1-1-2014

Two-Method Planned Missing Designs for Longitudinal Research

Mauricio Garnier-Villarreal
Marquette University, mauricio.garniervillarreal@marquette.edu

Mijke Rhemtulla
University of Amsterdam

Todd D. Little
Texas Tech University

Accepted version. *International Journal of Behavioral Development,* Vol. 38, No. 5 (2014): 411-422. DOI.

International Journal of Behavioral Development, Vol. 38, No. 5 (September 1, 2014): 411-422. DOI.

Two-method Planned Missing Designs for Longitudinal Research

Abstract

We examine longitudinal extensions of the two-method measurement design, which uses planned missingness to optimize cost-efficiency and validity of hard-to-measure constructs. These designs use a combination of two measures: a "gold standard" that is highly valid but expensive to administer, and an inexpensive (e.g., survey-based) measure that contains systematic measurement bias (e.g., response bias). Using simulated data on four measurement occasions, we compared the cost-efficiency and validity of longitudinal designs where the gold standard is measured at one or more measurement occasions. We manipulated the nature of the response bias over time (constant, increasing, fluctuating), the factorial structure of the response bias over time, and the constraints placed on the latent variable model. Our results showed that parameter bias is lowest when the gold standard is measured on at least two occasions. When a multifactorial structure was used to model response bias over time,

it is necessary to have the "gold standard" measures included at every time point, in which case most of the parameters showed low bias. Almost all parameters in all conditions displayed high relative efficiency, suggesting that the 2-method design is an effective way to reduce costs and improve both power and accuracy in longitudinal research.

Keywords intentionally missing data, missing data, planned missingness, structural equation modeling, simsem, two-method measurement

Planned missing data designs allow more data of higher quality to be collected than is typically possible. In these designs, either variables or occasions of measurement (or both) are omitted for a subset of participants, resulting in a predetermined pattern of missing data. These designs are most frequently employed when time constraints or attention concerns prevent a researcher from administering all measures (e.g., survey items) to each participant. When participants are randomly assigned to a particular pattern of missingness, the planned missing data meet the missing completely at random (MCAR) assumption and no bias results from missing data. Planned missing designs are used to increase validity (e.g., to reduce fatigue, response reactivity, and unplanned missingness) and decrease costs; however, they are less efficient than complete-case designs, which must be factored into the power estimates for these designs (Graham, 2012).

In contrast to other planned missing designs, the two-method design uses planned missing data to boost the power of a small study by dramatically increasing the total sample size. Beginning with a study in which a small sample is subjected to an expensive measurement scheme (e.g., direct observation), an additional inexpensive (but less valid) measure is administered to a large sample of participants. The result is a large sample of participants, most of whom have "missing data" on the original, expensive measure. With this design, expensive small-sample research is moved into the realm of affordable large-sample latent-variable research. Alternatively, the two-method measurement design can be used to increase the validity of large-sample research by including an additional small-sample measure that provides a gold-standard reference group.

The two-method design is predicated on the existence of two types of measures of a construct of interest. The first type must be considered a "gold standard" measure, which is characterized by very high construct validity and accordantly high assessment costs. Examples of gold-standard measures include measuring carbon monoxide in the blood to assess smoking behavior (Graham, Taylor, Olchowski, & Cumsille, 2006, Olchowski, 2007), individually administering intelligence tests such as the WISC (Wechsler, 2004) to assess cognitive ability, direct classroom observations of attention such as the Classroom Observations of Conduct and Attention Deficit Disorders (Atkins, Pelham, & Licht, 1985), or repeated cortisol measures to assess stress

(Gatti et al., 2009; Levine, Zagoory-Sharon, Feldman, Lewis, & Weller, 2007). The cost to administer such measures can be prohibitive, limiting the number of participants or occasions of measurement that can be feasibly tested. As a result, researchers may conduct studies that are under-powered, particularly for the demands of statistical modeling procedures such as structural equation modeling or multilevel modeling (MacCallum, Browne, & Sugawara, 1996).

The second type of measure is a less expensive but less valid proxy of the construct, which is easy to administer to a large number of participants. Examples of inexpensive measures include items on a self-report survey of smoking behavior, a paper-and-pencil test of cognitive ability, a teacher report of attentiveness, or a self-report measure of stress. In each case, these measures contain a source of method-related variance in addition to the construct they aim to measure: in the case of surveys, for example, self-report or teacher-report contain systematic variance that is unrelated to the measured construct (Ready & Wright, 2011). Having a source of systematic method-related variance means that the inexpensive measure is a biased representation of the construct, and, if used alone, it results in biased parameter estimates (Graham et al., 2006). Table 1 contains examples of constructs with corresponding measures that would be good candidates to use in the context of the two-method framework.

Table 1. Constructs and measures that fit the two-method planned missing framework.

Construct	Unbiased measure	Biased measure
Smoking	Cotinine, carbon monoxide (Taylor, Graham, Palmer, & Tatterson, 1998)	Self-reported smoking behavior (Brener et al., 2002)
Executive function	Performance tasks (Alloway, 2007; Armengol, 2002; Brophy, Taylor, & Hughes, 2002)	Behavior rating inventory of executive function (Gioia, Isquitch, Guy, & Kenworthy, 2000)
Depression	Electroencephalogram (EEG) (Yao et al., 2010), fMRI (Eugene, Joorman, Cooney, Atlas, & Gottlib, 2010)	Hamilton depression rating scale (Hamilton, 1967), Beck depression inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961)
Alzheimer's disease	Structural MRI (Schmand, Eikelenboom, & Van Gool, 2011), biomarkers (Genius, Klafki, Benninghoff, Esselman, & Wiltfang, 2012)	Stroop task switching (Hutchinson, Balota, & Ducheck, 2010)
Stress	Cortisol (Gatti et al., 2009; Levine et al., 2007)	Depression anxiety stress scales (Szabo, 2010)
Temperament	Observations (Seifer, Sameroff, Barret, & Krafchuk, 1994)	Children's behavior questionnaire (Rothbart, Ahadi, Hershey, & Fisher, 2001)

Faced with both types of measures, most researchers decide to use either the gold standard, recognizing that power may be attenuated, or use the inexpensive measure, recognizing that the construct will be contaminated with a source of systematic bias. In the two-method measurement design, both types of measure are administered. The variance that is shared across both measures reflects the construct of interest; the additional variance that is shared only among the indicators of the inexpensive measure reflects response bias associated with that measure.

The two-method measurement design uses latent variable structural equation modeling to separate the method variance (response bias) from the construct variance in the inexpensive measure. As shown in Figure 1, this design uses the

information from both the gold-standard measure and the inexpensive measure to estimate the focal construct. The model contains two types of observed variables: a set of indicators belonging to the gold standard (e.g., several measures of carbon monoxide blood levels or several time periods' worth of observations), and a set of indicators belonging to the inexpensive measure (e.g., several self-report indicators). Both types of measure load on the construct (e.g., smoking behavior). These loadings capture the proportion of variance in each indicator that is associated with the construct. A second construct, response bias, is also included as part of the measurement model; only the biased measure's indicators load on this construct. These loadings reflect the amount of shared variance in the inexpensive measure that is independent of the construct, that is, the information that reflects the response bias in the inexpensive measure. The gold-standard measure does not have an additional response bias correction construct because it is assumed to be measured without bias, hence, its status as a gold standard. The gold-standard indicators become the anchors that define the core construct. If a measure existed that was both inexpensive and unbiased, this design is not needed because the measure can be administered to a large sample with high construct validity.

