Psychological Test and Assessment Modeling, Volume 59, 2017 (1), 55-70

Design considerations for planned missing auxiliary data in a latent regression context

Leslie Rutkowski[1]

Abstract

Although variations of a multiple-matrix sampling approach have been used in large-scale assessments for the design of achievement instruments, it is only recently that item sampling has been used to extend content coverage of the student background questionnaire. In 2012, PISA implemented a so-called *3-form design*, whereby four sets of background questionnaire items were administered. This design reduced the time required to respond to each questionnaire by about 25% (30 minutes compared to 41 minutes, for all questions). An open problem for future rounds and assessments surrounds whether and how to deal with missing background data when unbiased and sufficiently precise achievement estimation is paramount. Imputation of background questionnaire data prior to estimating achievement is one means for treating these data; however, concerns over a sensible imputation model and preserving the quality of achievement estimates loom large. In the current paper, I take one step back and consider a precursor to statistical solutions for planned missing data. That is, I discuss possible questionnaire designs that create a more reasonable foundation from which to impute missing background questionnaire data. Among the design features discussed, I consider splitting constructs across questionnaires, planning missing among well-correlated constructs, and administering intensive questionnaires to a smaller subsample (so called "two-method" design). In each case, I consider the feasibility of each design against the backdrop of information gains and the multidimensional burden of preserving achievement distributions.

Key words: Planned missing designs, background questionnaires, rotated questionnaires

[1] *Correspondence concerning this article should be addressed to:* Leslie Rutkowski, PhD, Centre for Educational Measurement at University of Oslo, Postboks 1161 Blindern, 0318 OSLO, Norway; email: leslie.rutkowski@cemo.uio.no

A typical tension in the design and administration of large-scale educational assessments (LSAs) surrounds balancing respondent fatigue with a desire to collect as much information as possible. Indeed, policy makers and researchers are often interested in numerous aspects of education, putting pressure on testing organizations to increase the length of both achievement measures and non-achievement measures (administered through so-called *background questionnaires*), with sometimes undesirable consequences for respondents. And although reliable and valid measurement depends, at least partly, on sufficiently long instruments with good content coverage, diminishing returns necessarily result as time and human attention are limited. Solutions to this issue have historically been addressed using several versions of incomplete block designs, and have been geared toward the measurement of achievement domains of LSAs. As such, different models of incomplete matrix designs have been implemented in the National Assessment for Educational Progress (NAEP), the Trends in International Mathematics and Sciences Study (TIMSS), and the Programme for International Student Assessment (PISA). Multiple-matrix sampling (MMS; Shoemaker, 1973) is a method whereby a sample of examinees are administered a sample of items assigned to them in a complex and systematic fashion. This is in contrast to typical sampling procedures, which limits the sampling frame to a population of measurable units (in this case, examinees). Notable is that achievement is estimated for populations and sub-populations of examinees (rather than individuals). Since not all items are administered to every person, such an approach allows far greater coverage of the construct of interest, even if fewer people respond to any one item. As an example, the TIMSS 2011 eighth grade assessment was comprised of 434 total math and science items, distributed across 14 non-overlapping mathematics blocks and 14 non-overlapping science blocks. That is, the blocks exhaustively and mutually exclusively contained all available testing material. The blocks subsequently were arranged into 14 booklets containing two science and two mathematics blocks each, with no block-wise overlap within a booklet. In other words, no block appeared more than once within a booklet. This design, presented in Table 1, ensured linking across booklets since each block (and therefore each item) appeared in two different booklets. Further, the total assessment material (approximately 10 hours) was divided into more reasonable 90 minute periods of testing time for each student. Such an approach to item administration has consequences for achievement estimation, which I discuss subsequently.

The background questionnaire – administered to all participating students – is designed to elicit information on the context and correlates of learning (e.g., students' cultural and socio-economic background; their home context and experiences) as well as non-achievement constructs such as learning motivation and attitudes. In response to increased interest from researchers and policy makers, non-achievement measures have grown in importance in their own right, with commensurate development and refinement (OECD, 2016). These developments include limited implementation of item sampling in PISA 2012 in an effort to extend the content coverage of the student background questionnaire (e.g., in PISA 2012; OECD, 2014). In particular, PISA implemented a so-called *3-form design* (Graham, Hofer, & MacKinnon, 1996), whereby four sets of items (*X*, *A*, *B*, and *C*) are administered. Under this approach, set *X* is administered with combinations

Table 1:
TIMSS 2011 Booklet Design.

