly unrealistic scenario where data are MCAR, when modern missing data methods are employed the *accuracy* of the resulting parameter estimates is not a concern. That is, neither the amount of missingness nor the pattern of missingness will influence the accuracy of parameter estimates. In contrast, the amount and pattern of missingness can greatly affect the information available to estimate parameters. *Information* is a statistical concept that is inversely related to the standard errors of parameter estimates: The more information available to estimate a parameter, the smaller its standard error (and thus the smaller its confidence interval, and the greater the statistical power of a significance test on that parameter). Thus, information is directly related to *efficiency*, where a more efficient procedure or design results in smaller standard errors (and therefore greater power) than a less efficient procedure. Just as information is affected by sample size—a larger total sample size yields more information and thus more statistical power—it is also affected by missing data. All else being equal, more missing data means more missing information, and thus larger standard errors, larger parameter confidence intervals, and lower statistical *power*. However, the amount of missing information is not a straightforward function of the amount of missing data. Missing information is affected by which variables are missing (e.g., if two variables are perfectly correlated, deleting one of them will not affect the amount of information in the data) and the pattern of missingness (e.g., if no participants have complete data on both *X* and *Y*, there is no information available to estimate the *X*–*Y* correlation).

So when, as per the premise of the previous paragraph, would one ever have data that are truly MCAR? The answer is, when the researcher is in control of the missingness, that is, when the researcher decides who is missing what data. The idea behind planned missing designs, as mentioned at the beginning of the article, is to do just that: assign participants to provide some fraction of the total data, but to do so in a way so as to preserve information and maximize efficiency while reducing cost and participant burden. The remainder of this article introduces three such planned missing data designs that could be especially useful within educational and psychological research.

MULTIFORM DESIGNS

The *multiform design*, also known as a *split questionnaire design* or an *efficiency design*, is a highly adaptable approach that can be incorporated into a cross-sectional or longitudinal design. The idea is simple: To reduce the length of a survey or assessment, administer a subset of the items to each participant. For example, in the classic three-form design (Graham, Hofer, & Piccinin, 1994), items are assigned to one of four item sets (X, A, B, or C), which are then combined to create three short forms. Table 1 depicts the assignment of sets to forms: The X set is included in each form, and the other sets are each missing from one

TABLE 1
Three-Form Design

	Item Set			
Form	*X*	*A*	*B*	*C*
1				▨
2			▨	
3		▨		

Note. Gray represents missing data; white represents complete data.

form. Participants are then randomly assigned to receive one of the forms, thereby making missingness completely random (note that to ensure completely random missingness, assignment to forms must be done at the level of the student, not a higher level such as the classroom). A multiform design is appropriate whenever measures include multiple-item scales or tests, as exemplified by Echols (2015) in the assessment of peer-reported victimization in middle school. The design is particularly valuable when the length of the assessment is a concern, such as when fatigue effects may compromise the validity of the data.

A growing body of evidence suggests that data resulting from multiform designs are highly similar to those from corresponding complete data designs. For example, Smits and Vorst (2007) applied multiform missingness (specifically, within-block; see next) to complete survey data assessing study skills and achievement motivation. They compared two designs, including a three-form design with 33% missing data and a six-form design with 50% missing data, to the complete data characteristics. They reported that mean scale scores, reliability (Cronbach's alpha), and predictive validity estimates were very similar across all designs, and only standard errors were higher with missing data (due to the missing information, as expected).

Swain (2015) considered the effect of planned missingness on students' performance on standardized assessments in a situation when students were not highly motivated to succeed on the tests (i.e., test scores were used to assess the institution, not individual students). Students were assigned either to complete an entire assessment (2-hr test) or to complete one form of a six-form design with 50% of the items missing (1-hr test). Swain reported that the results were highly similar but that students in the planned missing conditions scored slightly but statistically significantly higher ($d = 0.172$), possibly as a result of reduced fatigue. Item analyses revealed no notable differential item functioning across the two designs.

Harel, Stratton, and Aseltine (2012) investigated whether planned missingness could be used to mitigate repeated testing effects in a pre–post control group experimental design. High school students were administered an assessment of their attitudes and knowledge about suicide, both before and 3 months after a suicide prevention intervention. Students in the control group did not receive the intervention. At pretest, some students were assigned the complete assessment, whereas others received only a small subset of the items. At posttest, all students were given the full assessment. This study revealed several notable findings. First, students who were given a longer assessment were more likely to leave questions blank, resulting in a higher rate of *unplanned* missing data. Second, students who completed the entire pretest assessment were 3 times more likely to skip the posttest assessment entirely (21% vs. 7%). Third, of the students who attempted the posttest questionnaire, those who had received the full questionnaire at baseline were again less likely to complete all items at posttest (25% vs. 15%). Finally, those in the no-intervention control group who received the complete assessment at pretest improved statistically significantly from pretest to posttest, despite not receiving the intervention, whereas those controls who received the truncated assessment at pretest showed no such learning. Thus, the use of a planned missing design at pretest can substantially reduce rates of unplanned missing data and lead to larger treatment effects by minimizing the effect of repeated testing.

BETWEEN-BLOCK VERSUS WITHIN-BLOCK STRATEGIES

Two main strategies are available to assign items to forms. *Between-block* designs assign whole scales to forms, whereas *within-block* designs distribute items from each scale across forms. Table 2 depicts a simple example in which four scales (S1, S2, S3, S4) contain four items each (I1, I2, I3, I4), for a total of 16 items (S1I1 through S4I4).