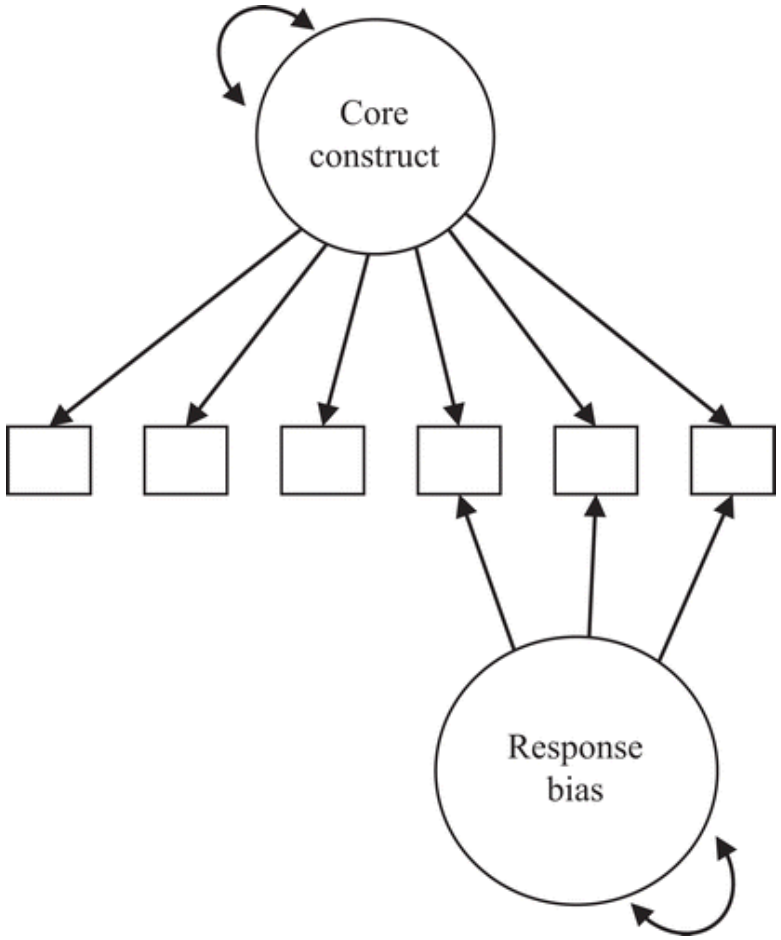

Figure 1. Two-method design structure.

The planned missing aspect of the two-method design involves how the measures are implemented. Only a limited number of participants can be administered the gold-standard measure, because it is expensive. A much larger number of participants is administered the inexpensive measure. By randomly assigning a small proportion (e.g., one-third) of participants to receive the expensive measure and administering the inexpensive measure to the entire (large) sample, the costs of conducting a large-sample study are kept at an affordable level.

Graham et al. (2006) conducted a series of simulation studies to find the ideal proportion of a sample who should receive the gold-standard measure in addition to the inexpensive measure, given a fixed total cost and desired level of efficiency. For example, given two measures of smoking behavior, an efficient design would be one where the effect of smoking (a latent variable measured by inexpensive and gold-standard measures) on health (a dependent variable) can be estimated with a high degree of precision (i.e., a small confidence interval). As more participants receive the expensive measure, the total sample size decreases rapidly (given a fixed total cost), so efficiency decreases as this proportion rises. At the same time, if too few participants receive the expensive measure, the response bias in the inexpensive measure becomes harder to estimate, decreasing efficiency. As such, maximum efficiency is achieved somewhere in between. Graham et al.'s (2006) simulations revealed that this ideal proportion depends on the degree of reliability of the two types of measure (as the inexpensive measure becomes more reliable, fewer gold-standard participants are needed), the amount of response bias in the inexpensive measure (as the inexpensive measure becomes more biased, more gold-standard participants are needed), the cost differential between the two types of measure (as the gold standard becomes more expensive in relation to the inexpensive measure, the most efficient design includes a smaller proportion of gold-standard participants), and the effect size relating the construct of interest to other constructs (as these structural relations become stronger, more gold-standard participants are needed). Their paper discusses how to estimate this proportion using information that is typically available prior to data collection.

To our knowledge, no one has applied the two-method measurement design to examine the behavior of a construct over time. Rhemtulla and Little (2012) proposed that the two-method design could be extended to longitudinal scenarios by including the gold-standard measure at a subset of measurement occasions (e.g., the first and last time point). The current project aims to study this idea via simulation. In two studies, we examine the performance of parameter estimates in a longitudinal structural equation model when the gold standard is measured at one, two, or all occasions in a four-wave study. To enable comparison, we hold the total number of gold-standard data points collected constant at 200 across all time points, allowing designs where either 200 expensive measurements are collected at a single time point, 100 expensive measurements are collected at each of two time points, or 50 expensive measurements are collected at all four time points. These conditions are expected to

produce different results depending on whether the degree of response bias actually changes over time (e.g., if self-report bias increases over the course of a smoking-cessation study). We explore three models of response bias: In Study 1, response bias is modeled using a single latent factor for all measurement occasions, with loadings on this factor either (a) fixed across occasions ("bias invariance constraint") or (b) allowed to differ across occasions, and in Study 2, response bias is modeled using (c) a distinct latent response bias factor for each measurement occasion. We examine the results in terms of bias and efficiency of parameter estimates.

Study 1: Method

We used the simsem (0.5–3) package (Pornprasertmanit, Miller, & Schoemann, 2013) within R (2.15.2, R Core Team, 2012) to generate data, impose missing values, and estimate a range of analytic models.

Population models

The data generation model was a panel model with four measurement occasions (see Figure 2). At each occasion there are three gold-standard indicators of the core construct of interest (e.g., these might be three outcomes from a blood test, three components of an intelligence test score, or three hours' worth of classroom attention observations) and three inexpensive/biased indicators (e.g., three items or parcels of items from a self-report measure). Lag-1 residual correlations (i.e., the correlation between the residual of the first indicator of the inexpensive measure at time 1 with that of the first indicator of the inexpensive measure at time 2) were .2. The regression coefficients over time for the core construct were .3 at lag-1 and .15 at lag-2. The unstandardized factor loadings of all 6 indicators on the core construct were .7. The total variance of each gold-standard indicator was 1. The total variance of the inexpensive indicators varied as a function of the amount of response bias. Rather than holding the total indicator variance constant, we held constant the ratio of reliable to residual variance at 49:51, where the reliable variance is composed of construct variance and response bias variance. The amount of reliable variance is a sum of the squared core-construct loading (always .7) and the squared response bias loading. For this reason, the parameter values mentioned previously are not standardized, the standardized factor loading of the inexpensive indicators are lower than the ones of the gold-standard indicators. The standardized regression coefficients are .287 from time 1 to time 2, .140 from time 1 to time 3, .292 from time 2 to time 3, .145 from time 2 to time 4, and .298 from time 3 to time 4. The standardized factor loadings of the gold-standard indicators are .7 at time 1, .715 at time 2, .724 at time 3, and .726 at time 4.