	Part 1		Part 2	
Booklet	Block 1	Block 2	Block 3	Block 4
1	M01	M02	S01	S02
2	S02	S03	M02	M03
3	M03	M04	S03	S04
4	S04	S05	M04	M05
5	M05	M06	S05	S06
6	S06	S07	M06	M07
7	M07	M08	S07	S08
8	S08	S09	M08	M09
9	M09	M10	S09	S10
10	S10	S11	M10	M11
11	M11	M12	S11	S12
12	S12	S13	M12	M13
13	M13	M14	S13	S14
14	S14	S01	M14	M01

Note. M indicates a mathematics block; S indicates a science block.

of item set *A*, *B*, or *C*, such that three background questionnaire booklets are assembled as *XAB*, *XBC*, or *XAC*, as in Figure 1. This had the effect that all students responded to set *X* and two-thirds of students responded to all other item sets. Regardless of form, the background questionnaires were designed to take 30 minutes to complete. This design reduced the time required to respond to each background questionnaire by about 25% (30 minutes compared to 41 minutes, for all questions). Such an approach to survey administration poses challenges due to the presence of missing data. In particular, under the 3-form design, about 25% of data are missing for each respondent. An advantage here is that the data are missing by design or, in the traditional typology of missing data, *missing completely at random* (MCAR). In other words, the missing mechanism does not depend

	Block			
Form	X	A	B	C
1	1	1	1	0
2	1	0	1	1
3	1	1	0	1

Note. 1 = questions asked; 0 = questions not asked.

Figure 1:
Background questionnaire design used in PISA 2012.

on the observed or missing values (Rubin, 1976). And although parameter estimates are not biased when data are MCAR, they exhibit decreased efficiency and power loss, if not properly treated. Under the design used in PISA 2012, for example, listwise deletion would result in significant loss of data when variables from blocks *A*, *B*, or *C* are used in an analysis. However, the overlap of blocks (e.g., each rotated block appears with every other block once in the design) ensures that booklets can, hypothetically, be linked, if such a goal exists, and pairwise deletion of data will allow for the computation of a complete correlation matrix between the background variables.

Although modern methods can hypothetically be brought to bear on these sorts of missing data problems, multiple imputation (MI) and full-information maximum likelihood (FIML) rely on correlations among variables to arrive at sensible parameter estimates (Schafer & Graham, 2002). And given the structure of this design, there is nothing built in that ensures reasonable correlations between any of the blocks. In other words, modern methods rely on relationships between, for example, block *A* and block *B* to sensibly impute data for block *B*. As one possible solution, construct items can be split across forms, rather than being left intact, to build a reasonable foundation upon which to impute or create scale scores (Graham et al., 1996). In what follows, these sorts of design possibilities are considered. In other words, I discuss possible questionnaire designs that create a more reasonable foundation from which to impute missing background questionnaire data, create scale scores on background constructs of interest, and use scales or individual variables for secondary analysis. Among the design features discussed, I consider splitting constructs across questionnaires, planning missing among well-correlated constructs, and administering intensive questionnaires to a smaller subsample (so called "two-method" design). In each case, I consider the feasibility of each design against the backdrop of information gains and the multidimensional burden of preserving achievement distributions. Importantly, I do not directly consider the problem of *treating* the missing data in one way or another. Rather, I take an approach that illustrates design features that are more or less conducive to forms of missing data treatment, scaling, and analysis.

Considerations for planned missingness in background questionnaires

As with any research design, employing planned missingness requires a careful account of the considerations that go into design selection. Although the following is not comprehensive, I discuss several key issues that are relevant to large-scale educational assessments regarding planned missingness in the student background questionnaires.