Each of the two strategies has pragmatic and statistical benefits, as well as drawbacks. Statistical considerations tend to favor the within-block strategy because it typically results in much less loss of information (Graham, Hofer, & MacKinnon, 1996). Each observation in a data set contains information, some of which is shared with other observations and some of which is unique. A variable that is perfectly correlated with another variable in the data set (or is a linear combination of several other variables) contributes no new information, so if it is missing then nothing is lost. At the other extreme, a variable that is perfectly

TABLE 2
Within-Block Versus Between-Block Designs

Within-Block Design				*Between-Block Design*			
X	*A*	*B*	*C*	*X*	*A*	*B*	*C*
S1 I1	S1 I2	S1 I3	S1 I4	S1 I1	S2 I1	S3 I1	S4 I1
S2 I1	S2 I2	S2 I3	S2 I4	S1 I2	S2 I2	S3 I2	S4 I2
S3 I1	S3 I2	S3 I3	S3 I4	S1 I3	S2 I3	S3 I3	S4 I3
S4 I1	S4 I2	S4 I3	S4 I4	S1 I4	S2 I4	S3 I4	S4 I4

Note. S1–S4 = Scales 1 to 4; I1–I4 = Items 1 to 4.

*un*correlated with all other variables in the data set contributes wholly unique information, so if it is missing then that information is entirely lost. Of course, most variables are somewhere in between, containing a mix of shared and unique information: The higher a variable's correlation with other variables in the data set, the less information loss will occur if it has missing values. Recall that more information loss means higher standard errors, larger confidence intervals, and lower power to test hypotheses.

Following this principle, Raghunathan and Grizzle (1995) advocated using the partial correlations between pairs of variables (estimated from pilot data) to assign items to subsets. Partial correlations reflect the information shared between two variables controlling for all other variables in the dataset. A high partial correlation between two variables thus means that these two variables share information over and above that which is shared by any other variable in the data set. To achieve maximum efficiency (i.e., minimum information loss), the partial correlations between items within a subset should be low, whereas those between subsets should be high. Raghunathan and Grizzle tested this principle in a simulation study and found that when the average within-set partial correlations were .1 and between-set partial correlations were .8, standard errors of regression coefficients were just 1.2% larger than they were with complete data. In contrast, when both within- and between-set partial correlations were low, standard errors were 32% larger than with complete data. Assignment of items to sets can thus make a substantial difference to efficiency.

Variables that measure the same construct (e.g., a set of items on a single scale) will tend to be highly correlated with each other; variables measuring different constructs (e.g., items on different scales) will not. Thus, to maximize information in a planned missing design, it helps if individual participants are missing only *some* items on each scale and the other items are observed. The missing items share a great deal of information with the observed scale items, resulting in a small amount of missing information. In contrast, when all of the items on a scale are set to missing, these share relatively little information with the remaining observed items on other scales, resulting in a greater amount of missing information.

More completely, which strategy is more efficient depends on what parameter is being estimated. Adigüzel and Wedel (2008) and Chipperfield and Steel (2009) presented algorithms for optimizing item assignment given estimates of the correlations among variables. These authors assumed that the parameters of interest were means and variances of scale scores, rather than relations between scales. For this purpose, their simulations found that the *between-block* strategy is most efficient. In latent variable designs, where constructs are modeled as latent variables with scale items as indicators, the most efficient strategy depends on what parameters are of most interest: Between-block designs lead to most efficient estimates of factor loadings and residual variances, whereas within-block designs lead to most efficient estimates of *structural parameters*, including regression coefficients and correlations among latent factors (Jorgensen et al., 2014; Rhemtulla, Savalei, & Little, 2015).

Practically speaking, the within-block strategy has the benefit of making an assessment more manageable for participants. Adigüzel and Wedel (2008) reported that within-block designs are perceived to be shorter, less boring, and less repetitive than both between-block and complete data designs because participants are spared some of the redundancy of responding to every item on each scale. On the other hand, a practical benefit of the between-block design is that it is possible to carry out some analyses using available case analysis (i.e., listwise deletion). Graham (2012) vigorously recommended the between-block strategy for this reason, despite the statistical advantages of the within-block design. Another benefit of the between-block strategy is that it avoids possible order effects. Johnson, Roth, and Young (2011) compared data on the National Survey of Fertility Barriers collected using a complete data design to a within-block multiform design. They found similar results across the two designs but noted that when items in the middle of a scale were omitted, this affected responses to items later on the scale. In the end, in addition to pragmatic issues of cost, educational researchers must weigh the inferential demands of their research questions against the needs and limitations of the individuals in the population under study in order to choose the most appropriate design to employ.

HOW TO USE THE X SET

A very important component of the multiform design is the X set, which contains variables that are administered to every participant. Although it is possible to use a multiform design with no such items (i.e., every item has some planned missingness), simulation results suggest that including an X set brings an advantage in terms of efficiency (Rhemtulla et al., 2015). The biggest efficiency advantage comes from assigning some items in every scale to the X set so that some information about every construct remains available. Indeed, a good strategy would be to put the most reliable items on each scale in the X set in order to be assured a set of robust variables with no missing data. Important covariates (e.g., demographic variables) might also be placed in the X set. A common recommendation is that variables that are particularly important to the research hypotheses should be included in the X set (Graham, Taylor, Olchowski, & Cumsille, 2006). For example, if a study is designed to investigate predictors of reading comprehension, it would make sense to put the entire reading comprehension measure in the X set to maximize power to detect predictive relations. Similarly, there may be other key variables that should be collected from all participants. Huff, Anderson, and Tambling (2015), for instance, gave the example of suicidal ideation as an item that should go in the X set because of its importance for clinical use.

