Psychological Methods
2006, Vol. 11, No. 4, 323–343

1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.4.323

Planned Missing Data Designs in Psychological Research

John W. Graham, Bonnie J. Taylor, and Allison E. Olchowski
Pennsylvania State University

Patricio E. Cumsille
Pontificia Universidad Católica de Chile

The authors describe 2 efficiency (planned missing data) designs for measurement: the 3-form design and the 2-method measurement design. The *3-form design*, a kind of matrix sampling, allows researchers to leverage limited resources to collect data for 33% more survey questions than can be answered by any 1 respondent. Power tables for estimating correlation effects illustrate the benefit of this design. The *2-method measurement design* involves a relatively cheap, less valid measure of a construct and an expensive, more valid measure of the same construct. The cost effectiveness of this design stems from the fact that few cases have both measures, and many cases have just the cheap measure. With 3 brief simulations involving structural equation models, the authors show that compared with the same-cost complete cases design, a 2-method measurement design yields lower standard errors and a higher effective sample size for testing important study parameters. With a large cost differential between cheap and expensive measures and small effect sizes, the benefits of the design can be enormous. Strategies for using these 2 designs are suggested.

Keywords: planned missingness, measurement efficiency, matrix sampling, multimethod measurement, structural equation modeling

Supplemental material: http://dx.doi.org/10.1037/1082-989X.11.4.323.supp

Cost effectiveness and design efficiency in various forms have long been considerations in research design. This article focuses on two approaches to design efficiency relating to measurement. We bring together well-known concepts of measurement and sampling on the one hand and add new analysis concepts relating to missing data analysis on the other.

In a practical sense, it is the recent development of missing data analysis tools such as multiple imputation (Rubin, 1987; Schafer, 1997) and the increasingly popular full information maximum likelihood (FIML) procedures in various structural equation modeling (SEM) packages that makes these new ideas feasible. Just a few years ago, the designs described in this article would not have been practical. However, with the recent advances in analysis with missing data, all this has changed. With the multiple imputation and FIML procedures now widely available, researchers are able to perform most mainstream analyses whether there are missing data or not. More important, these recent analytic procedures allow researchers to obtain unbiased parameter estimates and reasonable standard errors over a wide range of data circumstances, including the planned missing data designs described in this article.

John W. Graham and Allison E. Olchowski, The Methodology Center and Department of Biobehavioral Health, Pennsylvania State University; Bonnie J. Taylor, Department of Health Evaluation Sciences, Pennsylvania State University; Patricio E. Cumsille, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile.

Bonnie J. Taylor is now at the Office of Injury Prevention, New Mexico Department of Health, Sante Fe.

Portions of this research were presented as a Division 5 invited address at the 111th Annual Convention of the American Psychological Association, Toronto, Ontario, Canada, August 2003. This research was supported in part by National Institute on Drug Abuse Grant P50 DA 10075. We thank Linda M. Collins for several very helpful comments on earlier versions of this work.

Correspondence concerning this article should be addressed to John W. Graham, Department of Biobehavioral Health, E-315 Health & Human Development Building, Pennsylvania State University, University Park, PA 16802. E-mail: jgraham@psu.edu

Plan of This Article

In order to set the stage for our discussion of planned missing data designs, we have to say a little about the missing data analyses, without which planned missing data designs would not be feasible. Next, because we view planned missing data designs as members of the general class of efficiency-of-measurement designs, we discuss briefly efficiency designs in other research contexts. Then we discuss a class of efficiency-of-measurement designs referred to as *matrix sampling*. We focus on a special case of this general class of design that we have referred to as the

3-form design (Graham, Hofer, & Piccinin, 1994). We also touch on other designs in the same family of designs. We compare the 3-form design and the complete cases (1-form) design, discussing (a) the assumptions underlying the designs, (b) trade-offs, (c) power implications, and (d) practical considerations for deciding which design to use.

Finally, we describe another class of design for efficiency of measurement that we refer to as *two-method measurement*. This design involves one cheap, lower quality measure of a construct and one expensive, high quality measure of the same construct. We present simulated data showing that a planned missing data design involving many cases with just the cheap measures and few cases with both the cheap and expensive measures yields the best (lowest) standard errors for testing hypotheses of substantive interest. We discuss (a) assumptions underlying this class of designs, (b) trade-offs, (c) power implications, and (d) practical considerations for deciding which variation of the design to use.

Missing Data Analysis

The value of the planned missing data designs presented in this article hinges on one's ability to make use of analysis procedures that handle missing data. In this section, we talk very briefly about the current state of analysis with missing data. For a more thorough recent treatment of this topic, please see Schafer and Graham (2002). Also, one of the classic works in this area has recently been updated (Little & Rubin, 2002). For a less technical recent discussion of missing data analysis, please see Graham, Cumsille, and Elek-Fisk (2003).

The main thing to be clear about is that missing data analysis is very well established. The fears that researchers had in the beginning, nearly 2 decades ago, have all but disappeared. These fears have been replaced with a cautious confidence that is based on ever increasing experience with analyses involving the recommended procedures. In a very real sense, what was a rather difficult-to-sell analysis strategy just 10 years ago is becoming mainstream.

Kinds of Missingness

There are two kinds of missingness: *missing at random* (MAR) and *missing not at random* (Schafer & Graham, 2002). A special case of MAR missingness, which is often considered separately, is *missing completely at random* (MCAR). Missingness is essentially MCAR as long as $r_{ZY} = 0$, where Z is a variable that represents the cause of missingness and Y is the variable containing the missingness. For the present article, MCAR is the most relevant kind of missingness. For the two designs we consider in detail, the missingness is entirely under the researcher's control, making the MCAR assumptions entirely reasonable. For a more detailed discussion of the other forms of missingness, please refer to Schafer and Graham (2002).

Missing Data Analysis Approaches and Software

The two major approaches to analysis with missing data are multiple imputation and maximum likelihood, or FIML, procedures. Either of these approaches represents an excellent way of dealing with the missing data created by the designs described in this article. Multiple imputation (Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002; see also Graham et al., 2003; King, Honaker, Joseph, & Scheve, 2001; Raghunathan, Lepkowski, VanHoewyk, & Solenberger, 2001) takes a two-step approach to handling the missing data. The first step deals with the missing data and produces several imputed data sets. In the second step, the researcher uses standard complete cases analysis procedures to analyze each of the imputed data sets. The results are then combined into a single result that looks much like results from complete cases analyses. Several implementations of multiple imputation are currently available.

The maximum likelihood approach is to develop new procedures that allow the researcher to deal with the missing data and perform the usual parameter estimation, all in a single step. The maximum likelihood, or FIML, approach has been popularized in recent years in several SEM packages, including, in alphabetical order, Amos (Arbuckle, 1995; Arbuckle & Wothke, 1999), LISREL 8.5+ (Jöreskog, & Sörbom, 1996; also see du Toit & du Toit, 2001), MPlus (L. K. Muthén & Muthén, 1998), and Mx (Neale, 1991; Neale, Boker, Xie, & Maes, 1999). LTA (Collins, Hyatt, & Graham, 2000; Hyatt & Collins, 2000; Lanza, Collins, Schafer, & Flaherty, 2005), a latent class procedure, is also an example of a program using the FIML approach.