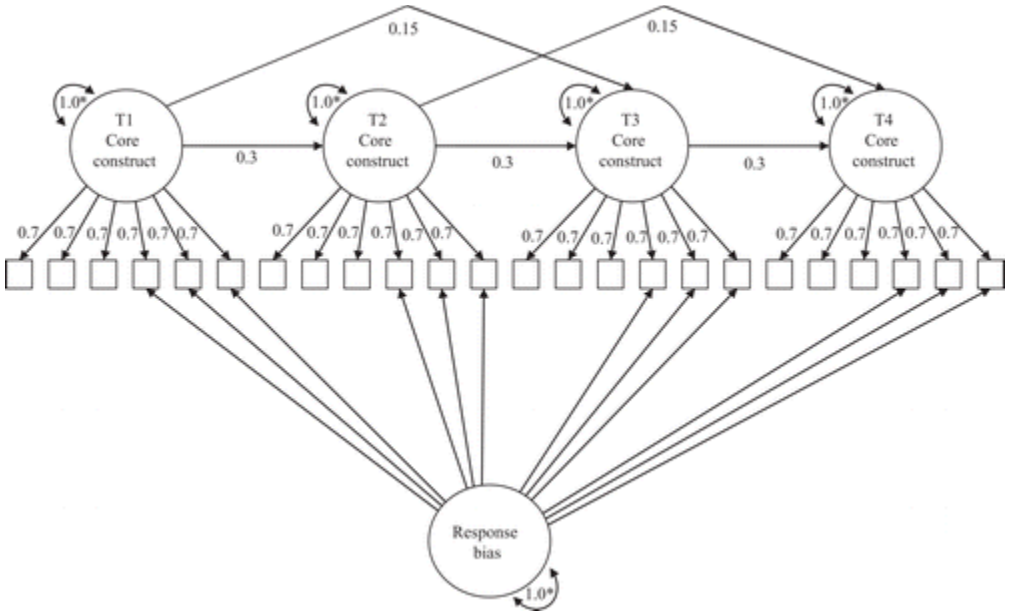

Figure 2. Data generation model for Study 1. Residual variances and residual covariances (at lag 1 and 2) are not shown.

Three population models were used; these differed in the amount of response bias in the inexpensive measure over time. In the *uniform bias* condition, this response bias was uniform over time, such that the inexpensive items had a .4 (unstandardized) loading on the response bias factor at every time point. In standardized terms, the inexpensive indicators had standardized loadings of .608, .624, .634, and .638 at times 1–4 on the core construct and .347, .342, .338, and .337 at each time point on the response-bias factor. In the *random bias* condition, the response bias loadings at each measurement occasion were drawn from a uniform distribution with a range of .3 to .6 (all three indicators within a time point always had the same loading). The standardized core-factor loadings are .640, .622, .601, and .589 at times 1–4. Finally, in the *increasing bias* condition, the response bias factor loadings were .3, .4, .5, and .6 at times 1–4, corresponding to standardized response bias factor loadings of .276, .342, .397, and .446, and core factor loadings of .643, .624, .596, and .560.

Data generation and missing data

For each of the three population model conditions (i.e., uniform bias, random bias, increasing bias), complete data on both measures were generated with a total sample size of $N = 500$. Each cell had 500 replications. For each complete dataset, missingness was imposed according to nine different missing data designs. The number of gold-standard measurements collected over time was invariant at 200 in every condition, but these 200 measurements were distributed differently across occasions, in nine different ways. As shown in Table 2, either 50 participants were measured at each of the four occasions, 200 participants were measured at a single occasion, or

100 participants were measured at each of two occasions (occasions 1 and 2 or occasions 1 and 4). For the purpose of computing relative efficiency, we ran partial complete-data conditions where the gold standard was measured on *all participants* at just one occasion, at occasions 1 and 2, and at occasions 1 and 4. These six partial complete data designs were used for relative efficiency computations only.

Table 2. Missing data designs.

Design	Occasions when gold-standard measure was collected	Total number of participants receiving gold-standard measure at each time point	Percentage of overlapping gold-standard measurements
GM all	1, 2, 3, 4	50	100
GM 1	1	200	100
GM 2	2	200	100
GM 3	3	200	100
GM 4	4	200	100
GM 1–2	1, 2	100	100
GM 1–4	1, 4	100	100
GM 1–4 HO	1, 4	100	50
GM 1–4 NO	1, 4	100	0

Note. GM = gold-standard measure. 1–4 indicates the occasions at which the gold-standard measure was administered. HO = half of the gold-standard subsample overlaps across occasions. NO = none of the gold-standard subsample overlaps across occasions.

We examined two designs where the 100 participants receiving the gold standard on the final occasion was different from the group of 100 who received it on the first occasion—either 50 of the participants who received the gold-standard measure at the first occasion received it again (in which case a new 50 participants also received it at the last occasion, so there was 50% overlap) or none of the participants who received the measure at the first occasion received it again (i.e., a new 100 participants received it at the last occasion, so there was 0% overlap). We included these variants because, for various practical reasons (e.g., attrition), it may be difficult to get the same subset of participants to receive the gold-standard measurement twice. Thus, we tested the robustness of these designs when the gold-standard group is not overlapping. We did not simulate attrition in this study, so the original participants who received the gold-standard measure were still included at the last time point, only they did not receive the gold-standard measure again. We deleted 5% of the data using a missing completely at random (MCAR) mechanism to approximate the effect of a small amount of unsystematic unplanned missingness.

Analysis models

The model shown in Figure 2 was fit to each complete and incomplete dataset; models were fit using the simsem package in R, which relies on lavaan (0.5–13; Rosseel, 2012) for structural equation model estimation. Full information maximum likelihood was used to estimate model parameters both with complete data and with missing data (see Baraldi & Enders, 2010). Factorial invariance was imposed across measurement occasions by constraining factor loadings for all indicators to be equal across measurement occasions, while allowing the factor variance and mean to

change over time. Each data set was also analyzed a second time with an additional set of "bias invariance" (BI) constraints that forced estimated loadings on the response bias factor to also be invariant across occasions. Note that this constraint imposes a misspecification of the model when the amount of response bias in the inexpensive measure varies, that is, in the *random bias* and *increasing bias* population conditions.

Outcomes

We first examine rates of nonconvergence and improper solutions, where nonconverged cases are those where the maximum likelihood algorithm failed to derive model estimates, and improper solutions are those that resulted in at least one negative residual variance estimate. These cases were removed before computing bias and efficiency. We used absolute relative bias (ARB) to assess the accuracy of the point estimates of the loadings (on the construct and on the response bias factor) and of the structural regression paths. As shown below, ARB was calculated as the absolute value of the ratio of estimation bias to the true parameter value, $ARB_\theta = \left|\frac{\bar{\hat{\theta}}-\theta}{\theta}\right| \times 100$, where $\bar{\hat{\theta}}$ is the population parameter value, and $\bar{\hat{\theta}}$ is the average estimated parameter value across all converged replications. ARB was computed for each model parameter, and then averaged across construct loadings, response bias loadings, and structural regression paths to produce three values of ARB for each model (one for each generating population model). The reason for using absolute values is that for many models both positive and negative bias appeared across the range of parameters (e.g., positive bias on the construct loadings at occasion 1 and negative at occasion 4) so averaging across these would obscure the degree of bias. The direction of bias is described in the text.

In addition to ARB, we computed relative efficiency (RE) for each parameter by comparing the empirical standard error of each parameter when estimated with and without missing data. The empirical standard error is the standard deviation of parameter estimates across replications (excluding those that did not converge or produced improper solutions). RE is the ratio of squared parameter empirical standard errors (i.e., the standard deviation of the estimates across all replications) of complete data to incomplete (planned missing) data, $RE_{\hat{\theta}} = \frac{ESE^2_{\hat{\theta},complete}}{ESE^2_{\hat{\theta},missing}} \times 100$, where $ESE^2_{\hat{\theta}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \bar{\hat{\theta}})^2$ is the estimated parameter value for replication *i*, and *N* is the number of converged proper solutions in a condition. RE ranges from 0 to 1, where a value of 1 means that the parameter estimates are as efficient as they would have been with no missing data, and lower values reflect efficiency loss and lower power in significance testing.