Another consideration is *how many* items to place in the X set. The X set could be the same size as the other sets (e.g., Pettigrew et al., 2015), but there is no particular reason that this must be the case. If the research hypotheses are not well specified, for example, a researcher might want to hedge her bets by including many potentially important variables in the X set. Or, if the sample size available is not very large and power is a concern, a more conservative amount of missingness would be appropriate. For example, Reitz et al. (2015) included 50% of their items in the X set. On the other hand, if sample size is less of a concern than the length of the assessment, fewer items in the X set means a greater reduction in test length, allowing more items in total to be measured.

HOW MANY FORMS?

The most common multiform design is the three-form design with four item sets (as shown in Table 1), but it is worth considering other variations. In general, the more item sets that are used, the greater the reduction in assessment length that can be achieved. Regardless of how many sets are used, each form should be composed of two sets in addition to the X set (e.g., one form is composed of sets A and B, another is A and C, and another is B and C); failing to do so would render it impossible to estimate relations between variables in different non-X sets (e.g., between items in the A and B sets). Thus, if s sets are used, each participant completes $2/s$ of the items (assuming all sets contain the same number of items) and the total number of forms is the number of combinations of two sets out of all s, $\frac{s!}{2(s-2)!}$.

Because each form represents a single combination of two item sets, the more sets and forms there are, the greater the proportion of missing data and the smaller the proportion of participants who provide data on any pair of variables in different item sets. For example, in a six-form design with item sets A to D, the relation between an item in the A set and an item in the B set is estimated based on the $N/6$ participants who get the XAB form. Thus, if relations among variables between sets will be an integral part of the analysis, it is probably wise to limit the number of sets (unless the total sample size is very large). Table 3 gives the number of forms, proportion of missing data, and rate of pairwise coverage when three through six non-X sets are used.

TABLE 3
Number of Forms and Proportion Missing for Three- to Six-Item Sets

Sets	*Forms*	*% Items Completed*	*Pairwise Coverage*
3	3	67%	1/3
4	6	50%	1/6
5	10	40%	1/10
6	15	33%	1/15

Note. The X set is not included. Sets = the number of non-X item sets; Forms = the number of combinations of two sets. % Items Completed = the proportion of total items completed by any one participant; Pairwise Coverage = the proportion of participants who provide data on any pair of two items in different sets.

EXTENDING THE MULTIFORM DESIGN LONGITUDINALLY

Several longitudinal studies have implemented multiform designs within one or more measurement occasions (e.g., Conrad-Hiebner, Schoemann, Counts, & Chang, 2015; Flay, Graumlich, Segawa, Burns, & Holliday, 2004; Hecht et al., 2003; Lin, Crnic, Luecken, & Gonzales, 2014). If a multiform design is to be used at multiple occasions, a design consideration that arises is how to assign participants to forms over time. The easiest approach may be to randomly assign participants to forms at every occasion, eliminating the need to keep track of which individuals get which forms at each time point. But it is also possible to assign participants to the same form at each occasion, or systematically rotate forms across occasions (these strategies may be more easily implemented when data are collected online). Jorgensen et al. (2014) investigated this issue using data simulated from a longitudinal latent variable mediation model, in which three-form missingness was applied to indicators of three latent constructs (X, M, Y) at three time points. They found that when the within-block strategy was used to assign items to forms (i.e., scales were split across item sets), assigning participants to the same or different forms over time made no difference to parameter estimate efficiency. However, when the between-block strategy was used (i.e., scales were kept together within item sets), then efficiency was improved by assigning participants different forms across occasions. In addition, if repeated-testing effects are a concern, then assigning participants to different forms over time may attenuate some of these effects or allow them to be modeled.

MULTIFORM DESIGNS: OTHER EDUCATIONAL ISSUES

Educational researchers might be interested in collecting data to assess or diagnose individual students as in a single-subject design, especially in settings where students have special needs and/or are considered potentially at risk. Huff et al. (2015) explored whether a multiform design could be used in clinical assessments when the goal is to use an individual's assessment score for categorization (below/above a clinical cutoff) or for classifying whether individuals improved or failed to improve over the course of therapy. This use of planned missingness is very different from the typical implementation of planned missing designs, where the goal is usually to do an analysis at the group level for making inference about a population—for example, to investigate group means or relations among constructs. When an individual score is needed, and especially when it is needed immediately (as when an assessment is given for the purpose of diagnosing/classifying a student), sophisticated missing data methods are not available. Huff et al. imposed 25% and 50% planned missingness on complete data and compared the scores to those based on complete data. They used item-mean imputation (i.e., averaging over the available items) to get a score for each individual, reflecting the assumption that the items are interchangeable. Unsurprisingly then, with real clinical scales, this method resulted in statistically significant differences between the planned missing design and complete data scores. Huff et al. advised that planned missing designs be used in this type of situation only when scales have high internal reliability.