Achievement estimation

Population and sub-population achievement estimates are, historically, the raison d'être for large-scale assessments. Consequently, the models and methods used to estimate achievement figure prominently during the development of each assessment cycle, where

considerable resources are dedicated to ensure that achievement is validly measured with sufficient precision and reliability. This is clearly evidenced in TIMSS and PISA technical reports (Martin & Mullis, 2012; OECD, 2014) and assessment frameworks (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009; OECD, 2013). As such, preserving achievement distributions for population and subpopulations of examinees is, arguably, the most important consideration when selecting a plausible design for background questionnaire rotation and possible subsequent imputation or scaling.

As noted above, NAEP, TIMSS, and PISA all use variations of multiple-matrix sampling for the achievement assessment. Further, the methods used to estimate achievement rely on a *type* of imputation (Mislevy, 1984; Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009) that uses information from the student background questionnaire, other key demographic variables, and responses to the achievement portion of the assessment. This method, referred to as *latent regression* (von Davier, Sinharay, Oranje, & Beaton, 2006), or *population modeling* (von Davier & Sinharay, 2014), reliably and sufficiently precisely estimates population and sub-population achievement. And although initial research suggests that background questionnaire rotation does not, in and of itself, compromise achievement estimates (Adams, Lietz, & Berezner, 2013), a number of questions remain with respect to whether and what to do with MCAR background questionnaire data. To that end, as with standard multiple imputation techniques, relationships between variables not included in the imputation model ("conditioning model" in this context) will be attenuated (Rubin, 1987). As such, any design should ensure that all background items receive sufficient exposure to accurately estimate the relationship between achievement and those variables, even in the absence of missing background data treatment. That is, there should be enough information in an individual background questionnaire item to estimate stable subpopulation achievement across levels of the item, even if the MCAR data are left *as is*. This would most reasonably be done using current operational approaches for handling missing background data, which are to code missing responses as a category of the variable.

Core block and power

Depending on the selected questionnaire design, a block of core items, administered to all examinees, could be incorporated. In deciding *what* to incorporate into the core block as well as the size of the core block, a number of issues should be considered. In the case of the rotated questionnaire in PISA 2012, the items common to all three questionnaire forms included measures such as general student demographics (grade, sex), truancy behavior, a collection of items used to derive a measure of economic and social cultural status (mother's and father's education and occupation and home possessions), and measures of immigrant status and home language use (OECD, 2013). These variables can be regarded as *input* variables in the input-processes-outcomes model of education (Purves, 1987) used by PISA or as *key reporting variables*. Regardless, there is a clear priority on obtaining the best possible estimates of achievement across levels of these variables. And given an interest by policy makers and educational effectiveness researchers in these sorts of variables, it is sensible to argue that they should be adminis-

tered to all students. For more details on the PISA 2012 student background questionnaire design, see OECD (2013).

Another perspective on planned missingness regards the expected strength of relationships among background variables and achievement and the power to find these effects (Enders, 2010). In particular, variables that are well-correlated with achievement can reasonably be assigned to blocks outside of the core since "the impact of missing data on power will diminish as the correlations among the variables increase in magnitude" (Enders, p. 27). In the quoted study, statistical power for $\rho = .30$ decreased from 1 to .90 for pairs of variables with 66% missing data. In contrast, when $\rho = .10$, power to detect a correlation diminished from .41 to .18 with 66% missing data. As a result, variables that are substantively important (to policy or research) but have imprecise or weak relationships with achievement should be included in the core. This is to maximize statistical power, as the covariance coverage between the core variables and achievement will be 1. In contrast, variables outside of the core will exhibit covariance coverage below 1, the magnitude of which depends on the selected design. I define covariance coverage in the next section and discuss it more thoroughly in the example designs.

Pairwise correlations

In terms of planned missingness designs, one useful measure to consider is the covariance coverage matrix, which gives the percentage of complete data for each pair of variables. This matrix is informative particularly with respect to the amount of data available for calculating pairwise correlations. Depending on the selected design, some variables will have high covariance coverage while others will have low or even zero covariance coverage. To the degree possible, decisions about questionnaire design should take this into account, especially the ways in which important policy and research variables are expected to relate to achievement and also among background items. The latter is important in considering methods for summarizing information on non-achievement constructs (e.g., affective and behavioral constructs) via measurement models, but also for relating background variables to one another. Covariance coverage among variables in the background questionnaire should also be considered when developing plans for treating missing background data. That is, well-correlated variables that are to be imputed should exhibit some degree of pairwise overlap across blocks to ensure that a sensible imputation model is possible.