Efficiency Designs: A Backdrop

The kinds of designs presented in this article are related to a general class of efficiency designs for research. Random sampling of respondents is perhaps one of the best known efficiency designs (e.g., see Thompson, 2002; Thompson & Collins, 2002). Two other types of design that are often described in the efficient design literature are *optimal designs* (e.g., see D. B. Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997; Atkinson & Donev, 1992; Cox, 1958; Gelman, 2000; McClelland, 1997) and *fractional factorial* (and related) *designs* (e.g., see Box, Hunter, & Hunter, 1978; Collins, Murphy, Nair, & Strecher, 2005; West & Aiken, 1997). An essential characteristic of these designs is that they tend to be variants of experimental designs. They often focus on the independent variable and involve efficiencies deriving from aspects of assignment to conditions in experimental studies. These designs attempt to limit the scope of the research without placing undue limits on what actually is studied.

Efficiency Designs for Measurement: I. Matrix Sampling

Early Designs

Another class of efficiency design focuses on measurement, per se. In this section, we talk about measurement designs that fall generally under the heading of *matrix sampling* (e.g., Shoemaker, 1973). The simplest version of this type of measurement design involves administering different items to different respondents, for example, as shown in Table 1 (Johnson, 1989, 1992; Lord, 1962; Munger & Loyd, 1988; Shoemaker, 1973). Researchers using this type of design were interested in reducing the amount of time it takes respondents to complete the survey. Indeed, studies involving this type of design have seen improved response rates, for example, from mail surveys (e.g., Adams & Darwin, 1982; Munger & Loyd, 1988). Researchers making use of simple matrix sampling designs have been interested mainly in estimating means. However, this design is less useful for estimating correlations between items (see Graham et al., 1994; Raghunathan & Grizzle, 1995).

More recent variations of the matrix sampling approach have extended the simple version such that there are some data not just for all items but also for most or all pairs of items. McArdle (1994) suggested the *fractional block design.* This design was an improvement over the simple matrix sampling designs; in addition to allowing estimation of means for all variables, the researcher is able to estimate correlations for most (but not all) pairs of variables. A limitation of this type of design is that it requires use of specialized SEM analyses (P. D. Allison, 1987; B. Muthén, Kaplan, & Hollis, 1987).

Johnson (1992) suggested the *balanced incomplete blocks* (BIB) *spiral* design. A desirable feature of this design is that means may be estimated for all variables, and correlations may be estimated for all pairs of variables. What makes the BIB spiral design *balanced* is that the same number of individuals respond to each item set and to each pair of item sets. With the design described by Johnson (1992), each pair of item sets was presented in exactly one form. One drawback to this design is that the balancing works only in the case of seven item sets and seven forms (however, see van der Linden, Veldkamp, & Carlson, 2004, for an example of a much larger BIB design).

Table 1
Example of Simple Multiple Matrix Design

	Blocks of items						
Form	A	B	C	D	E	F	G
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	0	1	0	0
6	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1

Note. 1 = questions asked; 0 = questions not asked. Letters A–G refer to different sets of items.

The 3-Form Design

Another approach to the matrix sampling concept, which has been in use since 1982 (Graham et al, 1984; Graham, Johnson, Hansen, Flay, & Gee, 1990; Hansen & Graham, 1991; Hansen, Johnson, Flay, Graham, & Sobel, 1988), is the *3-form design* (Graham et al., 1994; Graham, Hofer, & MacKinnon, 1996; Graham, Taylor, & Cumsille, 2001). This design was also developed as a way of limiting the time necessary for respondents to complete the survey. Or, more precisely, the goal was to ask more questions than could be answered by a single respondent. Also, with this design, it was important to be able to estimate all correlations as well as means and variances. The logic of its development went something like the following.

Suppose a researcher determines that subjects are willing to answer 30 questions. But suppose the researcher would like to ask a few more questions, say 40 questions. One solution to this common dilemma is to drop 10 items. Making the decision to drop 10 items is really a planned missing data design, the consequences of which are (a) that the researcher cannot ask any research questions involving the 10 dropped items and (b) that the analyses of the remaining 30 items involve the straightforward application of common complete cases procedures.

An alternative is to produce three different forms, such that a different set of 30 items is presented in each form. This limits to 30 the number of items given to each subject, a study requirement, but still allows for collection of data for all 40 items, which can be used to answer important research questions. Such a design is presented in Table 2 and has been described by Graham and colleagues (Graham et al., 1994, 1996, 2001). With this planned missing data design, items are divided into four item sets (X, A, B, and C), and questions are presented to subjects as shown in Table 2. With this version of the 3-form design, items in the X set, which are essential to the hypotheses under study, are asked of everyone. Items in the X set are often asked first, although that is not a requirement. In addition, the item sets in the three forms are rotated so that different item sets appear last in each form. For example, in the Adolescent Alcohol Prevention Trial (Hansen & Graham, 1991), items were presented as follows: Form 1, XAB; Form 2, XCA; Form 3, XBC.

Table 2
The 3-Form Design, With X Set

	Item set			
Form	X	A	B	C
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1

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

The Split Questionnaire Survey Design

Raghunathan and Grizzle (1995) suggested a similar variation of matrix sampling, which they called the *split questionnaire survey design* (SQSD). The SQSD, as they described it, involved 11 different forms, 1 of which contained all of the survey items (note that in all further discussion of the SQSD, we omit this latter form). An example of the SQSD, eliminating that 1 (complete data) form, appears in Table 3. This version has six item sets, including one set, X in Table 3, that was included in all 10 of the forms shown. The remaining five item sets (A, B, C, D, and E in Table 3) were presented to respondents in pairs, such that each combination of the five item sets, taken two at a time, was presented in exactly one form of the questionnaire. An important characteristic in common between the 3-form design and the SQSD (and other designs in this family) is that in each of these designs, all possible combinations of *L* item sets (excluding the item set X), taken two at a time, are represented in the various forms of the design. In the case of the 3-form design, all combinations of three item sets, taken two at a time, are represented in the three forms (see Table 2). For the SQSD, all combinations of five item sets, taken two at a time, are represented in the 10 forms (see Table 3).

The main focus in this article is the 3-form design. We compare that design with the 1-form (complete cases) design. As an example to be used throughout this part of the article, we start with the scenario described above: Considering all relevant factors (e.g., subject fatigue), we assume that the researcher has the resources to present each subject with 30 questions but would like to leverage those resources in order to collect data on 40 questions, 33% more than what is presented to each subject. We consider this to be a modest leveraging of resources. The conclusions we draw here apply to other situations in which the researcher wants to collect data on 33% more questions than can be answered by each subject. In a later section, we touch on other scenarios in which the researcher wishes to leverage resources to a greater extent, for example, by collecting data on twice as many questions as can be answered by a single subject.

Details and Refinements of the 3-Form Design

Is the X set needed with the 3-form design? With the version of the 3-form design described so far, one set of items, X, is asked of all subjects. The main reason for including the X set is that the questions in this set are central to the research, and there may be power concerns for hypotheses involving these variables. The idea is to have a design that allows at least some hypotheses to be tested with the full sample size. That is, having the X set provides a kind of hedge against the possibility that a hypothesis critical to the research is actually tested with lower power than expected. Although situations may arise for which the X set may be safely excluded, we argue that including the X set is almost always a good idea.