LONGITUDINAL WAVE MISSING DESIGNS

Cross-sectional designs can be used to measure differences among age groups or grade levels, but if one is interested in *change*, these designs are problematic. For one, cohort and age differences are confounded, so it can be impossible to know whether an observed difference in age groups is due to intraindividual change or to preexisting differences between the groups. In addition, though cross-sectional designs can be used to study group mean differences, they do not support the examination of interindividual differences in rates of change. Longitudinal designs, therefore, in which participants are measured at several time points, are typically regarded as the gold standard method for studying change. Notwithstanding, such designs also come with practical and theoretical disadvantages. The expense and organization required to collect longitudinal data can lead to designs in which too many measures are included, putting a high burden on participants. Finding participants who are willing to participate over a long periods can lead to pronounced selection effects (Bell, 1953). High rates of attrition may dramatically change the sample characteristics over the course of the study. The mere fact of being repeatedly measured could, depending upon the outcome variable, also induce change in participants' behavior over time.

For the aforementioned reasons, longitudinal designs in which data are not collected from all participants at all measurement occasions, herein referred to as *wave missingness*, can have considerable benefits in terms of cost (fewer participants must be measured at each occasion), as well as repeated testing effects (each participant is measured fewer times). Missingness can be distributed across occasions in any way that suits the research goals: A greater proportion of the total sample might be measured at the early waves, a constant proportion might be measured at each wave, or a greater proportion might be measured at certain key time points. The optimal assignment of missingness to measurement occasions depends on the research questions and the analysis models.

Although wave missing designs have been studied via simulation (Graham, Taylor, & Cumsille, 2001; Mistler & Enders, 2012; Rhemtulla, Jia, Wu, & Little, 2014; Wu, Jia, Rhemtulla, & Little, 2015), planned wave missing designs do not yet appear to have been implemented in educational research. Wave missingness does, however, routinely arise in educational research as a consequence of restructuring data from other popular designs (e.g., accelerated longitudinal designs and test–retest designs), as we hope to make clear next. These data configurations do not constitute true planned missing data designs because participants are not randomly assigned to missing data patterns, but they share many characteristics. We begin this section by considering some of these configurations in which complete data designs can be transformed into quasi-planned-missing designs; we term this *configural wave missingness*. Then we describe the research on planned wave missing designs and discuss how these designs could be applied to educational research.

CONFIGURAL WAVE MISSINGNESS

Accelerated longitudinal designs combine the benefits of cross-sectional and longitudinal designs by tracking several cohorts over a short time span (Bell, 1953). The resulting data have an overlapping configuration that facilitates inferences regarding longitudinal change over a time span longer than the duration of the study. For example, Poteat, O'Dwyer, and Mereish (2012) investigated the trajectory of adolescents' use of homophobic epithets through high school. Six overlapping cohorts of students (i.e., those in Grades 7–12) were measured on four occasions over 2 years. The resulting data were stretched out and overlapped to get, effectively, a spliced-together picture of the developmental trend from beginning to end of high school. Table 4 depicts their research design in terms of data collection and analysis, whereby 2 years of data collection are stretched out to cover a 5-year developmental change trajectory.

Accelerated longitudinal designs are not true planned missing designs because participants are not randomly assigned to cohorts. As such, *cohort effects* can arise when not all cohorts conform to the same change trajectory. Miyazaki and Raudenbush (2000) gave an example of an accelerated longitudinal design in which cohort effects were present. They identified seven cohorts of participants in the National Youth Survey according to their age at the survey's onset (11–17 years), each of which was followed for the 5-year duration of the study. In total, these data described longitudinal change from 11 to 21 years in

TABLE 4
Accelerated Longitudinal Design

	Data Collection		*Analysis*				
Cohort	*Year 1*	*Year 2*	*Grade 7/8*	*Grade 9*	*Grade 10*	*Grade 11*	*Grade 12*
1		Grade 7/8	Year 2				
2	Grade 7/8	Grade 9	Year 1	Year 2			
3	Grade 9	Grade 10		Year 1	Year 2		
4	Grade 10	Grade 11			Year 1	Year 2	
5	Grade 11	Grade 12				Year 1	Year 2
6	Grade 12						Year 1

attitudes favorable toward deviant behavior. Their analyses revealed significant differences among the cohort trajectories, such that older cohorts showed less steep increases in favorability than younger cohorts. The authors noted that the more overlap between the cohorts (e.g., each cohort in the NYS data overlapped with another cohort on 4 of 5 years), the higher the power to find such differences. If statistically significant cohort effects do exist, then inferences about change, including predictors of change, cannot necessarily be generalized across all cohorts. That said, Mizayaki and Raudenbush also argued that accelerated designs have several advantages over true longitudinal designs, including that they accrue less attrition; that age and history effects can be disentangled; and that contamination associated with frequent repeated measurements of true longitudinal designs, especially if the time points are close, can be mitigated.

Similar to accelerated longitudinal designs, *wave-to-age restructuring* of longitudinal designs takes data collected at a number of fixed waves (e.g., Grades 1–4) and restructures them into a set of variables that reflect students' ages (e.g., ages 6–10) instead of grade level. The resulting data set looks very similar to that of the accelerated longitudinal design, but instead of discrete cohorts the missing data patterns now correspond to different ages of students relative to their classmates. Beyond wave and age, data can also be restructured according to alternative metrics of time, such as months or half years (e.g., to model more narrow age ranges), stages of development, or number of hours of instruction on a task (Bollen & Curran, 2006).