Prioritizing background information relative to program research and policy questions

In selecting a planned missingness design, important considerations are the research and policy priorities of the study. For example, stated research aims of PISA 2012 were understanding patterns of educational effectiveness and a focus on educational equity, as two examples (OECD, 2013). Similarly, the 2011 Trends in International Mathematics and Science Study (TIMSS) framework listed investigating equity and the effect of

changing demographics as areas of interest (Mullis et al., 2009). As a result, the questionnaires are designed to sufficiently measure variables that can reasonably be used to achieve these aims. Although neither study explicitly prioritizes research aims as they pertain to background questionnaire development, this is an area that should receive considerable attention when deciding on a sensible planned missingness design. This is particularly relevant for variables or constructs that are politically sensitive or of special policy interest, where tolerance for missingness or reduced covariance coverage might be low for some important stakeholders. In addition to the acceptability of missing data, the importance of particular variables or constructs should be included in any design plan to ensure that high-priority measures will have sufficient data to estimate correlations with achievement and with other important variables.

A few design examples for planned missingness

In what follows, I illustrate several possible planned missingness designs and discuss these designs as they relate to the issues outlined above. The first design example considered here is the PISA 2012 design, illustrated in Figure 2. Of the designs considered here, it is the only one with related evidence regarding the impact of planned missingness in the background questionnaires on population and subpopulation achievement estimates (Adams et al., 2013). Specifically, the authors used PISA 2006 data to simulate a two-form design. They employed several different conditioning models to estimate population achievement. The study differed in several regards from the operational approaches used in 2012, including using a two-form rather than a three-form design. Further, they used background scales rather than individual variables. Nonetheless, the authors consistently found that this rotation design had very little effect on overall achievement estimates or on associations between the background variables and achievement, regardless of whether the variables were in the conditioning model. Although this leaves some open questions regarding the effect of operational procedures on achievement estimates, there is some evidence that achievement is stable under such an approach.

As with any design that features a core block, variables that are included in the core will have the highest potential statistical power, given that they are presented to all examinees. And as noted above, the variables included in the core generally correlate well with achievement. For example, increases in the books in the home variable (a common proxy for socioeconomic status) predict an approximately 24 point increase in PISA scores among OECD countries, with standard errors around .5 to .8 points. In turn, the effect is roughly equivalent to one-quarter of a standard deviation on the PISA scale. In some respects, this and similar variables could theoretically be moved to the rotated blocks, as the impact on statistical power would likely not change conclusions around the effect of these variables. Many of the core-block variables, however, are typically of high policy value and figure prominently in discussions around educational equity (e.g., Hanushek & Luque, 2003) and achievement gaps (D. Rutkowski, Rutkowski, & Plucker, 2012). As a result, systematic partially missing data on these variables would possibly be met with resistance, pointing to a need to either create scale scores for these variables or use multiple imputation to fill in the missing data.

Figure 2 contains the information for covariance coverage for this design. Here, we can see that two-thirds of data are available for any pairwise correlations that involve the core block (*X*) and any other block or with achievement. This coverage deteriorates when pairwise relationships are of interest between rotated blocks since only one-third of data are not missing. Assuming list-wise deletion, the same is true of any analysis (e.g., an ordinary least-squares regression) where variables from two rotated blocks are involved as, for example, predictors in a model of achievement. The problem is most severe in an analysis that involves variables from all rotated blocks, where no non-missing data would remain under list-wise deletion. With this in mind, analyses that do not treat missing data in a meaningful way will feature highly restricted sample sizes ($n = 0$, at the extreme). A further problem arises when considering the possibility of creating scale scores via measurement models. Although the forms are linked, in that block *A* appears in form 1 and 3, block *B* appears in form 1 and 2, and block *C* appears in form 2 and 3, it is not possible to generate values on latent variables for examinees that did not respond to a set of items. That is, any examinee that received form 1 would have no scores for latent variables that derive from block *C*, using standard factor analytic or item response theory approaches. One exception to this is the Bayesian approach that is used to estimate achievement (Mislevy et al., 1992), which can also be used to create scale scores. However, with variables that are not theoretically related and possibly poorly-correlated, the degree to which such an approach would work is an open question.