The complete data comparison condition differed depending on when the gold-standard measure was included. For example, the condition where 200 participants

received the gold standard at occasion 3 and the factorial invariance constraint was applied was compared to a design with complete data on the gold standard at occasion 3 and the factorial invariance constraint applied.

Simulation 1: Results

Rates of convergence and improper solutions

When the bias invariance (BI) constraint was imposed (in addition to factorial invariance constraints, which were always imposed), convergence was always 100% and there were no improper solutions. For conditions where the BI constraint was not imposed, the percent of replications that failed to converge or converged with an improper solution is listed in Table 3. In addition, there were 3 replications total that resulted in a negative standard error estimate; these were also removed but are not reported in Table 3. When the BI constraint was not imposed, convergence rates ranged from 76.2% to 98.4%. Convergence was highest when the gold standard was measured at all four measurement occasions and lowest when it was measured at occasions 1 and 2. It was also lower when the gold standard was measured at occasions 1 and 4 with less than complete overlap. In every condition, rates of improper solutions were higher than rates of non-convergence.

Table 3. Percent of no convergence and improper solutions.

	Equal bias		Random bias		Increasing bias	
Design	No conv	Neg var	No conv	Neg var	No conv	Neg var
GM all	0.0	5.2	0.2	3.0	0.2	2.2
GM 1	1.6	6.2	6.2	13.4	3.2	18
GM 2	2.6	6.6	2.8	5.6	4.8	7.6
GM 3	3.0	6.6	0.6	2.6	0.6	2.4
GM 4	10.0	10.2	2.4	3.2	0.8	0.8
GM 1–2	1.8	8.4	2.6	13.2	2.6	14.4
GM 1–4	0.8	1.8	1.6	3.2	0.8	1.6
GM 1–4 HO	10.0	10.8	7.0	5.4	3.2	3.6
GM 1–4 NO	13.0	10.8	6.6	8.8	2.8	4.2

Note. No conv = no convergence; Neg var = at least one negative variance estimate, GM = gold-standard measure, 1–4 indicates the occasions at which the gold-standard measure was administered (see Table 2), HO = half of the gold-standard subsample overlaps across occasions, NO = none of the gold-standard subsample overlaps across occasions. No condition where the Equality of Bias Constraint was applied resulted in any non-convergences or improper estimates.

Absolute relative bias of parameter estimates

Table 4 displays the absolute relative bias (ARB) for the response bias loadings, construct loadings, and structural regression paths by each generating population (i.e., response bias is stable, randomly varying, or increasing over time) and by each of the nine planned missing data designs modeled with and without the BI constraint, for a total of 18 designs. As recommended by Muthén, Kaplan, and Hollis (1987), ARB less than 10% is considered negligible. Values higher than 10% are bolded in Table 4.

Table 4. Percent average absolute relative bias.

		Equal bias			Random bias			Increasing Bias		
Design	*df*	Bias loading	Construct loading	Structural paths	Bias loadings	Construct loadings	Structural paths	Bias loadings	Construct loadings	Structural paths
GM all	220	5.9	6.3	1.2	4.0	5.1	3.7	2.3	5.2	4.3
GM 1	64	2.7	4.4	9.6	4.4	4.0	**11.6**	3.8	4.1	6.3
GM 2	64	3.5	4.7	4.4	1.9	4.7	5.8	2.8	5.4	3.7
GM 3	64	2.8	4.9	9.3	1.9	5.6	5.1	1.7	5.7	4.0
GM 4	64	2.5	4.8	**11.3**	1.5	5.2	8.9	1.2	5.1	9.4
GM 1–2	109	3.5	5.1	3.1	3.1	4.6	4.5	2.7	4.6	3.5
GM 1–4	112	2.3	5.3	7.0	3.2	5.1	5.9	3.1	5.3	5.5
GM 1–4 HO	112	3.5	6.5	6.6	2.6	5.3	5.7	1.8	5.3	4.5
GM 1-4 NO	112	3.6	6.4	6.3	2.2	5.5	5.2	2.1	5.3	5.7
GM all BI	229	2.9	5.4	2.7	**18.0**	2.6	**18.1**	**22.3**	3.5	**26.2**
GM 1 BI	73	0.8	4.1	3.9	**21.2**	2.9	**20.5**	**24.5**	2.8	**28.4**
GM 2 BI	73	2.0	5.1	2.2	**17.9**	2.3	**14.4**	**22.0**	2.3	**20.9**
GM 3 BI	73	1.8	5.5	2.6	**18.3**	3.2	**14.9**	**23.6**	4.5	**21.7**
GM 4 BI	73	1.4	5.7	3.2	**19.0**	5.4	**21.5**	**25.3**	9.1	**33.3**
GM 1–2 BI	118	1.8	4.5	2.9	**19.9**	2.9	**17.0**	**23.0**	2.6	**24.2**
GM 1–4 BI	121	1.4	4.9	3.6	**18.5**	3.1	**10.6**	**21.8**	3.0	**11.5**
GM 1–4 BI HO	121	1.7	5.0	3.2	**18.0**	3.1	**17.2**	**22.4**	3.1	**24.9**
GM 1–4 BI NO	121	1.6	5.0	3.4	**17.8**	3.1	**17.3**	**22.6**	3.1	**25.0**

Note. GM = gold-standard measure, 1–4 indicates the occasions at which the gold-standard measure was administered (see Table 2). BI = bias invariance constraint applied, HO = half of the gold-standard subsample overlaps across occasions, NO = none of the gold-standard subsample overlaps across occasions. Relative bias values greater than 10% are bolded.

Equal bias over time

When the generating population had a constant level of response bias over time, the only appreciable ARB appeared when the gold standard was measured only at occasion 4 and the BI constraint was not imposed (11.3% in the structural paths). In all other conditions, ARB was less than 10%, because this model was always correct: no matter when response bias was estimated, the available information was enough to characterize the response bias in the inexpensive measure at all four measurement occasions.

When the amount of bias in the inexpensive measure did not vary, the assumption underlying the BI constraint was true, and imposing this constraint resulted in more accurate response bias loadings and structural parameter estimates than imposing only the factorial invariance constraint. ARB of loadings on the response bias factor ranged from 0.8% to 2.9% (mean = 1.7%) when the BI constraint was imposed,

compared to 2.3% to 5.9% (mean = 3.4%) when it was not. On the construct loadings, ARB ranged from 4.1% to 5.7% (mean = 5.0%) when the BI constraint was imposed, and from 4.4% to 6.5% (mean = 5.4%) when it was not. ARB of structural parameters ranged from 2.2% to 3.9% (mean = 3.1%) when the BI constraint was imposed and from 1.2% to 11.3% (mean = 6.5%) when it was not.

The number or location of occasions when the gold-standard measure was included had no noticeable effect on ARB of any parameter estimates when the BI constraint was imposed and a small effect when it was not. With only factorial invariance constrained, models where the gold standard was measured at 2 occasions rather than 1 had slightly higher ARB in the loadings (5.1%–6.4% for 2 occasions; 4.4%–4.9% for 1 occasion) and slightly lower ARB in the structural parameters (3.1%–7.0% for 2 occasions; 4.4%–11.3% for 1 occasion). Compared to 1 or 2 occasions, when the gold standard was measured at all 4 occasions, ARB on the loadings was high (6.3%) and ARB on the structural parameters was low (1.2%).