Finally, *test–retest designs* can be another source of configural wave missing data when the data are reconfigured to account for the time lag between the initial test and retest. McArdle and Woodcock (1997) described a scenario in which students were given a test–retest memory assessment during which they learned picture–name associations at the initial test and were asked to recall these associations during a retest occurring 1 to 8 days later. Although it would be possible to analyze these data by ignoring the differential time lag between test and retest, such an analysis would fail to take advantage of the potentially rich information that can be extracted from the time lag data. Table 5 shows how the data can be restructured to look like a planned missing design in which all participants are measured at baseline (t_0) and at one other time point (t_1–t_8). Structuring the data in this way allows the researcher to model a change trajectory over the course of time (in this case, 8 days), and even to disentangle the effect of time from the effect of practice (McArdle & Woodcock, 1997).

PLANNED WAVE MISSINGNESS

In planned wave missing designs, participants are randomly assigned to a particular pattern of missing data across the waves of a longitudinal study. In contrast to the configural wave missing designs described in the previous section, planned wave missing designs allow the researcher to choose patterns of missingness that lead to efficient parameter estimation while controlling cost and considering the effects of repeated measurement. Planned wave missing designs have several benefits over configural wave missing designs. First, because the patterns of missingness are randomly assigned, they cannot be related to meaningful effects in the data, such as cohort or age effects. Second, whereas accelerated longitudinal designs and wave-to-age transformations tend to create a lot of missing data at the first and last measurement occasions, planned wave missingness allows the missing data to be concentrated in the middle waves, leading to more stable estimation of linear and quadratic growth trajectories (Graham et al., 2001; Mistler & Enders, 2012).

TABLE 5
Developmental Time-Lag Design

Time Point								
t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
test	retest							
test		retest						
test			retest					
test				retest				
test					retest			
test						retest		
test							retest	
test								retest

Latent growth curve models are an especially important context for these types of data designs. Such models allow researchers to examine the nature of, individual differences in, and determinants of change over time. Jordan, Kaplan, and Hanich (2002), for example, studied the growth trajectories in math and reading achievement at four data collection waves across Grades 2 and 3 and found different growth trajectories for students with learning difficulties in math and reading. With regard to the effect of different patterns of wave missingness on the efficiency of parameter estimates in latent growth curve models, Graham et al. (2001) considered a linear growth trajectory with five measurement occasions, where the linear slope was predicted by a program versus control grouping variable. They compared several missingness patterns with respect to the efficiency of the slope-on-group regression coefficient. Their most efficient design had no missing data on the initial occasion, 30% missing data on the last occasion, and 50% missing data on the middle three occasions, for a total of 36% missing data. This design resulted in the same power to detect the regression effect as a complete data design would have with 83% of the participants. Mistler and Enders (2012) considered the effectiveness of similar designs in estimating the slope mean in linear and quadratic growth curve models, as seen in Table 6. These authors found that designs with missing data confined to the middle measurement occasions produced more efficient estimates than those that also included missing data on the first and last occasions. The results of these two studies together suggest that the end points of a curve may contain more information than the middle points when estimating a growth trajectory.

Rhemtulla et al. (2014) examined the effect of wave missing designs on multivariate latent growth curve models (McArdle, 1988). In these models, researchers may investigate whether the rate of change in one construct over time (e.g., vocabulary acquisition) is correlated with the rate of change of another construct (e.g., reading frequency) by modeling multiple growth curves simultaneously. The authors considered the most flexible growth pattern in which the shape of the growth trajectory is freely estimated (as opposed to fixing a linear or quadratic rate of change), a so-called *unspecified* trajectory (e.g., Hancock, Harring, & Lawrence, 2013). In a simulation study, wave missingness was imposed according to the optimal pattern reported by Graham et al. (2001), and bias, efficiency, and power were compared to a complete data design. In contrast to the results of Graham et al. and Mistler and Enders (2012), Rhemtulla et al. found that the efficiency loss that resulted from imposing wave missingness was often *greater* than it would have been under a complete data design with reduced sample size. With 36% missing data in the wave missing design, standard errors of the slope parameter estimates (i.e., slope means, variances, and covariance) were less than 50% as efficient as with complete data. The authors concluded that wave missing designs should be used with caution if the rate and shape of change, and correlations among the rates of change, are desired. These findings are particularly concerning in light of research suggesting that tests of slope covariances tend to be extremely underpowered even with complete data (Hertzog, Lindenberger, Ghisletta, & von Oertzen, 2006; Hertzog, von Oertzen, Ghisletta, & Lindenberger, 2008). More methodological research is clearly needed to reconcile these various findings.

TABLE 6
Wave Missing Designs for Linear and Quadratic Growth

	Measurement Occasion (Wave)					
Group	*1*	*2*	*3*	*4*	*5*	*6*
A				▨	▨	
B			▨		▨	
C		▨			▨	
D			▨	▨		
E		▨		▨		
F		▨	▨			

Note. This table represents an example of the type of design discussed by Mistler and Enders (2012). Gray is missing data; white is complete data.

It is important to keep in mind that each of the aforementioned findings is specific to the parameters and models that were studied. Other model parameters and other models would be expected to have a different optimal missingness pattern. For example, a growth process that is characterized by a spline function (e.g., two linear functions connected at a transition point) relies heavily on information at the point where the change in direction happens. To reliably estimate the shape of change, a planned missing design would ideally have complete data at the point of change, with missingness distributed among other time points (Hogue, Pornprasertmanit, Fry, Rhemtulla, & Little, 2013). For education researchers examining the transition from middle school to high school, for example, where a spline model would seem especially appropriate to capture potentially abrupt trajectory changes, collecting complete data around the eighth-grade to ninth-grade transition point would be well advised.