How large should the X set be? In our examples, and in most of our applications to date, all of the item sets (X, A, B, C) contained roughly the same number of variables. However, this is not a requirement. The X set could be rather larger than the other sets (e.g., A, B, and C). Making the X set larger allows more effects to be tested with the full sample, but at the cost of testing fewer effects overall. Making the X set smaller than the others means that more effects may be tested overall, but at the cost of testing fewer effects with the full sample.

What should be included in the X set? Variables involved in the most important study hypotheses would normally be included in the X set. Past studies making use of the 3-form design have been experimental studies looking at the effects of a prevention program on later substance use and abuse. Because substance use was the main dependent variable and because power for such tests is often rather limited, we have always included such variables in the X set. One might also examine the pattern of expected effect sizes. Variables associated with smaller effect sizes might be placed in item sets that would be tested with larger samples.

Table 3
Ten-Form, Six-Set Variation of the Split Questionnaire Survey Design, With X Set

	Item set					
Form	X	A	B	C	D	E
1	1	1	1	0	0	0
2	1	1	0	1	0	0
3	1	1	0	0	1	0
4	1	1	0	0	0	1
5	1	0	1	1	0	0
6	1	0	1	0	1	0
7	1	0	1	0	0	1
8	1	0	0	1	1	0
9	1	0	0	1	0	1
10	1	0	0	0	1	1

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

Placement of the X set. The most important reason for having an X set is that virtually everyone should provide answers to the questions in this set. For this reason, it makes sense that the X set should be near enough to the front of the survey that even the slowest or least motivated readers will get as far as the items in the X set. However, if the researcher is concerned about possible order effects, there is no reason why the X set should necessarily be absolutely first in the survey. It probably should just not be last. There is also no reason why the items in the X set should be all together in one part of the survey. Finally, it is possible that the X set could be located in different parts of the survey for different forms. Table 4 shows one such example, in which the X set has different locations in different forms but is never last in the survey.

Variations of the 3-form design. An interesting variation of the 3-form design asks all questions of all subjects but rotates item sets so that different forms have different item sets last (e.g., Flay et al., 1995; Taylor, Graham, Palmer, & Tatterson, 1998). With the Flay et al. (1995) version, the orders for the three forms looked like this: XABC, XCAB, and XBCA. With the Taylor et al. (1998) variation, the X set of items was divided up (on conceptual grounds) to yield the following orders: Form 1, X_1ABX_2C; Form 2, X_1CAX_2B; and Form 3, X_1BCX_2A. With this variation, virtually everyone was able to complete questions as far as the X_2 set, but some subjects did not have time to complete questions in the item set that followed X_2.

The goal of the Graham et al. (1994, 1996, 2001) version of the 3-form design was to select a number of items that most subjects could complete. With the Flay–Taylor variant of the 3-form design, many more subjects will fail to complete the questionnaire, mainly as a result of slow reading. Thus, with this variation, it is good to ask questions early in the questionnaire about reading speed and any other variable that might explain why some subjects will leave items blank at the end of the questionnaire (e.g., conscientiousness). By collecting data on the cause of missingness (e.g., reading speed) and including these variables in the missing data model, the missingness approximates MAR, and biases associated with this kind of missingness, however small, are controlled (Collins, Schafer, & Kam, 2001; Graham & Hofer, 2000; Graham & Schafer, 1999; Graham et al., 1994, 1996, 2003; Little & Rubin, 2002; Rubin, 1987; Schafer, 1997; Schafer & Graham, 2002).

Table 4
Alternative Orders for the 3-Form Design

Form	Item set order
1	XAB
2	XCA
3	XBC
4	AXB
5	CXA
6	BXC

Two caveats about the 3-form design. The first caveat is that it may be desirable to have complete data for at least a subset of cases. With the original variation of the 3-form design, the design yields data for all possible combinations of two items. Thus, all covariances are well estimated. However, with the original version of the design, some combinations of three or more variables do not have data. This could lead to limitations with some two- or three-way interactions; there could also be limitations relating to estimation of certain partial correlations. The Flay–Taylor variant of the 3-form design is one good way to solve this problem. However, this variant may be less useful for larger designs in this family of designs (e.g., the 10-form SQSD), because of the substantial increase in the number of items in such designs. For all of the designs, it may be practical simply to pay some respondents an extra amount to complete the entire survey; however, the logistics of this option will be awkward at times.

The second caveat is a practical one. When one makes use of the 3-form design, the question arises as to whether items within scales should appear all within a single item set or separated across the item sets. Graham et al. (1996) showed that the latter choice was clearly better from a statistical standpoint. Standard errors (for the regression coefficient of one scale predicting another) were always lower when scale items were divided up across the item sets (e.g., one scale item in Item Set A, one in Item Set B, one in Item Set C) than when they all appeared within a single item set. The main reason for this is that items within scales are generally more highly correlated with one another than they are with items from other scales. This fact can lead to better handling of the missing data and lower standard errors, if one divides up the scale items across item sets (see also Collins et al., 2001).

However, if the number of variables in the whole model exceeds the number that can reasonably be handled by any of the available missing data procedures, the strategy that is better in a statistical sense turns out to be a big headache in a logistical sense: For every analysis, one must use the best FIML or multiple imputation procedures. Reasonable compromises such as forming scale scores on the basis of available variables (see Schafer & Graham, 2002) are inadvisable under these circumstances. Thus, we now recommend that scale items be kept together within item sets (perhaps interspersed throughout the item set), even if it means somewhat less efficient estimation. Of course, this recommendation may change as the missing data analysis procedures advance to the point of being able to handle large numbers of variables. This issue may be less of a problem when the Flay–Taylor variant of the 3-form design is used, but even with that design, we would now recommend keeping scale items together within a single item set.

Power Considerations in the 3-Form Design

The main idea of the 3-form design (as with other designs in this family) is to leverage one's resources: One is able to collect data for more variables with the 3-form design than with the 1-form design. But it is not as simple as that. When one leverages one's resources in any context, there are always trade-offs. Understanding these trade-offs is necessary for making an informed decision about which design to use. Gaining a good understanding of the relevant trade-offs is a bit complicated. In this section, we attempt to walk the reader through these complications.

1. We show the number of correlations that may be tested with each design, along with the sample sizes and statistical power with which they are tested.

2. We introduce devices to help interpret basic power figures. We introduce two visual devices to help highlight the conditions under which the 3-form design performs less well (in terms of power) than the standard complete cases design (the ✗ mark) and the conditions under which the 3-form design performs better (the ✓ mark). Also, we use the concept of the *power ratio* (power based on the complete cases design divided by the power based on the 3-form design) to help quantify differences in power between the two designs.

3. Finally, we argue that the conditions under which the 3-form design performs less well (✗) are relatively rare. But more important, we show that how many potential undesirable outcomes there are with the 3-form design is largely under the researcher's control.

Sample sizes and the number of correlation effects tested. That one is able to collect data for more variables with the 3-form design than with the 1-form design is, to an extent, a benefit in and of itself. Having more variables implies that more means and standard deviations may be estimated. With most study sample sizes, estimates for all means and standard deviations will be stable and useful. However, one must also take into account the fact that with the 3-form design, some data are collected from less than the full sample. Thus some correlations (and related statistics) are tested with the full sample, some are tested with two thirds of the full sample, and some are tested with one third of the full sample.