Though none of these numbers represent substantial ARB, the model that has least ARB in structural parameters should be preferred. Thus, the most accurate design when the inexpensive measure has equal bias over time appears to be one where the gold standard is measured at occasion 1 (or two occasions on a fully overlapping group of participants), with both factorial invariance and equality of bias constraints.

Random bias over time

When the amount of bias in the inexpensive measure varied randomly over time, substantial bias appeared in both the response bias loadings and the structural paths when the BI constraint was applied. That is, when the amount of bias was constrained to be equal over time, but the true degree of bias varied randomly over time, loadings on the response bias construct showed around 19% ARB, and the average ARB of all structural paths ranged from 10.6% to 21.5% ARB. The bias in structural paths was not uniform across conditions or across structural parameters; for example, the regression path from time 1 to 2 was always positively biased, whereas that from time 2 to 3 was negatively biased (not shown).

When the BI constraint was not imposed, the only appearance of ARB greater than 10% was in the structural paths when the gold standard was measured only at the first occasion (11.6%). Each condition where the BI constraint was imposed had about 2% lower ARB in construct loadings compared to when it was not imposed, but construct loading ARB was never higher than 6% in any condition.

The number or location of occasions when the gold-standard measure was included had little effect on ARB when the BI constraint was applied, except that smallest ARB in structural paths occurred in the condition where the gold standard was measured at occasions 1 and 4 on a fully overlapping group of participants (10.6%). When the BI

constraint was not imposed, structural parameter estimate ARB was smallest when the gold standard was measured at all four occasions (3.7%), slightly greater when the gold standard was measured at 2 occasions (ranging from 4.5% to 5.9%), and highest when the gold standard was measured at just one occasion (ranging from 5.1% to 11.6%). The proportion of overlapping cases had no effect.

In sum, when the degree of bias in the inexpensive measure varies randomly over time, the most accurate models are those that do not constrain bias to be equal over time and where the gold standard is measured at every occasion, although designs where the gold standard is measured at any one or any two occasions also performed well.

Increasing bias over time

When the amount of bias in the inexpensive measure increased systematically over time, the pattern of parameter ARB strongly resembled that of the random bias condition. When the BI constraint was not imposed, no model had substantial ARB. When the BI constraint was imposed, however, models showed 21.8%–25.3% ARB in response bias loadings, and 20.9%–33.3% ARB in structural paths (see Table 4), with one exception: when the gold standard was measured at occasions 1 and 4 with fully overlapping cases, ARB in the structural parameter estimates was just 11.5%. Apart from this one condition, the location or number of occasions at which the gold standard was measured did not affect ARB. ARB in construct loadings was not large, ranging from 2.3% to 9.1%. As with the random bias condition, the assumption of equal bias variance across measurement occasions in this condition is false, and the consequences of imposing the BI constraint are substantial.

The ARB in the structural path estimates showed the same pattern as when the population bias varied randomly. This bias was not uniform across conditions, or across structural parameters; the regression path from occasion 1 to 2 was always positively biased, as was the path from occasion 1 to 3; and the other three paths were always negatively biased.

When the BI constraint was not imposed, ARB of the response bias loadings ranged from 1.2% to 3.8%, ARB in construct loadings ranged from 4.1% to 5.7%, and ARB in regression paths among latent variables ranged from 3.5% to 9.4%. There was no noticeable effect of the number or location of occasions when the gold standard was measured. This finding suggests that, even if bias systematically differs over time, whether the gold standard is measured at a single time point or at all 4 time points, bias in the inexpensive measure can be accurately estimated. In this situation, the most accurate design is any design where factorial invariance is imposed and response bias is freely estimated over time.

Relative efficiency of parameters estimates

Table 5 displays the RE of each type of parameter for each condition. Recall that RE was computed by taking a ratio of the complete data sampling variance of each parameter estimate to the missing data sampling variance, where the complete data model contained the same variables as the incomplete data model. RE measures the degree to which information is retained as a result of missing data. For example, RE of .60 means that the parameter estimate is 60% as efficient as it would have been if every variable had complete data.

Table 5. Percent average relative efficiency.

Table 5. Percent average relative efficiency.

		Equal bias			Random bias			Increasing bias		
Design	% subsample	Bias loading	Construct loading	Structural paths	Bias loading	Construct loading	Structural paths	Bias loading	Construct loading	Structural paths
GM all	10	16.0	18.4	22.4	17.9	22.8	22.2	22.3	24.4	23.8
GM 1	40	45.6	56.6	36.0	43.5	58.9	31.9	45.3	60.3	48.9
GM 2	40	43.3	48.1	36.0	45.9	61.1	44.3	48.0	63.8	51.1
GM 3	40	45.4	42.7	12.1	49.6	62.4	47.5	53.3	65.5	45.4
GM 4	40	40.3	36.9	15.2	47.1	53.5	39.6	50.8	50.9	37.3
GM 1–2	20	25.5	34.1	33.5	27.4	41.8	31.1	32.5	41.2	35.1
GM 1–4	20	22.8	37.0	25.1	22.4	41.1	27.7	23.6	42.7	33.8
GM 1–4 HO	20	25.1	34.2	24.4	28.6	42.6	30.2	34.4	43.5	35.8
GM 1–4 NO	20	24.9	36.8	23.5	29.9	47.6	28.2	35.3	43.9	32.7
GM all BI	10	21.8	27.4	43.7	19.7	27.5	41.5	18.4	27.4	38.4
GM 1 BI	40	55.4	66.0	77.8	50.0	63.2	80.0	46.9	61.3	78.0
GM 2 BI	40	58.1	67.6	80.1	56.0	72.5	78.3	55.2	71.0	75.6
GM 3 BI	40	46.9	66.8	78.5	60.3	72.8	81.1	59.5	73.5	78.9
GM 4 BI	40	50.3	68.3	77.6	56.0	70.0	77.8	57.2	70.4	72.5
GM 1–2 BI	20	38.0	44.3	64.9	26.4	42.9	64.7	28.3	42.3	62.0
GM 1–4 BI	20	29.6	46.5	68.2	32.3	47.2	69.3	32.5	46.6	69.3
GM 1–4 BI HO	20	37.2	44.7	65.8	36.1	47.6	64.0	34.5	47.9	60.3
GM 1–4 BI NO	20	36.4	45.2	66.0	39.0	48.8	64.2	36.5	49.5	60.1

Note. GM = gold-standard measure, 1–4 indicates the occasions at which the gold-standard measure was administered (see Table 2), FIC = factorial invariance constraint, FBC = factorial invariance and equality of bias constraints, HO = half of the gold-standard subsample overlaps across occasions, NO = none of the gold-standard subsample overlaps across occasions. % subsample refers to the percent of the sample who received the gold-standard measure at the time points when it was measured.

It is helpful to interpret RE in light of the percentage of observations that are missing. Recall that in order to compare missing data designs to complete data designs using the same model, the complete data design that is being compared differs according to the number of occasions on which the gold standard is measured. For example, a design with 50 participants measured on the gold-standard measure at each time point is compared to a complete data design where all 500 participants are measured on the gold-standard measure at each time point, representing a 10% subsample of the complete data design. A design with 200 participants getting the gold-standard measure at just one time point is compared to a complete data design where all 500 participants get the gold-standard measure at only that time point, representing a 40% subsample of the complete data design. In the former case, RE values greater than 10% represent a "savings" in terms of efficiency-per-observation. In the latter case, RE values greater than 40% represent such a savings. The subsample percentage for each model is given in Table 5 as a baseline for comparison.