Wu et al. (2015) proposed an algorithm that searches for the most efficient pattern of planned missing data for each parameter in a given model. When planning research with a particular analysis in mind, the algorithm simulates data based on initial estimates of the relations among variables, imposes a wide range of missingness patterns, and compares the results in terms of the efficiency of each parameter estimate. This tool may be useful for researchers who have a precise idea of what analyses they plan to run and which parameters will be of greatest interest.

TWO-METHOD MEASUREMENT DESIGNS

A third and potentially very useful planned missing data design, the *two-method measurement design*, was proposed by Graham et al. (2006) for a very specific situation. Imagine that a researcher has access to two types of measures of a construct—one of these is a gold-standard, which is unbiased but expensive in cost and/or effort (e.g., a direct observation of student attentiveness), and the other is a systematically biased and typically more error laden, but quick and less expensive, approximation (e.g., teacher-report of student attentiveness). Data are gathered on the set of inexpensive measures, whereas only a fraction of the participants are randomly assigned to receive the gold standard measure(s) as well. The motivating idea behind two-method measurement is that including an excellent but expensive type of measure within a small subset of the larger sample allows the degree of bias in the inexpensive measures to be calibrated and thus statistically controlled. Bias is modeled using a structural equation model that separates latent bias from the latent construct.

Consider the model shown in Figure 1, where the goal is to assess the effect of classroom attentiveness in Grade 1 on children's reading achievement in Grade 2. Three of the classroom attentiveness measures are direct assessments, observing children in the classroom over strategically chosen periods; the other two measures are less reliable and potentially biased teacher reports. Using the observational assessments alone, although ideal, would be prohibitively costly and as such lead to a small sample size and inadequate power to assess the impact of classroom attentiveness on reading achievement. On the other hand, using the teacher measures alone would be expected to confound the desired construct of classroom attentiveness with a method construct (we simply refer to this as teacher-report bias), yielding an inaccurate evaluation of the key structural relation of interest. Hence, it is desirable to parse the teacher-report bias method factor from classroom attentiveness so that the effect of the latter on the subsequent reading achievement score may be accurately assessed. This may be accomplished using a model such as that shown in Figure 1, where 125 students are randomly assigned to have all four measures gathered and an additional 375 students are randomly assigned to have only the inexpensive measures gathered. Together, the gold standard measures help to purge classroom attentiveness of the reporting bias inherent in the teacher-report measures, whereas the additional subjects receiving only the teacher-report measures help to increase the power to detect the now accurately calibrated relation between classroom attentiveness in first grade and reading achievement in second grade.

It should be emphasized that the most important aspect of two-method measurement is that the inexpensive measure is a systematically biased measure of the same construct that is assessed by the expensive measure. In the context of education research, the expensive measure will often be a direct assessment or observation, whereas the inexpensive measure could be a paper-and-pencil or computer-based measure. Examples might include an individualized administration of the Wechsler Intelligence Test for Children versus a written intelligence test, or an intensive repeated observation of classroom aggression versus a teacher report. This means that the two measures must measure the same thing (i.e., a single underlying trait, aptitude, or characteristic gives rise to both these measures), and the inexpensive measure must be contaminated by some kind

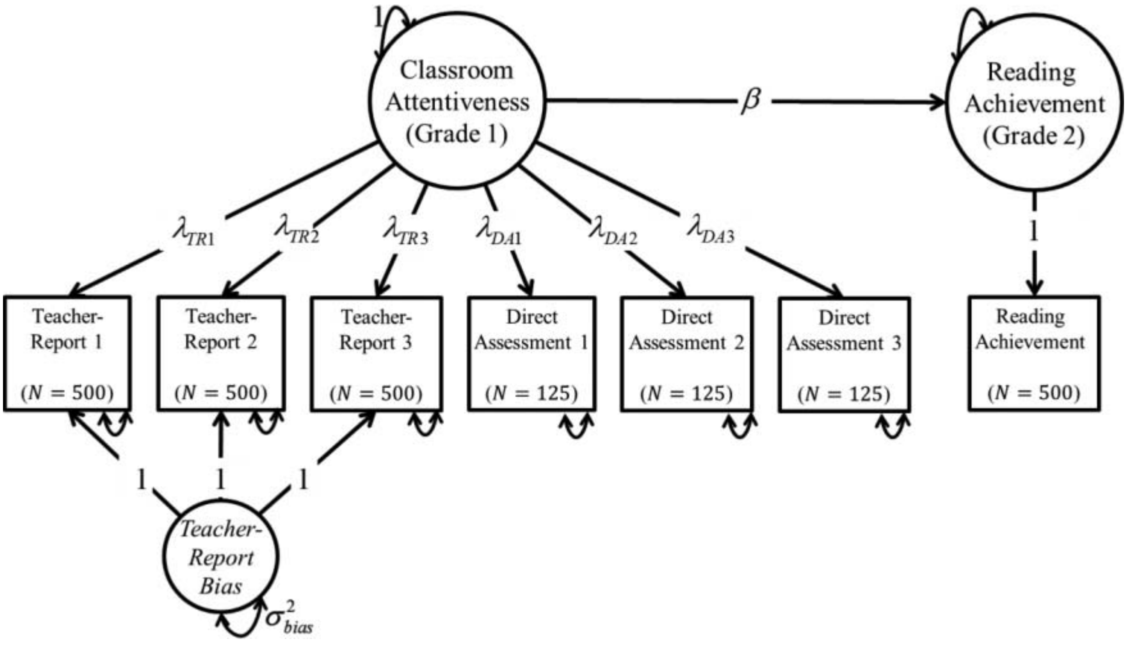
1
Classroom Attentiveness (Grade 1)
β
Reading Achievement (Grade 2)
λ_{TR1}
λ_{TR2}
λ_{TR3}
λ_{DA1}
λ_{DA2}
λ_{DA3}
1
Teacher-Report 1 (N = 500)
Teacher-Report 2 (N = 500)
Teacher-Report 3 (N = 500)
Direct Assessment 1 (N = 125)
Direct Assessment 2 (N = 125)
Direct Assessment 3 (N = 125)
Reading Achievement (N = 500)
1
1
1
Teacher-Report Bias
σ^2_{bias}