In terms of prioritizing variables relative to program goals and policy and research questions, the current design, in general, offers the possibility of putting the most critical items or groups of items into the core block. Further, given that each rotated block occur in two forms, there is at least the potential for correlating all variables with achievement and all variables with one-another, given the covariance coverage noted above. Nevertheless, if policy or research priorities are in conflict with the design for some constructs, a possible modification could attend to this while allowing for a modest increase in measured material. In particular, a reasonable variation would be an extended, larger core with smaller rotated blocks. The PISA 2012 questionnaire dedicated 10 minutes to the core and 10 minutes to each of the rotated blocks, for a total questionnaire time of 30

		Block			
	Ach	X	A	B	C
Ach	100%				
X	100%	100%			
A	66%	66%	100%		
B	66%	66%	33%	100%	
C	66%	66%	33%	33%	100%

Note. Ach = achievement.

Figure 2:
Covariance coverage for standard 3-form design.

minutes (and 41 minutes of material, offering an approximately 33% increase in content coverage). One example of such a modification could be a 20 minute core with three five-minute rotated blocks. This increases the total measurable material to 35 minutes (a modest 17% increase). One possible advantage (or consequence) of this approach and an open area of research is if and how this could impact posterior achievement distributions. As the number of fully observed variables in the conditioning model increases, it is reasonable to expect that sub-population differences will be more accurately estimated across those variables.

As a second example of a rotated design that adheres to the three-form approach is one that I refer to as *three-form-A*, found in Figure 3, to distinguish it from the standard three-form described above. In the three-form-A design, each rotated block should be divided into three sub-blocks. For example, block *A* should be divided into three approximately equal sub-blocks, A_1, A_2, and A_3. Ideally, some set of variables in each of the sub-blocks should be reasonably well-correlated with variables in the other sub-blocks. This provides some sensible means by which to impute missing data or create scale scores using traditional methods. The distribution of items across sub-blocks could reasonably be accomplished by splitting up items that comprise scales, also offering a theoretical basis for imputation. For example, the PISA 2012 *math self-efficacy* scale is comprised of six items that ask students about their confidence in doing math tasks. This scale could be evenly distributed across the three blocks and, ideally, be situated with other items within a given sub-block that correlated reasonably well (e.g., items from the scale *math self-concept*). At the scale level, these two variables have an estimated weighted correlation, averaged across countries, of .496. Although country-wise item correlations will likely differ from the scale averages, especially in light of measurement error, this gives some idea of the ways in which variables can be combined to optimize the possibility of sensibly imputing or creating scale scores.

In terms of achievement estimation, there is little reason to believe that there are advantages over the standard three-form design, as there is no more or no less information in the conditioning model. Further, without any changes to the core block, this does not confer advantages related to statistical power. This might, however, serve as a possible means to move some items out of the core block and into rotation, given that it is a more reasonable place from which to impute. Regarding pairwise correlations, the covariance

	Block									
Form	X	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3
1	1	1	1	0	1	1	0	1	1	0
2	1	0	1	1	0	1	1	0	1	1
3	1	1	0	1	1	0	1	1	0	1

Note. 1 = questions asked; 0 = questions not asked.

Figure 3:
Three-form-A design.

coverage matrix can be found in Figure 4. Given that the basic design is the same, there is little difference in terms of covariance coverage of variables across blocks. That is, the core and achievement exhibit complete coverage, all sub-blocks have 66% covariance coverage with achievement and the core. And across rotated blocks (e.g., *A* and *B*) there is 33% covariance coverage. In contrast, within rotated block covariance coverage decreases from 66% to 33%. Given this design feature, it is clearly important to consider the importance of pairs (or groups) of variables appearing in the same sub-block. In thinking through implications, it is important to consider the relevance of lower covariance coverage to statistical power, research priorities, or political issues.