With eight exceptions, all parameters in all conditions displayed higher RE than subsample percentage. Five exceptions occurred when bias was stable across time, when the gold standard was measured at just one occasion, and when the BI constraint was not imposed. Four of these five exceptions affected the structural paths, the other affected construct loadings.

The choice of constraint had a large effect on efficiency. In almost every condition and for every parameter type, the model with the BI constraint imposed produced more efficient estimates than the one without. For example, when the gold standard was measured on just 20% of the sample at 2 occasions, RE of structural parameter estimates ranged from 60.1% to 69.3% with the BI constraint imposed, compared to 24%–26% without it. Construct and response bias loadings were also about 6% more efficient when the BI constraint was imposed.

Whether bias was equal, varied randomly, or systematically increased over time, patterns of RE were almost identical. Imposing the BI constraint was almost always the most efficient strategy, but recall that these conditions also resulted in the most extreme ARB when the degree of bias in the inexpensive measure varied over time. High efficiency and high ARB are, of course, a poor combination, as it results in small confidence intervals around inaccurate parameter estimates, encouraging a false impression of accuracy. When the BI constraint was not imposed, RE rates suggest that all of these designs still represent a high cost savings in terms of efficiency per piece of data collected. The greatest efficiency relative to the subsample gains appear in the condition where the gold standard is measured at 2 occasions, followed by 1 occasion, followed by all 4 occasions.

Method: Simulation 2

In Study 1, the generating model and analysis models included just one response bias factor for all four measurement occasions. In reality, response bias is unlikely to be perfectly correlated across measurement occasions. Instead, it is more likely that response bias also follows a longitudinal structure similar to that of any core construct, correlated across time but nonetheless distinct. Thus, in simulation 2, we respecified the population generating model to include a separate response bias factor at each time point and replicated the random bias population condition of Study 1 using this more complex model (see Figure 3). The response bias factor correlations were specified to follow a simplex structure with lag 1 correlations of .7, lag 2 correlations of .49, and lag 3 correlations of .343. Corresponding to this change in the population model, we investigated two analysis models: a) the same structure as the data generation model with a response bias factor at each time; and b) the model with one response bias factor that was used in Study 1 (Figure 2).

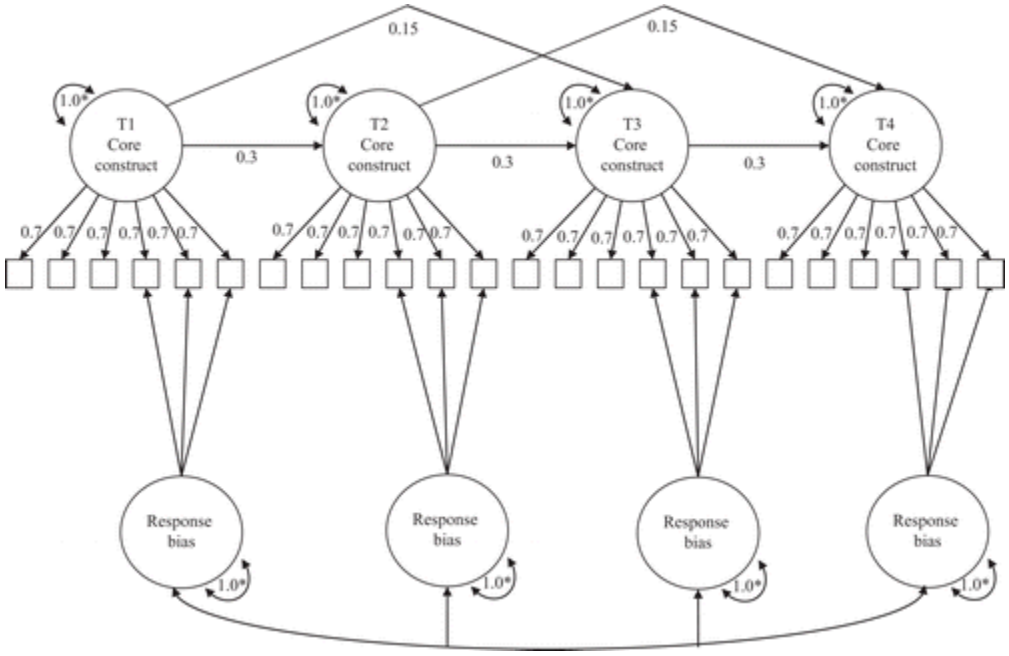

Figure 3. Data generation model for Study 2. Residual variances and residual covariances (at lag 1 and 2) are not shown.

For simulation 2 we only included the results of the analysis models that include the gold-standard measure at every time point. In order to capture a proper change in the response bias factor it is necessary to have the gold-standard measure at every time point; models lacking this characteristic would give biased parameter estimates, even with complete data. We confirmed this hypothesis by testing the design with complete data at one, two, or all time points. This analysis revealed the same trend: the only models with ARB lower than 10% were the ones that included the gold-

standard measure at every measurement occasion. Thus, to model separate response bias constructs at each wave, the gold standard must be measured at every wave. With the gold standard at only a subset of occasions, the only plausible model is a single response bias factor to account for bias across all occasions.

Rates of convergence and improper solutions

Table 6 shows the percent of nonconvergence and improper solutions. The improper solutions include results with at least one negative variance estimate, or at least one negative standard error estimate.

Table 6. Percent of no convergence and improper solutions.

	1 Bias factor			4 Bias factor			
Design	No conv	Imp sol	Total est pr	No conv	Imp sol	Total est pr	Conv imp
GM all	0.2	9.4	9.6	6.0	31.8	37.8	87
GM all BI	0.0	0.0	0.0	20.4	15.8	36.2	78.2

Note. No conv = no convergence; Imp sol = improper solutions are at least one negative standard error estimate or at least one negative variance estimate; Total est pr = total estimation problems, Conv imp = convergence rate including the improper solutions where negative variance estimates were no more than 2.58 standard errors from zero, GM = gold-standard measure, BI = bias invariance constraint applied.

The analysis models with one bias factor had low percent of improper solutions (9.4%) and almost no convergence failures (0.2%). When the BI constraint was applied, the models had no convergence failures. The analysis model with four response bias factors, in contrast, showed severe estimation problems. When only the FI constraint was applied, there were a total of 37.8% estimation problems, with most of them being improper solutions (31.8%). When the BI constraint was also applied, there was a total of 36.2% estimation problems, and most of these (20.4%) were failures to converge.

The problems of nonconvergence and improper solutions are due to empirical underidentification (Kenny & Kashy, 1992). This issue is common in multi-trait multi-method (MTMM) models, and the multifactorial response bias structure resembles an MTMM model. To gauge the seriousness of the estimation problems, we randomly selected four data sets that did not converge properly with four response bias factors in the simulation and attempted to fix the estimation problems. By trying different sets of starting values, one data set converged properly. Another data set converged properly after removing one indicator of the inexpensive measure at one time point; and we were unable to get the two remaining data sets to converge without improper solutions by adding starting values, additional constraints, or by excluding individually problematic indicators. This informal investigation suggests that in applied situations it may be possible to raise the likelihood of a proper model solution somewhat, but not in every case. These models would certainly require more patience and labor, with unknown results.