Information about sample sizes and the number of correlation effects tested are presented in Figure 1. In general, the number of testable effects is the same as the number of unique correlations within and between sets. Assuming k variables within each item set, there are $k(k - 1)/2$ different two-variable correlation effects within each item set and k^2 two-variable correlation effects across any two sets.

The bottom part of each bar of Figure 1 represents the number of effects testable with the full sample size ($N = 300$ in this case). The middle bar (3-form design) represents the number of effects testable with the intermediate sample

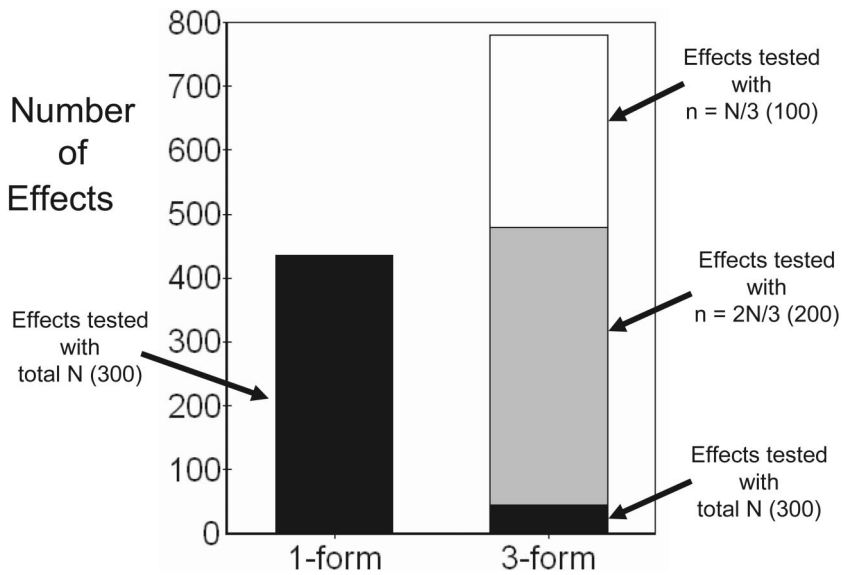

Figure 1. Number of testable hypotheses for the 1-form (complete cases) and 3-form designs. For the 1-form design, 435 effects are testable with the full sample size; for the 3-form design, 45 effects are testable with the full sample, an additional 435 effects are testable with two thirds of the full sample, and 300 effects are testable with one third of the full sample. A total of 780 effects are testable using the 3-form design, representing a 79.3% increase over the 1-form design.

size. The top part of the bar (3-form design) represents the number of effects testable with the smallest sample size. Under the conditions of our example (30 questions asked of each subject; 40 questions asked altogether), the 3-form design allows testing of 780 effects (correlations), whereas the comparable 1-form design (complete cases design that asks just 30 questions) allows testing of only 435 effects. However, as shown in Figure 1, with the 1-form design, all 435 of the effects are tested with the full sample size (N = 300 in the figure). With the 3-form design, 45 effects are tested with the full sample (e.g., N = 300), 435 effects are tested with two thirds the full sample (e.g., N = 200), and 300 effects are tested with one third the full sample (e.g., N = 100).[1]

Because some effects (correlations) are tested with less than the full sample size, these effects will, of course, be tested with less power. Table 5 presents power information for the one- and 3-form designs for a study with N = 300 and effect sizes ranging from ρ = .05 to ρ = .30. To put these effect sizes in context, Cohen (1977) referred to ρ = .10, ρ = .30, and ρ = .50 as *small*, *medium*, and *large* effects, respectively. The trade-off with the 3-form design is that more effects can be tested, but many of them are tested with smaller sample sizes. The effects listed across the top of Table 5 represent classes or categories of effects, where XX refers to the relationship between two variables from the X set, XB refers to the effect of a variable from the X set on a variable from the B set, and so on.

Effects in Table 5 with an ✗ mark indicate that the effect is tested with good power (.80 or greater) in the 1-form design, but with power < .80 in the 3-form design. Note that one can calculate the ratio of the power for rejecting the false null hypothesis for the two designs by dividing the power for the 1-form design by the power for the 3-form design (the power ratio). These calculations are provided in Table 5 for the two most relevant columns. Effects in the last two columns of Table 5 (those in effect categories XC, CC and AC, BC) are not testable at all with the 1-form design. Effects in these columns marked with a check (✓) are effects that can be tested with good power (.80 or greater) in the 3-form design. Because these effects are not testable at all with the 1-form design, they represent an advantage of the 3-form design.

Effects tested with the full N = 300 are tested with good power when effect sizes are $\rho \geq .16$. Effects tested with N = 200 are tested with good power when effect sizes are $\rho \geq .20$. Effects tested with N = 100 are tested with good power when effect sizes are $\rho \geq .28$.

✗ and ✓ results for a study with N = 300. For a study with N = 300, there are four effects with the ✗ mark in the XA, XB, AA, BB category, those for effect sizes ρ = .16, ρ = .17, ρ = .18, and ρ = .19 (see Table 5). For this study sample size, there are 12 such effects in the AB category, for effect sizes from ρ = .16 to ρ = .27. For the XC, CC effect category, all effects of $\rho \geq .20$ are tested with good power with the 3-form design. For the AC, BC effect category, all effects $\rho \geq .28$ are tested with good power with the 3-form design. None of the effects in these two categories are testable using the 1-form design.

✗ and ✓ results for studies with N = 1,000 and N = 3,000. Detailed power calculations, such as those shown in Table 5, for study sample sizes N = 1,000 and N = 3,000 are available in the online supplemental data for this article. For a study with N = 1,000, there are two effects with the ✓ mark in the XA, XB, AA, BB category, those for effect sizes ρ = .09 and ρ = .10. For this study sample size, there are six such effects in the AB category, for effect sizes from ρ = .09 to ρ = .14. For the XC, CC effect category, all effects of $\rho \geq .11$ are tested with good power with the 3-form design. For the AC, BC effect category, all effects $\rho \geq .15$ are tested with good power with the 3-form design. None of the effects in these two categories are testable using the 1-form design.

For a study with N = 3,000, there is only one effect with the ✗ mark in the XA, XB, AA, BB category, that for effect size ρ = .05. For this study sample size, there are four such effects in the AB category, for effect sizes from ρ = .05 to ρ = .08. For the XC, CC effect category, all effects of $\rho \geq .06$ are tested with good power with the 3-form design. For the AC, BC effect category, all effects $\rho \geq .09$ are tested with good power with the 3-form design. None of the effects in these two categories are testable using the 1-form design.

Power Reconsidered: Discounting the Apparent Power Disadvantage of the 3-Form Design

The ✗ and ✓ effects shown in Table 5 do help determine which design one should use. However, some researchers will take a conservative stance here. If there are too many ✗ effects with the 3-form design, some may prefer the 1-form design despite the number of ✓ effects. That is, some researchers will pay too much attention to ✗ effects shown in Table 5. Thus, although the information provided in Table 5 is useful, we argue that it is not the whole story. To counter the possible conservative bias, we present three arguments as to why the potentially bad outcomes with the 3-form design (i.e., lower power indicated by the ✗) are not so bad after all.