FIGURE 1 Structural equation model for two-method measurement design data.

of systematic measurement bias (e.g., self-report bias). If the inexpensive measure is unreliable but not systematically biased, it will typically be more efficient to use only the inexpensive measure and gather data on a larger sample. As for the relative amount of each type of data needed, useful insights into the trade-offs in terms of the measures' cost ratio and reliability are provided by Graham et al. (2006) and Graham (2012).

LOOKING AHEAD

For researchers considering employing planned missingness designs, two final issues are worth mentioning: determining adequate sample size, and combining multiform and two-method measurement designs with designs that impose missingness over multiple time points. First, as with all studies, one should be able to determine the sample size needed to have adequate statistical power. A complicating factor of planned missing data designs is that routine power analysis methods for sample size planning cannot easily be applied. Calculations based on complete data must be adjusted to account for missing data, but this adjustment is not straightforward; as mentioned earlier, the power to test the statistical significance of a particular parameter estimate depends on the covariance structure of the data (i.e., the strength of relations among variables), the pattern of missing data, and the missingness mechanism.

Perhaps the most direct way to estimate power accurately, taking all these factors into account, is to use a Monte Carlo simulation approach (see, e.g., Enders, 2010), which can be done in software such as Mplus (Muthén & Muthén, 1998–2016) and the open-source software R (R Core Team, 2015). This method has five steps: (a) specify a hypothesized population, including all model parameter values; (b) draw a large number of samples of some arbitrarily chosen size N from this hypothesized population; (c) impose the proposed pattern of planned missing data on each sample; (d) carry out an alpha-level significance test for a given parameter of interest within each sample (e.g., a test of a regression coefficient, or a test of model fit); and (e) count the proportion of samples in which the test returned a statistically significant result—this value represents an estimate of power for samples of size N. The procedure then cycles repeatedly, trying different sample sizes until achieving a desired level of power (e.g., .80). The whole method is then repeated in turn for each of the other parameters of interest, ultimately choosing the largest of the resulting sample sizes across all key parameters.

A further benefit of the simulation approach to sample size planning is that it is fully flexible with respect to not only planned missing data patterns but also foreseeable *unplanned* missingness. For example, a researcher planning a four-wave study with wave missingness could investigate the effect of attrition in addition to the planned missingness. Based on previous substantive research in her field, this researcher could evaluate whether she would still have acceptable power with 10%, 20%, or 30% attrition, with results possibly persuading her to increase her sample size and/or to use less planned missing data.

The second issue to be mentioned as we look ahead has to do with the designs themselves. So far we have addressed designs that impose missingness patterns at a single time point, including multiform and two-method measurement designs, and designs that impose missingness over multiple time points, including accelerated longitudinal and more general wave missing designs. Although not common, one could, in fact, draw upon aspects of multiple such designs simultaneously, should it be useful to balance cost and efficiency. Reitz et al. (2015), for example, studied adolescent sexual development over 2 years, with four waves of data each 6 months apart. The authors implemented a three-form design in the initial two time points and collected complete data at the last two time points, the latter helping to mitigate the loss of power due to attrition over time. Because participants ranged in age from 10 to 18 at the study onset, this design would support an accelerated longitudinal analysis of the data in which a continuous developmental trajectory from ages 10 to 18 could be modeled. As another example, Garnier-Villarreal, Rhemtulla, and Little (2014) considered a longitudinal extension of the two-method measurement design, in which the inexpensive/biased measure was administered over four waves of data collection and the expensive measure was included at one, two, or all four occasions. Simulations revealed that, as long as the degree of measurement bias was about the same at each measurement occasion, including the expensive measure at just the first occasion was enough to get the accuracy and efficiency benefits of two-method measurement approach.

The two examples just presented help to illustrate the tremendous potential for planned missing designs, whereby clever adaptations can save research expense while ensuring the validity of, and statistical power necessary to detect, relations of key research interest. We look forward to the many interesting methodological developments to come, and more important their informative implementations within the education and educational psychology literature.

REFERENCES

Adigüzel, F., & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*, 608–617. http://dx.doi.org/10.1509/jmkr.45.5.608

Bell, R. Q. (1953). Convergence: An accelerated longitudinal approach. *Child Development*, *24*, 145–152. http://dx.doi.org/10.2307/1126345

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation modeling perspective*. Hoboken, NJ: Wiley.