Given these estimation problems, we analyzed the results with improper solutions included, because excluding a large proportion of the replications would lead to

skewed results. Following Kolenikov and Bollen (2012) and Savalei and Kolenikov (2008) a negative residual variance may not be evidence of model misspecification. In particular, as some of our residual variances were close to zero, sampling variability would be expected to lead to some negative variance estimates. Thus, we present results both excluding all improper solutions, as well as including those improper solutions where the negative residual variance was less than 2.58 standard errors from zero (the 99% confidence interval). The last column on Table 6 shows the percentage of convergence including improper solutions. The smallest percentage of replications used in an analysis, given this criterion, was 78.2% for the model with bias invariance constraint.

Absolute relative bias of parameter estimates

Table 7 displays the ARB for the response bias loadings, construct loadings, and structural regression paths by each response bias factor structure (i.e., 1 bias factor, 4 bias factors, and 4 bias factors with improper solutions included). Because ARB less than 10% is considered negligible (Muthén et al., 1987), values higher than 10% are bolded.

Table 7. Percent average absolute relative bias.

		1 Bias Factor			4 Bias factor			4 Bias factors imp		
Design	*df*	Bias loadings	Construct loading	Structural paths	Bias loadings	Construct loadings	Structural paths	Bias loadings	Construct loadings	Structural paths
GM all	220	**53.9**	7.6	**11.5**	3.7	5.1	6.2	**44.4**	5.7	5.9
GM all BI	229	8.3	5.3	**22.0**	**20.9**	3.8	7.0	**19.3**	3.9	7.1

Note. 4 Bias factors imp = results that include the proper convergent and those improper solutions where negative variance estimates were no more than 2.58 standard errors from zero, GM = gold-standard measure, BI = bias invariance constraint applied. Relative bias values greater than 10% are bolded.

One bias factor

The model with response bias loadings unconstrained across time show non-negligible ARB on response bias loadings, with an average ARB of 53.9%. The construct loadings had negligible ARB of 7.6%. In the case of the structural paths, the model's ARB is small but non-negligible (11.5%).

The model with the BI constraint applied showed negligible ARB in the response bias loadings of 8.3%, no substantial ARB in the construct loadings (5.3%), but high levels of ARB (22%) in the structural paths. The ARB in the structural paths was not uniform across structural parameters; the regression path from time 1 to time 2 tended to be negatively biased, as did the path from time 2 to time 3, and the path from time 3 to time 4. The regression paths from time 1 to time 3, and from time 2 to time 4 were always positively biased.

Four bias factors

When four bias factors were modeled, construct loadings were accurately estimated (ARB was never higher than 10%). For the response bias loadings, the model with unconstrained bias loadings had negligible bias (3.7%), but non-negligible bias when the BI constraint was applied (20.9%). In the case of structural paths both models (with and without the bias invariance constraint) showed negligible bias, with response bias loadings unconstrained (6.2%) and constrained (7.0%).

These results are limited to the low number of proper convergent results. When improper solutions were included in the analysis, both models have high ARB on the response bias loadings (unconstrained bias loading = 44.4%, constraint response bias loadings = 19.3%), but the ARB was never higher than 10% for the construct loadings or structural paths.

Relative efficiency of parameters estimates

Table 8 displays the RE of each type of parameter for each condition. Recall that a RE higher than the percentage of subsample of the complete data design represents a "savings" in terms of efficiency-per-observation. All parameters in all conditions displayed higher RE than subsample percentage, indicating a savings in efficiency.

Table 8. Percent average relative efficiency.

		1 Bias factor			4 Bias factor		
Design	% subsample	Bias loading	Construct loading	Structural paths	Bias loading	Construct loading	Structural paths
GM all	10	15.9	27.9	23.3	18.6	15.1	12.3
GM all BI	10	14.1	34.3	48.4	19.7	16.1	12.1

Note. GM = gold-standard measure, BI = bias invariance constraint applied. % subsample refers to the percent of the sample who received the gold-standard measure at the time points when it was measured.

For the model with four response bias factors and the response bias loadings unconstrained, the RE for the response bias loadings was 18.6%, RE for the construct loadings was 15.1%, and RE for the structural paths was 12.3%. For the model with the BI constraints imposed, the RE for the response bias loadings was 19.7%, RE for the construct loadings was 16.1%, and RE for the structural paths was 12.1%. The analysis model with only one response bias factor had lower RE than the four response bias factor model on the response bias loadings (unconstraint bias loadings = 15.9%, constraint bias loadings = 14.1%), but it has higher RE for the construct loadings (unconstrained bias loadings = 27.9%, constrained bias loadings = 34.3) and structural paths (unconstrained bias loadings = 23.3%, constrained bias loadings = 48.4).

General discussion

We examined the performance of the two-method measurement design in longitudinal models. Simulation 1 showed that when the amount of bias in the inexpensive measure does not change over time, all designs produce unbiased estimates. Under more realistic conditions, when the bias randomly varies or increases across time, models that impose an equality of response bias constraint result in substantial bias, especially in the structural paths (i.e., regression coefficients between factors over time, which are typically the parameters of most interest). Imposing the BI constraint results in least accurate parameter estimation when the amount of bias in the inexpensive measure increases over time. When the gold-standard measure is included at two or more occasions, this bias is smaller. Measuring the gold standard at both the first and last occasions on the same group of participants resulted in the most accurate parameter estimates, but the degree of parameter bias was always unacceptably high as long as the BI constraint was imposed. When the BI constraint was not imposed, no substantial bias appeared in any condition.

Almost every factor in simulation 1 resulted in parameter estimates that were more efficient per piece of data collected than a complete data design. In all cases, models that imposed the BI constraint resulted in much more efficient estimates – but when response bias is not equal over time, substantially biased parameter estimates emerged. When the factorial structure of the response bias across time is multifactorial (simulation 2) the gold-standard measure and a response bias factor must be included at every time point. These features are necessary to estimate properly the change in the response bias across time.

This multifactorial structure increases the rate of improper solutions and convergence failures due to empirical underidentification. Even when the multifactorial model is more theoretically appropriate, it is in practice difficult to estimate. When the multifactorial model converges, it is likely to result in less biased estimates (especially for the structural paths) whereas imposing the BI constraint led to higher ARB on the response bias loadings, while it didn't impacted the construct loadings, and structural paths.

Recommendations

When choosing the details of the two-method measurement design and analysis model for longitudinal research, it is important to consider whether the degree of bias in the inexpensive measure is likely to be relatively stable over time, or whether it is likely to change; and it is relevant to consider the factorial structure of the response bias over time. Given that the overlapping of the gold-standard measure didn't had a noticeable effect it is possible for the researchers to think of designs where the

participants randomly receive the gold-standard measure, this depending on the resources and nature of the gold-standard measure.

When bias is equal over time, the most accurate and efficient design is any one where both the factorial invariance and equality of bias constraints are imposed across measurement occasions. Whether the gold standard is measured at 1, 2, or all waves, makes little difference when the response bias is unifactorial. When the bias in the inexpensive instrument changes over time, however, the most accurate and efficient design is one where the gold standard is measured at more than one occasion, and only factorial invariance over time is constrained.

When the response bias factor is expected to change over time, the only recommended design is to include the gold-standard measure at every time point. Other designs are biased even with complete data. The only way to estimate properly the four response bias factors is with this design. The other option is to estimate the model with only one response bias factor, but it still requires the gold-standard measure at each wave. Imposing factorial invariance (as we did for every model studied) never led to biased parameter estimates because factorial invariance was true in the data generating model. When the factorial structure differs from the analysis structure model even factorial invariance models can lead to biased results. Before using the factorial invariance constraint, therefore, it is important to test that factor loadings and intercepts are actually invariant over measurement occasions. To test if the factorial invariance constraint is appropriate, we recommend estimating a Confirmatory Factor Analysis (CFA) on the target factor (with the unconstrained response bias factor included in the model) without constraints, then estimating a model with the factor loadings constrained to be equal over measurement occasions while the factor variance at each occasion is freely estimated.