First notice that the badness of these ✗ effects (i.e., lost power) is rather variable. We employ the concept of the power ratio to help quantify the difference in power be-

[1] The number of variables per item set will, of course, vary from study to study. However, the ratios of testable effects shown in Figure 1 remain approximately the same as what is shown, as long as there are equal numbers of items in the four item sets.

Table 5
Statistical Power for 1- and 3-Form Designs by Effect Size for a Study With N = 300

Effect size (ρ)	Effects										
	XX		XA, XB, AA, BB			AB			XC, CC		AC, BC	
	1-form (*N* = 300)	3-form (*N* = 300)	1-form (*N* = 300)	3-form (*N* = 200)	Power ratio	1-form (*N* = 300)	3-form (*N* = 100)	Power ratio	1-form (*N* = 0)	3-form (*N* = 200)	1-form (*N* = 0)	3-form (*N* = 100)
.05	.14	.14	.14	.11	1.27	.14	.08	1.75	0	.11	0	.08
.06	.18	.18	.18	.14	1.29	.18	.09	2.00	0	.14	0	.09
.07	.23	.23	.23	.17	1.35	.23	.11	2.09	0	.17	0	.11
.08	.28	.28	.28	.20	1.40	.28	.13	2.15	0	.20	0	.13
.09	.35	.35	.35	.25	1.40	.35	.15	2.33	0	.25	0	.15
.10	.41	.41	.41	.29	1.41	.41	.17	2.41	0	.29	0	.17
.11	.48	.48	.48	.34	1.41	.48	.20	2.40	0	.34	0	.20
.12	.55	.55	.55	.40	1.38	.55	.22	2.50	0	.40	0	.22
.13	.62	.62	.62	.45	1.38	.62	.26	2.38	0	.45	0	.26
.14	.68	.68	.68	.51	1.33	.68	.29	2.34	0	.51	0	.29
.15	.74	.74	.74	.57	1.30	.74	.32	2.31	0	.57	0	.32
.16	.80	.80	.80	.62✗	1.29	.80	.36✗	2.22	0	.62	0	.36
.17	.84	.84	.84	.68✗	1.24	.84	.40✗	2.10	0	.68	0	.40
.18	.88	.88	.88	.73✗	1.21	.88	.44✗	2.00	0	.73	0	.44
.19	.91	.91	.91	.77✗	1.18	.91	.48✗	1.90	0	.77	0	.48
.20	.94	.94	.94	.81	1.16	.94	.52✗	1.81	0	.81✓	0	.52
.21	.96	.96	.96	.85	1.13	.96	.56✗	1.71	0	.85✓	0	.56
.22	.97	.97	.97	.88	1.10	.97	.60✗	1.62	0	.88✓	0	.60
.23	.98	.98	.98	.91	1.08	.98	.64✗	1.53	0	.91✓	0	.64
.24	.99	.99	.99	.93	1.06	.99	.68✗	1.46	0	.93✓	0	.68
.25	.99	.99	.99	.95	1.04	.99	.72✗	1.38	0	.95✓	0	.72
.26	—	—	—	.96	1.04	—	.75✗	1.33	0	.96✓	0	.75
.27	—	—	—	.97	1.03	—	.78✗	1.28	0	.97✓	0	.78
.28	—	—	—	.98	1.02	—	.81	1.23	0	.98✓	0	.81✓
.29	—	—	—	.99	1.01	—	.84	1.19	0	.99✓	0	.84✓
.30	—	—	—	.99	1.01	—	.86	1.16	0	.99✓	0	.86✓

Note. Dashes indicate power > .995. Power ratio = power for 1-form design divided by power for 3-form design; XX = correlation for two items from X set; XA = correlation for one item from X set, one item from A set; AA = correlation for two items from A set; and so on (see Table 2). ✗ is used to highlight effects for which power ≥ .80 for the 1-form design and power < .80 for the 3-form design. ✓ is used to highlight effects for which power ≥ .80 for the 3-form design and power = 0 for the 1-form design.

tween the two designs. With this ratio, it is easy to show that (a) for XX effects, there is no power differential between the designs; (b) for XA, XB, AA, BB effects, the power differential is small—the 1-form design averages only 23% more power than the 3-form design for effects denoted ✗ in Table 5 (range = 18%–29%)—and (c) for AB effects, the power differential is more substantial—the 1-form design averages 69% more power than the 3-form design for effects designated ✗ in Table 5 (range = 28%–122%). On the basis of these numbers, we argue that the true disadvantage of the 3-form design is found for just one kind of effect: AB correlations (one item from the A set and one from the B set).

Second, note that these AB effects represent a relatively small portion of the total number of effects. For starters, AB effects represent just 23% of the total number of effects (435) in the 1-form design. But also consider the fact that the various effect sizes in empirical data are not uniformly distributed. For example, in one wave of one cohort of the Adolescent Alcohol Prevention Trial study (Hansen & Graham, 1991), we found that the 1,540 correlations based on 56 scales had a mean of .194 (SD = .139; Mdn = .17) and were distributed with slightly positive skew (1.01) and kurtosis (0.89). On the basis of these data, 31% of the AB effects in the Adolescent Alcohol Prevention Trial study had effect sizes in the r = .16–.27 range (those with ✗ marks in Table 5).

Multiplying these two percentages together (23% of the total correlations are in the AB class of effects; 31.3% of these are expected to be in the problem effect size range), we see that only 7.2% (31.3% × 23%) of the total number of effects are expected to be a problem. With larger study sample sizes, the problem is even smaller. For a study with N = 1,000, only 4.3% of the total number of effects are expected to be a problem. For a study with N = 3,000, only 2.9% of the total number of effects are expected to be a problem (the percentage of problem effects is smaller with larger Ns because the range of AB effects with ✗ marks is smaller in such studies).

Third, let us look more closely at the AB category of effects. Most important, the researcher controls to a large extent which effects fall in this category. Let us suppose that the researcher identifies 50 correlations as being most important to the study goals. If only 7.2% of the study correlations are expected to represent a power disadvantage, this means that by chance alone, only 50% × 7.2% = 3.6% of these most important effects are expected to have a power disadvantage with the 3-form design. However, placement of items (and thus correlations) into the X, A, B, and C item sets is not a random process. In fact, if there are 50 correlations defined in advance as most important, the researcher can virtually guarantee that all 50 will be tested with adequate power.

Researcher Control Over Trade-Offs

There certainly are trade-offs with the 3-form design, as described above. Fortunately, these trade-offs are largely under the researcher's control. If a particular regression hypothesis is central to the study, then it makes sense to place the variables relating to the hypothesis into item sets that will be tested with at least the intermediate sample size. Even smallish effect size hypotheses are generally tested with good power under these circumstances. If the hypothesis in question is expected to have a very small effect size, it may even be advisable to place all the relevant variables in the X set, where the full sample will be available for the hypothesis test.