Chipperfield, J. O., & Steel, D. G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, *25*, 227–244.

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.

Echols, L. (2015). Social consequences of academic teaming in middle school: The influence of shared course taking on peer victimization. *Journal of Educational Psychology*, *107*, 272–283.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Enders, C. K. (2013). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 493–519). Charlotte, NC: Information Age.

Flay, B. R., Graumlich, S., Segawa, E., Burns, J. L., & Holliday, M. Y. (2004). Effects of 2 prevention programs on high-risk behaviors among African American youth: A randomized trial. *Archives of Pediatrics & Adolescent Medicine*, *158*, 377–384. http://dx.doi.org/10.1001/archpedi.158.4.377

Garnier-Villarreal, M., Rhemtulla, M., & Little, T. D. (2014). Two-method planned missing designs for longitudinal research. *International Journal of Behavioral Development*, *38*, 411–422. http://dx.doi.org/10.1177/0165025414542711

Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-4018-5

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*, 197–218. http://dx.doi.org/10.1207/s15327906mbr3102_3

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). *Analysis with missing data in drug prevention research* [Monograph]. Rockville, MD: National Institute on Drug Abuse. http://dx.doi.org/10.1037/e495862006-003

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in the analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-011

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

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 309–341). Charlotte, NC: Information Age.

Harel, O., Stratton, J., & Aseltine, R. (2012). Designed missingness to better estimate efficacy of behavioral studies (Tech. Rep. 11–15). Storrs: Department of Statistics, University of Connecticut.

Hecht, M. L., Marsiglia, F. F., Elek, E, Wagstaff, D. A., Kulis, S., Dustman, P., & Miller-Day, M. (2003). Culturally grounded substance use prevention: an evaluation of the keepin' it R.E.A.L. curriculum. *Prevention Science*, *4*, 233–248. http://dx.doi.org/10.1023/a:1026016131401

Hertzog, C., Lindenberger, U., Ghisletta, P., & Oertzen, T. V. (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological Methods*, *11*, 244. http://dx.doi.org/10.1037/1082-989x.11.3.244

Hertzog, C., von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling*, *15*, 541–563. http://dx.doi.org/10.1080/10705510802338983

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*, 310–325.

Huff, S. C., Anderson, S. R., & Tambling, R. B. (2015). Testing the clinical implications of planned missing data designs. *Journal of Marital and Family Therapy*, *42*, 313–325. doi:10.1111/jmft.12129

Johnson, D. R., Roth, V., & Young, R. (2011). Planned missing data designs in health surveys. In *Tenth conference on health survey research methods*. Hyattsville, MD: National Center for Health Statistics.

Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology*, *94*, 586–597. http://dx.doi.org/10.1037//0022-0663.94.3.586

Jorgensen, T. D., Rhemtulla, M., Schoemann, A., McPherson, B., Wu, W., & Little, T. D. (2014). Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies. *International Journal of Behavioral Development*, *38*, 397–410. http://dx.doi.org/10.1177/0165025414531094

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*, 571–582. http://dx.doi.org/10.1016/j.infbeh.2014.07.001

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: Wiley. http://dx.doi.org/10.1002/9781119013563

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The handbook of multivariate experimental psychology* (Vol. 2, pp. 561–614). New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4613-0893-5_17

McArdle, J. J., & Woodcock, R. W. (1997). Expanding test-retest designs to include developmental time lag components. *Psychological Methods*, *2*, 403–435.

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

Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, *5*, 44–63. http://dx.doi.org/10.1037//1082-989x.5.1.44

Muthén, L. K., & Muthén, B. O. (1998–2016). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Pettigrew, J., Graham, J. W., Miller-Day, M., Hecht, M. L., Krieger, J. L., & Young, J-S. (2015). Adherence and delivery: Implementation quality and program outcomes for the seventh-grade keepin' it REAL program. *Prevention Science*, *16*, 90–99. http://dx.doi.org/10.1007/s11121-014-0459-1

Poteat, V. P., O'Dwyer, L. M., & Mereish, E. H. (2012). Changes in how students use and are called homophobic epithets over time: Patterns predicted by gender, bullying and victimization status. *Journal of Educational Psychology*, *104*, 393–406. http://dx.doi.org/10.1037/a0026437

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*, 54–63. http://dx.doi.org/10.2307/2291129

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Reitz, E., van de Bongardt, D., Baams, L., Doornwaard, S., Dalenberg, W., Dubas, J., . . . Dekovic, M. (2015). Project STARS (Studies on Trajectories of Adolescent Relationships and Sexuality): A longitudinal, multi-domain study on sexual development of Dutch adolescents. *European Journal of Developmental Psychology*, *12*, 1–14. http://dx.doi.org/10.1080/17405629.2015.1018173

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*, 423–434. http://dx.doi.org/10.1177/0165025413514324

Rhemtulla, M., Savalei, V., & Little, T. D. (2015). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, *81*, 60–89. doi:10.1007/s11336-014-9422-0

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

Swain, M. S. (2015). *The effects of a planned missingness design on examinee motivation and psychometric quality* (Unpublished doctoral dissertation). Harrisonburg, VA: James Madison University.

Wu, W., Jia, F., Rhemtulla, M., & Little, T. D. (2015). Search for efficient complete and planned missing data designs for analysis of change. *Behavioral Research Methods*. Advance online publication. doi:10.3758/s13428-015-0629-5