Our results suggest that bias invariance constraints should be imposed only in cases where there is good evidence that bias does not change over time. Substantial change in model fit between a model in which BI is not constrained and the one in which it is constrained would be evidence of changing bias; nonsignificant change between models would support the BI constraint. It is also important to test the factorial structure of the response bias over time. We recommend starting with the more appropriate theoretical structure and working to resolve estimation problems by changing starting values. Many improper convergence problems can be worked around by changing starting values or placing justifiable constraints on parameters. Models with a multifactorial structure of the response bias require due diligence.

Limitations and future research

A principal limitation is that there are no applied data (that we know and have access to) that show how the type of response bias we consider here might actually behave over time. Given this limitation we chose three possible ways in which bias may

behave over time, and 2 different factorial structures. Further applied research will give a better perspective on the behavior of the bias across time. In this research both the gold-standard measure and the biased measure had three indicators, but the number of indicators per type of measure could vary. In many cases the gold-standard measure may consist of one or two indicators. It is of special interest to see how these models may be affected by sample size and percent of sample that receives the gold-standard measure. Further research could give guidance about the minimum sample necessary to have a stable two-method design. It may be that smaller samples need a higher percent of the sample receiving the gold-standard measure. Another factor to include in future research is the effect of attrition, which is common in longitudinal studies.

*This article accepted during Marcel van Aken's term as Editor-in-Chief.

Funding

This study was supported by grant NSF 1053160 (Wei Wu & Todd D. Little, co-PIs) and by the Center for Research Methods and Data Analysis at the University of Kansas (when Todd D. Little was director; 2009–2013).

References

1 Alloway, T. (2007). Automated working memory assessment. London, UK: Pearson.

2 Armengol, C. (2002). Stroop test in Spanish: Children's norms. The Clinical Neuropsychologist, 16, 67–80.

3 Atkins, M. S., Pelham, W. E., Licht, M. H. (1985). A comparison of objective classroom measures and teacher ratings of attention deficit disorder. Journal of Abnormal Child Psychology, 13, 155–167.

4 Baraldi, A. N., Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psychology, 48, 5–37.

5 Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., Erbaugh, J. K. (1961). An inventory for measuring depression. Archives of General Psychiatry, 4, 561–571.

6 Brener, N. D., Kann, L., McManus, T., Kinchen, S. A., Sundberg, E. C., Ross, J. G. (2002). Reliability of the 1999 youth risk behavior survey questionnaire. Journal of Adolescent Health, 31, 336–342.

7 Brophy, M., Taylor, E., Hughes, C. (2002). To go or not to go: Inhibitory control in "hard to manage" children. Infant and Child Development, 11, 125–140.

8 Eugene, F., Joormann, J., Cooney, R. E., Atlas, L. Y., Gotlib, I. H. (2010). Neural correlates of inhibitory deficits in depression. Psychiatry Research: Neuroimaging, 181, 30–35.

9 Gatti, R., Antonello, G., Prearo, M., Spinella, P., Cappellin, E., De Palo, E. F. (2009). Cortisol assays and diagnostic laboratory procedures in human biological fluids. Clinical Biochemistry, 42, 1205–217.

10 Genius, J., Klafki, H., Benninghoff, J., Esselman, H., Wiltfang, J. (2012). Current application of neurochemical biomarkers in the prediction and differential diagnosis of Alzheimer's disease and other neurodegenerative dementias. European Archives of Psychiatry and Clinical Neuroscience, 262, 71–77.

11 Gioia, G. A., Isquitch, P. K., Guy, S. C., Kenworthy, L. (2000). Behavior rating inventory of executive function. Child Neuropsychology, 6, 235–238.

12 Graham, J. W. (2012). Missing data: Analysis and design. New York, NY: Springer.

13 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.

14 Hamilton, M. (1967). Development of a rating scale for primary depressive illness. British Journal of Social and Clinical Psychology, 6, 278–296.

15 Hutchinson, K. A., Balota, D. A., Ducheck, J. M. (2010). The utility of stroop task switching as a marker for early-stage Alzheimer's disease. Psychology and Aging, 25, 545–559.

16 Kolenikov, S., Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? Sociological Methods & Research, 41, 124–167. doi:10.1177/0049124112442138

17 Kenny, D. A., Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin, 112, 165–172. doi:http://dx.doi.org.www2.lib.ku.edu:2048/10.1037/0033-2909.112.1.165

18 Levine, A., Zagoory-Sharon, O., Feldman, R., Lewis, J. G., Weller, A. (2007). Measuring cortisol in human psychobiological studies. Physiology & Behavior, 90, 43–53.

18 MacCallum, R. C., Browne, M. W., Sugawara, H. W. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, 1, 130–149.

19 Muthén, B., Kaplan, D., Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. Psychometrika, 52, 431–3462.

20 Olchowski, A. E. (2007). Assessing the impact of physical conditioning, dietary intake, body fat, and tobacco use on blood pressure parameters: A two-method measurement design approach (Doctoral dissertation). Retrieved from *ProQuest Dissertations and Theses*, 180-n/a; http://search.proquest.com/docview/304836966?accountid=14556. (304836966).

21 Pornprasertmanit, S., Miller, P., Schoemann, A. (2013). simsem: simulated structural equation modeling. R package version 0.5–3. Retrieved from http://www.simsem.org

22 R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from

23 Ready, D. D., Wright, D. W. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. American Educational Research Journal, 48, 335–360.

24 Rhemtulla, M., Little, T. D. (2012). Planned missing data designs for research in cognitive development. Journal of Cognition and Development, 13, 425–438.

25 Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48, 1–36.

26 Rothbart, R., Ahadi, S., Hershey, K., Fisher, P. (2001). Investigations of temperament at three to seven years: The children's behavior questionnaire. Child Development, 72, 1394–1408.

27 Savalei, V., Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. Psychological Methods, 13, 150–170. doi:10.1037/1082-989X.13.2.150

28 Schmand, B., Eikelenboom, P., Van Gool, W. A. (2011). Value of neuropsychological tests, neuroimaging, and biomarkers for diagnosing Alzheimer's disease in younger and older age cohorts. Journal of the American Geriatrics Society, 59, 1705–1710.

29 Seifer, R., Sameroff, A. J., Barret, L. C., Krafchuk, E. (1994). Infant temperament measured by multiple observations and mother report. Child Development, 65, 1478–1490.

Szabo, M. (2010). The short version of the depression anxiety stress scales (DASS-21): Factor structure in a young adolescent sample. Journal of Adolescence, 33, 1–8.

30 Taylor, B. J., Graham, J. W., Palmer, R. F., Tatterson, J. W. (1998, June). Interpreting latent variable models involving self report and objective measures. Paper presented at the annual meeting of the Society for Prevention Research, Park City, UT.

31 Wechsler, D. (2004). The Wechsler intelligence scale for children – fourth edition. London, UK: Pearson Assessment.

32 Yao, S., Liu, M., Liu, J., Hu, Z., Yi, J., Huang, R. (2010). Inhibition dysfunction in depression: Event-related potentials during negative affective priming. Psychiatry Research: Neuroimaging, 182, 172–179.