We suggest, for example, that variables associated with the most important effects (correlations) under study be placed in the X set. This is especially true when these most important correlations may have somewhat low effect sizes. At the very least, researchers should place both variables for important correlations within one item set (i.e., both variables within the X, A, B, or C item sets) or one variable in the X set and one in the A, B, or C item set. With these strategies, the correlation will appear in the categories XA, XB, AA, BB or XC, CC in Table 5. Further, we recommend that researchers place all items for scales together in the same item set (X, A, B, or C), perhaps interspersed throughout that item set. Finally, we recommend that to the extent possible, researchers place into the same item set (X, A, B, or C) sets of scales that tend to be examined together.

Power Considerations With Statistical Interactions

The situation with interactions is complicated in several ways, but what we have said above relating to regular two-variable correlations also applies in large part to the correlations between the product of two variables and a third variable. In general, the number of such correlations possible is $[k\,(k-1)/2](k-2)$. If we assume k = 10 variables (e.g., scales) in each item set, then there are 30 items total in the X, A, and B item sets, and 12,180 possible two-way interaction effects involving these three item sets. All of these interactions are testable with the complete cases design and with the 3-form design. With the 3-form design, more than half of these effects are tested with at least two thirds of the full sample size. As we have argued above, when correlations are tested with at least two thirds of the full sample size, the difference in power between the one- and 3-form designs is not large. An additional 14,460 possible two-way interaction effects involving variables from the C item set are testable with the 3-form design, but not with the complete cases design.

Efficiency Designs for Measurement: II. Two-Method Measurement

Our *two-method measurement* design involves two kinds of measures: relatively cheap, noisy (e.g., less valid) measures of a construct and expensive, more valid measures of the same construct. It is key that the expensive and cheap measures can be thought of, and modeled, as two or more measures of a single construct. It is generally accepted that it is desirable to have multiple measures of a construct (e.g., Bollen, 1989), for example, because one is better able to model the error structure of such measures (e.g., Kaplan, 1988). We suggest that these multiple measures can be selected to have other desirable properties as well.

A major benefit of two-method measurement is that one can often use a combination of the cheap and expensive measures to get more power than with the expensive measures alone and more construct validity than with the cheap measures alone. What makes this possible is that the valid (expensive) measures may be used to help model the response bias associated with the cheap measures. It is important to note that this benefit can be achieved even when a relatively small proportion of the cases have data for the expensive measure. The benefit of the two-method measurement design can also be framed in terms of its favorable cost-to-benefit ratio, especially as it relates to statistical power to test hypotheses involving the constructs of interest.

Definition of Response Bias

The model we use can be thought of as a *response bias correction* model. But what do we mean by *response bias*? We conceive of response bias as being the opposite of construct validity. A valid measure measures what it was intended to measure (e.g., Cook & Campbell, 1979). To the degree that a measure actually measures something different from what we intended, the measure is not valid; that is, it is biased.

This kind of bias can be thought of as the reason why two measures are more highly correlated than can be explained by the fact that the measures are both indicators of the same common factor. This view of response bias is similar to what Hoyt (2000) described as *halo errors*. It is also similar to what Berman and Kenny (1976; see also Graham & Collins, 1991) referred to as *correlational bias*.

The Bias Correction Model

Before introducing the missing data part of our approach, it is important to be clear about the statistical model that will allow the planned missing data approach to work. Given our definition of bias, we can conceive of a statistical model designed to correct for this bias in substantive regression analyses. The model we use, which is shown in Figure 2, has some features in common with the models described by Kenny and others in the multitrait–multimethod context (Eid, 2000; Graham & Collins, 1991; Kenny, 1976; Kenny

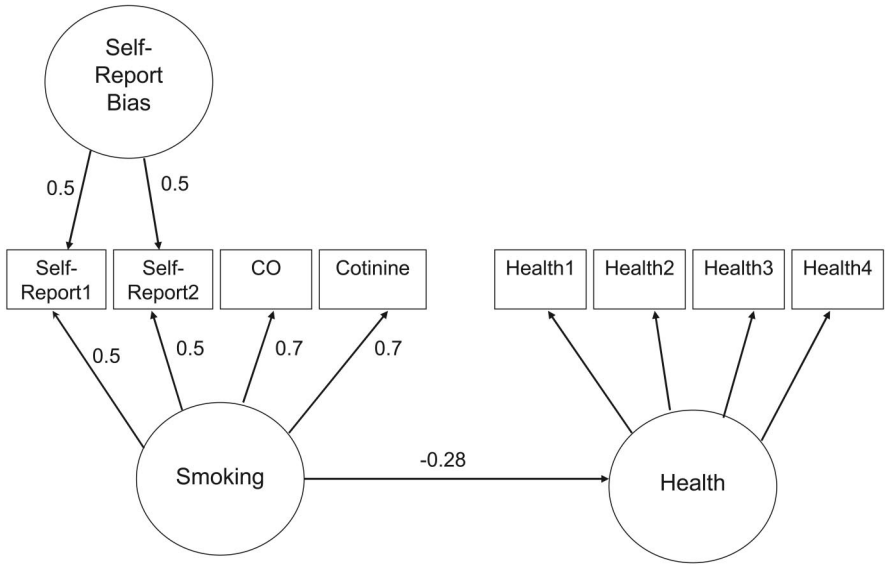

Figure 2. Three-factor structural equation model examining the effect of smoking on general health. Smoking represents the four-item common factor comprising two cheap measures (self-report) and two expensive measures (carbon monoxide [CO] and cotinine). The two self-report measures are also specified to load on the Self-Report Bias factor. Health represents the four-item latent outcome variable (with four generic indicators of health). The latent variable Smoking is specified to predict the outcome variable Health. Population parameter values are shown for Cell A of Simulation 1. In order to reduce complexity, we excluded residual error variances from the figure.

& Kashy, 1992; Marsh, 1989; Widaman, 1985). Our model is an SEM model following Palmer, Graham, Taylor, and Tatterson (2002; see also Graham & Collins, 1991; Hawkins et al., 1997). The idea of our model is to allow for two sources of correlation between the two cheap measures. One source is the common factor, or latent variable, for which the two (cheap) measures are manifest indicators; the other source is the response bias. This model could actually include a latent variable representing response bias (e.g., Graham & Collins, 1991; Palmer et al., 2002), as depicted in Figure 2, or equivalently, the model could simply involve estimating the residual covariance between the two cheap measures (e.g., Kenny, 1976; Kenny & Kashy, 1992; Marsh, 1989).

Applications of Two-Method Measurement: Cigarette Smoking Research

Examples of uses of the two-method measurement design come from several research domains. In the present article, we focus on cigarette smoking research. In the study of cigarette smoking, researchers have supposed that respondents in some (e.g., adolescent) populations are reluctant to tell the truth about their cigarette smoking. Thus self-report measures of smoking have been suspect, as are important correlations with cigarette smoking based on self-reports. In response to this problem, researchers have often collected data on biochemical validators, such as carbon monoxide (CO), saliva thiocyanate, or saliva cotinine (Biglan, Gallison, Ary, & Thompson, 1985; Heatherton, Kozlowski, Frecker, Rickert, & Robinson, 1989; Kozlowski & Herling, 1988; Pechacek, Murray, Luepker, Mittelmark, & Johnson, 1984; Taylor et al., 1998). The model shown in Figure 2 is designed to correct these problems and produce more realistic correlations (and regression coefficients) involving cigarette smoking. For the model shown in Figure 2, all of the smoking relevant variables are specified to load on the latent variable Smoking. In addition, the two self-report variables are specified to load on the latent variable Self-Report Bias. The latent variable Smoking is specified to predict the latent variable Health. In this example, CO and cotinine are the expensive, but valid, measures of cigarette smoking, and self-reports of cigarette use are the cheap measures.