One issue not discussed until now is the matter of context or position effects that can arise due to rotation (Lord & Novick, 1968; Sirotnik, 1970). Typically, it is assumed that item performance "does not depend on the context in which the item occurs" (Lord & Novick, p. 252). Nevertheless, it can and does happen that item characteristics are a function of where the item is located and what surrounds it (Frey & Bernhardt, 2012; OECD, 2002). To account for this issue, one solution is to use a balanced booklet design where every block occurs with every other block and in every position. Clearly, neither the three-form nor the three-form-A designs employ this sort of balancing. Although in the standard three-form, some control over context effects is realized because each block occurs with every other block, the block positions are fixed. Yet, large-scale assessment sample sizes are typically quite large – in the thousands for TIMSS and PISA (Martin & Mullis, 2012; OECD, 2014); much larger for NAEP (US DOE, 2011). And, in the case of TIMSS and PISA, computerized testing is replacing paper and pencil versions as the standard, making it a relatively straightforward matter to present a given form in random order for each examinee. Such an approach would offer better control over context and booklet effects while still minimizing response burden for examinees.

		Block									
	Ach	X	A1	A2	A3	B1	B2	B3	C1	C2	C3
Ach	100%										
X	100%	100%									
A1	66%	66%	100%								
A2	66%	66%	33%	100%							
A3	66%	66%	33%	33%	100%						
B1	66%	66%	33%	33%	33%	100%					
B2	66%	66%	33%	33%	33%	33%	100%				
B3	66%	66%	33%	33%	33%	33%	33%	100%			
C1	66%	66%	33%	33%	33%	33%	33%	33%	100%		
C2	66%	66%	33%	33%	33%	33%	33%	33%	33%	100%	
C3	66%	66%	33%	33%	33%	33%	33%	33%	33%	33%	100%

Figure 4:
Covariance coverage for three-form-A design.

	Block									
Form	X	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3
1	1	1	1	1	1	0	0	0	1	1
2	1	1	1	1	0	1	0	1	0	1
3	1	1	1	1	0	0	1	1	1	0

Note. 1 = questions asked; 0 = questions not asked.

Figure 5:
Three-form-B design

A third design that I include here is a further modification of the three-form design that extends the core to include some key constructs while reducing information on something of lower priority. This design is represented in Figure 5 and I refer to it as *three-form-B*. Under this design, the core is expanded to include all of block *A* (although this choice is arbitrary for the current example). To account for the expansion of the core, the exposure of some sub-blocks within a block is reduced (*B*, in this example) while the third block is left as in the three-form-A design. This design confers some advantages, in that the core is expanded, total response burden is controlled, and the total measurable material remains the same. Nevertheless, this design also has some consequences for statistical power and pairwise comparisons. In particular, power is reduced for analyses involving any *B* sub-blocks as they only appear once. And because each sub-block in *B* appears only once, no pairwise correlations are possible across sub-blocks within *B*. As with three-form-A, this design is not expected to offer any advantages in terms of preserving posterior achievement distributions over the standard three-form design. An additional consequence of this design is the inability to control context effects, even under randomization (as proposed above). This is due to the fact that not all sub-blocks appear with all other sub-blocks. For example, B_1 never appears with C_1.

In terms of creating the conditions for sensible imputation, this design obviates the need to impute core blocks, offering an advantage over the three-form-A design. Further, as long as items in *C* are divided across sub-blocks to optimize the cross-block correlations, there is no disadvantage. In contrast, imputation for sub-blocks in *B* would have to rely on reasonably strong correlations outside of the block (e.g., within the core or block *C*). Although this is not an insurmountable problem and could be a good option in situations where items in *A* take high-priority at the acceptable reduction in coverage of *B*, the distribution of items across sub-blocks within *B* would have to done with considerable care.

A final design that I consider here is the so-called *two-method design* that assigns the entire questionnaire battery to a small group. This is a method described and demonstrated in several places (e.g., Graham et al., 1996; Graham, Taylor, Olchowski, & Cumsille, 2006; Graham & Shevock, 2012) and uses a high-quality but expensive measure, which is administered to a small sub-sample of study participants. Several examples in these citations feature self-reports supplemented by more expensive measures (e.g, biochemical measures, extensive interviews) designed to correct for response bias that can arise