Some Other Applications

Numerous other applications of the two-method measurement design are possible. For example, research related to nutrition and exercise often involves measurement of adiposity. In this context, the cheap (but flawed) measure commonly used is body mass index (e.g., see Davis & Gergen, 1994; Goodman, Hinden, & Khandelwal, 2000). More valid, but much more expensive, measures of adiposity include densitometry, or hydrostatic weighing (e.g., Going, 1996), and dual energy X-ray absorptiometry (e.g., Avesani et al., 2004; Fisher, Johnson, Lindquist, Birch, & Goran, 2000; Lohman, 1996; Roubenoff, Kehayias, Dawson-Hughes, & Heymsfield, 1993).

In the area of nutrition, researchers frequently collect data from many subjects on a relatively inexpensive (but flawed) short self-administered nutrition survey. A more intensive (and much more expensive) face-to-face nutrition interview is then administered to a smaller number of subjects via a trained interviewer (Williet, 1990). The extensive nutrition survey incorporates visual cues and prompting by the interviewer, yielding much more reliable and valid nutrition data. This idea also applies to survey research more generally. It may be reasonable to combine a self-administered survey, which could be given inexpensively to a large number of subjects, with a more in-depth, face-to-face interview, which could be administered to a subset of the subjects.

Artificial Data Illustrations for Two-Method Measurement

We now illustrate the benefits of our two-method measurement model with three brief simulations (note that our simulations involve artificial data but are not the kind of simulations often called *Monte Carlo* studies). In Simulation 1, we varied the valid reliability and validity of the cheap measures and the valid reliability of the expensive measures (we define *valid reliability* as the valid component of reliable variance; we distinguish this from reliable variance that may be attributable to response bias or another construct). In Simulation 2, we varied the cost differential between the cheap and expensive measures. In Simulation 3, we varied the effect size of the major latent variable regression coefficient of interest.

Simulation 1

Our artificial data illustrations involve a latent variable model with two substantive latent variables, Smoking and Health (along with the Self-Report Bias latent variable). This model is depicted in Figure 2, along with key population parameter values for one cell (Cell A) of the artificial data example. The simulated Health latent variable was measured with four manifest variables. The simulated Smoking latent variable (also referred to below as the *valid common factor*) was also measured with four manifest variables, two items from a hypothetical self-administered smoking survey (administered to all subjects) and two hypothetical biochemical items (e.g., CO from a breath sample and cotinine from a saliva sample) administered to a random sample of the same subjects.

For Simulation 1, the population correlation between the smoking and health latent variables was $\rho = -.40$ ($\beta = -.28$). For illustration, we assumed that \$15,050 was avail-

able for measurement of the smoking measures. On the basis of the actual data collection costs of one study (Taylor et al., 1998), we estimated collection of self-report measures to be $7.30 per subject and collection and analysis of the breath and saliva samples to be an additional $16.78 per subject.

With these assumed costs, it is possible to collect complete self-report and biochemical data for $N = 625$ people. For exactly the same costs ($15,050), it is also possible to collect more self-report data and less biochemical data. For illustrative purposes, we increased the sample size for self-report measures in $N = 100$ increments and reduced the number of biochemical measures to keep the overall cost constant (for all five cells in Simulation 1, with an increase of 100 self-reports, 43.5 fewer biochemical measures must be collected in order to keep the overall cost constant). For each sample size combination, we simulated data under five conditions. In each case, we retained the idea that the expensive (e.g., biochemical) measures were better than the cheaper self-reports, in that they were more valid and/or more reliable. More precisely, we examined several cases in which we varied (a) the valid reliability (i.e., the factor loadings on the hypothetical smoking factor) of the two kinds of measure and (b) whether there was systematic self-report bias (i.e., reliable variance on the self-report bias construct).

Simulation Cell A (self-reports biased; expensive measures had higher valid reliability). For this cell, self-reports contained some response bias, which was represented in the population by the nonzero (.50) factor loadings on the Self-Report Bias factor. This factor was specified to be uncorrelated with other factors in the model. This level of bias has the effect of doubling the observed correlation between self-report variables in this cell from $r = .25$ to $r = .50$. The expensive measures also had more valid reliability: The factor loadings on the Smoking factor were .50 for the self-report items and .70 for the biochemical measures. Key population values for this cell of the simulation appear in Figure 2. The annotated LISREL code that generated the population covariance matrix for this (and all) cells can be found in the online supplemental data for this article. In Cell A, the valid common factor and bias factor each accounted for 25% of the variance in the self-reports.

Simulation Cell B (self-reports biased; expensive and cheap measures had equally high valid reliability). For this cell, self-report bias was created as described for Cell A. However, in this case, factor loadings on the Smoking factor were reasonably high (.70) for both types of measures. In this cell, the valid common factor accounted for 49%, and the response bias factor accounted for 25% of the variance in the self-reports.

Simulation Cell C (self-reports biased; expensive measures had lower valid reliability). Self-report bias was created as described for Cell A, and the expensive measures had less valid reliability: The factor loadings on the Smoking factor were .70 for the self-report items and .50 for the biochemical (expensive) measures. For this cell, the valid common factor accounted for 49%, and the response bias factor accounted for 25% of the variance in the self-reports.

Simulation Cell D (self-reports biased; expensive and cheap measures had equally low valid reliability). Self-report bias was created as described for Cell A. In this case, factor loadings on the Smoking factor were a bit low (.50) for both types of measures. For this cell, the valid common factor and bias factor each accounted for 25% of the variance in the self-reports.

Simulation Cell E (self-reports not biased; expensive measures had more valid reliability). For this cell, there was no Self-Report Bias factor; the self-reports loaded only on the Smoking factor. In this case, factor loadings on the Smoking factor were .70 for biochemical items and .50 for the self-reports.

Main dependent variable. The variable of primary interest in Simulation 1 was the standard error for the regression coefficient between the two factors (note that there is no parameter estimation bias with the missing data procedures used). This regression coefficient would carry the main substantive interest for the test of the hypothesis of whether smoking had an impact on later health. The standard error for this regression coefficient is inversely related to the power for testing the regression coefficient.

The standard errors reported in Figure 3 were simulated in the sense that the population covariance matrix for the tested model was analyzed, as if it contained real data, under the missing data circumstances to be described. The complete cases model in each cell (A, B, C, D, E) was a straightforward one-group model analyzed using LISREL 8.5 (Jöreskog & Sörbom, 1996; the annotated LISREL code for this model may be found in the online supplemental data for this article). For each cell, this model was tested using the population covariance matrix as input, with *N* set to 625 (i.e., "no = 625" in the DA [Data] statement). For Cells A, B, C, and D, the residual covariance was estimated between the two simulated self-report items to model the bias contained in these cells. The models for Cell E were the same as those in the other cells except that no residual covariances were estimated (and no bias was present in the population; note, however, that exactly the same results were obtained for Cell E whether this residual covariance was estimated or not).