from social desirability or other error sources. Ideally, the more expensive measures are more valid and have better reliability than the cheaper measures. I propose a modification of this design for the international assessment context. In particular, the modified two-method (M2M) design would combine a three-form design along with a small subsample that responds to the entire questionnaire. Here, we can think of the additional time needed to complete the questionnaire as the "expense," with the benefit that there are complete responses for some sub-set of examinees. This design is represented in Figure 6. Notice that there is one additional form (form 4) that includes all blocks. Assuming that each block takes approximately the same amount of time to complete, this form will take 33% longer to complete than any of the other three forms. Further, the representation here might suggest that booklet 4 is administered 25% of the time; however, this is an open area of research in this particular context. Given the findings from other related research (Graham et al., 2006), it is likely that booklet 4 could be administered to a much smaller sub-sample, for example rotated at a rate of one in seven booklets. Although I provide one example of the M2M design here, it is also possible that the complete booklet could be combined with one of the other modified 3-form designs.

In terms of modern missing data methods, this design confers some advantage, especially in the case of the three-form-B design, since all correlations are estimable from the data. Further, under any design, there is some basis for including three-way interactions, where relevant, in the imputation model. In terms of advantages or disadvantages for posterior achievement distributions, this is an open research question; however, if well-executed, there is no reason to believe that additional information would be harmful. For all rotated blocks statistical power would increase, the degree to which depends on the frequency of inclusion of booklet 4. This design offers no additional information for pairwise correlations; however, the covariance coverage of all pairwise correlations increases for blocks *A*, *B*, and *C*. Although this increase depends, again, on the frequency of booklet 4 rotation, we offer one example where it is rotated evenly or 25% of the time. In the case of covariance coverage of the core or achievement with the rotated blocks, this would increase from 66% to 75%. Likewise, covariance coverage across rotated blocks would increase from 33% to 50%. Less frequent rotation would result in lower increases in coverage.

	Block			
Booklet	X	A	B	C
1	1	1	1	0
2	1	0	1	1
3	1	1	0	1
4	1	1	1	1

Note. 1 = questions asked; 0 = questions not asked.

Figure 6:
Modified two-method design.

An issue of primary concern under this design is the sampling method used in international assessments. In particular, TIMSS samples whole classrooms while PISA samples a group of 15-year-olds that are administered the test together. As a result, implementing this background questionnaire design would require that administration of the longer forms is done for the entire classroom or sampled group within a school. In this way, the administration burden could be confined to test administrators in selected schools. If no such approach is taken the longer booklets would be administered in the usual fashion, meaning that the additional administrative burden and teaching disruption will extend to every sampled school, as some percentage of students will receive the longer booklet. It is reasonable to expect that such a proposal would be met with resistance from school leadership and teachers. As such, examinee sampling and background questionnaire design would have to be considered in parallel.

Discussion and conclusion

A key priority in national and international assessments, such as NAEP, TIMSS, and PISA is unbiased and sufficiently precise population and sub-population achievement estimates while also covering broad content in a limited testing time. To that end, these assessment programs have long employed planned missingness in the achievement test, referred to as multiple matrix sampling (Shoemaker, 1973). And specialized methods are used to ensure that achievement distributions are stable and well-estimated for populations of test takers (Mislevy et al., 1992). As interest in the context and correlates of achievement have grown, so too has an interest in measuring more non-achievement domains. This creates a classic tension between a desire to measure many domains broadly while also limiting the response burden for study participants.

Given an interest in expanding the measured background domains in large-scale assessments, I take up this discussion and describe several questionnaire designs that can potentially be employed in the large-scale assessment context when planned missingness is included in the background questionnaires. I recognize an interest in carefully preserving achievement distributions and I identify open questions regarding if and how to treat missing background questionnaire data. However, I do not tackle these issues directly. Rather, I take a step back and consider several designs that can reasonably be employed in the large-scale assessment setting. Here, I emphasized variations on the three-form design, which typically features a core set of questions administered to all study participants and several blocks of questions, some subset of which are assembled into booklets. Under this approach and in-line with the methods for achievement booklet design, individual examinees respond to a subset of the total amount of measurable material.

The designs covered here may not fit every situation and there are many other designs that could have been considered; however, those presented offer are a reasonable starting point and they highlight several important considerations and areas that remain open for further research. The basic three-form design discussed here has also received considerable methodological attention in psychological science settings (Graham & Shevock, 2012; Graham et al., 2006), particularly in developmental research (Little & Rhemtulla,

2013; Mistler & Enders, 2012). Although it is a relatively simple design that is straightforward to implement, it offers several advantages including increased content coverage, reduced response burden, and flexibility in terms of design and administration. And depending on the goals of the study under consideration, variants on the basic design can be used to prioritize particular policy or research questions.

In reviewing each design, I considered whether there were apparent advantages or disadvantages for (a) estimating population achievement distributions; (b) statistical power; (c) pairwise correlations and covariance coverage. I also briefly discussed issues around order or context effects. I considered four designs, including the standard three-form design and three variants. As one empirical example, I discuss the case of PISA 2012, which implemented a standard three-form design in the student background questionnaire. In a limited study, this general approach was found to have no meaningful effect on population or sub-population achievement distributions (Adams et al., 2013). However, there is a need for much more research on the impact of questionnaire rotation in achievement distributions, particularly given the importance of achievement estimates to policy and research.

The current manuscript begins to examine some possibilities with respect to background or context questionnaire rotation along with some of the issues that should be considered in the selection of a design. One issue that clearly stands out is what (if anything) should be done with missing data that arise out of questionnaire rotation. In the 2012 PISA cycle, the data were left as is, with any choices about missing data treatment left to the end analyst. Although this is certainly a reasonable approach, particularly given the state and availability of modern missing data methods, it seems that a second, viable approach would be for the testing organizations to impute or produce scale scores for a limited set of mutually agreed upon variables. Such an approach is in line with what is currently done operationally with achievement scales; however, a sensible imputation model with well-described limitations would need to accompany such a solution. Importantly, however, further research is needed to understand what impact this approach might have on achievement scales overall and on relationships with achievement and how best to implement this sort of method. Nevertheless, providing imputed data, even on a limited subset of high-priority scales, would limit the operational burden implied by treating missing data at the organization level, while also providing complete data for some scales. Finally, if the operational decision is to leave “holes” in the public-use data, an open research question remains: what is an optimal planned missingness design strategy that would allow end-users to sensibly treat missing non-achievement data and to best model achievement and non-achievement data?

References

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education*, *1*(1), 5. https://doi.org/10.1186/2196-0739-1-5

Enders, C. K. (2010). An introduction to missing data. In *Applied missing data analysis* (pp. 1–36). New York: Guilford Press.

Frey, A., & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling*, *54*(4), 397–417.

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., & Shevock, A. E. (2012). Planned missing data design 2: Two-method measurement. In *Missing data: Analysis and design* (pp. 295–323). New York: Springer Science & Business Media.

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

Hanushek, E. A., & Luque, J. A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review*, *22*(5), 481–502. https://doi.org/10.1016/S0272-7757(03)00038-4

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*(4), 199–204.

Lord, & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Welsley Publishing.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381. https://doi.org/10.1007/BF02306026

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

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

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Boston: TIMSS & PIRLS, International Study Center, Lynch School of Education, Boston College. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED512411

OECD. (2002). *PISA 2000 Technical Report*. Paris: OECD Publishing.

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving, and financial literacy*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

OECD. (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

OECD. (2016). *PISA 2015 assessment and analytical framework*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264255425-en

Purves, A. C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review*, *31*(1), 10–28. https://doi.org/10.2307/1188220

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*. Hoboken, NJ: Wiley.

Rutkowski, D., Rutkowski, L., & Plucker, J. A. (2012). Trends in education excellence gaps: a 12-year international perspective via the multilevel model for change. *High Ability Studies*, *23*(2), 143–166. https://doi.org/10.1080/13598139.2012.735414

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

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling* (Vol. xviii). Oxford, UK: Ballinger.

Sirotnik, K. (1970). An investigation of the context effect in matrix sampling. *Journal of Educational Measurement*, *7*(3), 199–207. https://doi.org/10.1111/j.1745-3984.1970.tb00717.x

US DOE. (2011). NAEP - Mathematics 2011: Target Population and Sample Size. Retrieved March 11, 2016, from http://www.nationsreportcard.gov/math_2011/target_pop.aspx

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Mongraph Series*, *2*, 9–36.

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Boca Raton, FL: CRC Press.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C.R. Rao and S. Sinharay (Ed.), *Handbook of Statistics* (Vol. Volume 26, pp. 1039–1055). Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169716106260322