We tested each of the remaining models using the multiple group capabilities of LISREL 8.5, using the missing data strategy described by P. D. Allison (1987) and B. Muthén et al. (1987; the LISREL code for this model may also be found in the online supplemental data for this article). This strategy has been used in the past for examining the effects of missing data on parameter estimation (Graham et al., 2001).

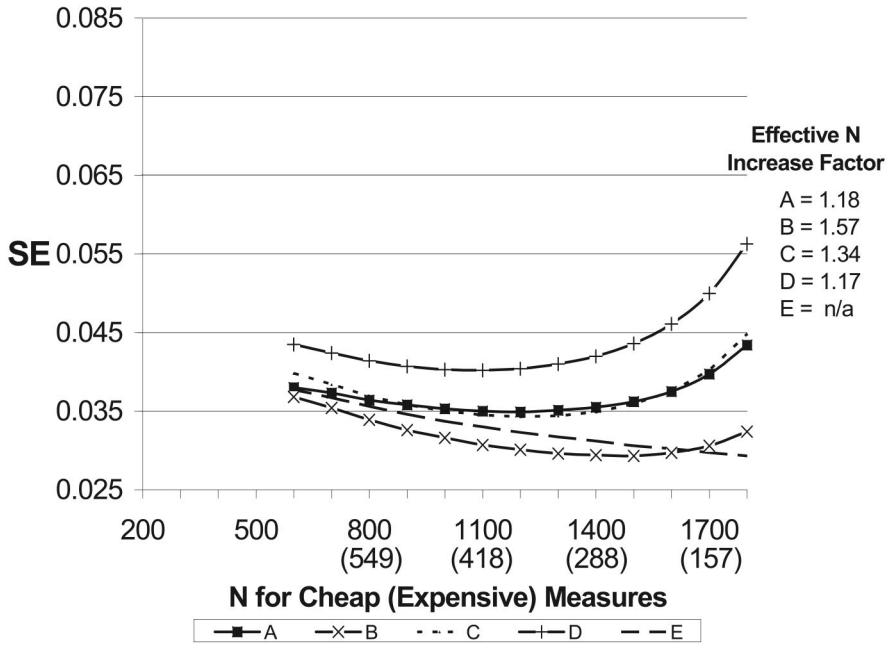

Figure 3. Simulation 1 standard errors for the regression coefficient of the Smoking factor predicting the Health factor. *N*s displayed along the *x*-axis refer to number of study subjects presented with the cheaper, less valid measures; corresponding *N*s for subjects also presented with the more expensive, more valid measures (e.g., biochemical measures) are shown in parentheses. All standard errors were based on analyses of the population covariance matrix under the various missing data conditions of Simulation 1. Effective *N* Increase Factors were calculated by dividing the effective *N* by the complete cases *N*. Curves A–E correspond to Simulation Cells A–E. The valid reliability and validity of the cheap measures and the valid reliability of the expensive measures were varied as follows. Cell A = factor loadings of .50, .70 (cheap, expensive) + self-report bias; Cell B = factor loadings of .70, .70 (cheap, expensive) + self-report bias; Cell C = factor loadings of .70, .50 (cheap, expensive) + self-report bias; Cell D = factor loadings of .50, .50 (cheap, expensive) + self-report bias; and Cell E = factor loadings of .50, .70 (cheap, expensive) + no bias.

Normally one would analyze missing data, including data from the two-method measurement design, using multiple imputation (Rubin, 1987; Schafer, 1997) or one of the SEM packages with a FIML feature for handling missing data. However, with simulation work, the advantage of using the multiple-group SEM procedure over using the more standard SEM–FIML approach, is that the population values can be analyzed directly. With FIML, one must input raw data (including missing and nonmissing values). Such data sets, even when simulated, cannot conform to the population parameters except on average. Making use of the multiple-group strategy, one is able to simulate the standard errors using just one (two-group) analysis. Simulating the standard errors using FIML procedures would require tens or even hundreds of thousands of replications for the entire simulation.

For the two-group models, the overall sample size ranged from 700 to 1,800 in increments of 100. For example, when the overall $N = 700$, 592 cases had complete data, and another 108 cases had self-report data but no data for either of the expensive biochemical measures. The 592 cases with complete data were modeled as Group 1 and were handled in the usual way for any SEM model. The input matrix for this group was the population covariance matrix.

The 108 cases with partial data were set up as follows. The input covariance matrix was the same as the population matrix, except that all covariances involving the two missing variables were set to 0. The variances for the two missing variables were set to 1 (see P. D. Allison, 1987; Graham et al., 1994). All factor level parameters in the model (factor variances, covariances, regressions) were estimated and constrained to be equal across the two groups. All item-level parameter estimates (factor loadings, residuals, residual covariances), where data were available, were also estimated and constrained to be equal across the two groups. Factor loadings involving the two missing variables were estimated normally in Group 1 and fixed at 0 in Group 2. Item residual variances involving the two missing variables were estimated normally in Group 1 and fixed at 1 in Group 2. This procedure yields maximum likelihood param-

eter estimates in the missing data case (P. D. Allison, 1987; Graham et al., 1994; B. Muthén et al., 1987).

Results for Simulation 1

One of the most striking results shown in Figure 3 is that for all five conditions of the simulation, the best measurement design (i.e., the design yielding the lowest standard error) was a missing data design. In each case, the standard errors for at least eight of the planned missing data designs were lower than the standard error for the corresponding complete cases design costing exactly the same.

At first blush these results may seem counterintuitive. But consider the following: For the cost scenario presented in Figure 3, the sample size for the self-report measures was increased in increments of $N = 100$. In order to maintain the same overall costs, for every 100 cases of self-reports added, we must collect 43.5 fewer biochemical measures. The degree to which the standard error gets smaller because of the increase of $N = 100$ self-reports is greater than the degree to which the standard error gets larger because of the decrease of $N = 43.5$ biochemical measures.

A second result that emerges from Figure 3 is that not all planned missing data designs yielded better (lower) standard errors. For the first four cells of the artificial data illustration (Cells A, B, C, and D), as the number of planned missing values increased, the standard errors went down to some minimum and then increased again, eventually becoming larger than the complete cases standard error. Although it is not shown in Figure 3, the standard error approaches infinity as the number of expensive measures approaches zero. Modeling the response bias becomes less and less efficient as the complete cases sample size gets small; in the extreme, when there are no expensive measures, the bias is not estimable because of algebraic underidentification.

A third result of Simulation 1 is that the curves for Cells B and C are relatively deeper than the curves for Cells A and D (see Figure 3). The most important difference between these two sets of cells was the (valid) reliability of the self-report measures (.70 in Cells B and C; .50 in the other cells). One conclusion is that the planned missing data design is more obviously better than the complete cases design of the same cost when the (valid) reliability of the self-report (or inexpensive) measures is relatively high.

Depth of the curves, percentage of decrease in standard error. The difference in the depth of the curves is captured nicely in Figure 3. However, more important than the depth, per se, is the degree to which the key standard error is reduced from the complete cases design to the best missing data design. For example, for Cell A, the standard error for the best missing data design was 8.2% lower than the corresponding complete cases design. For Cell B, the standard error for the best missing data design was 20% lower than the corresponding complete cases design.

Effective N Increase Factor. A more intuitive way of describing the depth of these curves is to find the complete cases *N* that produces the same standard error found in the best missing data design. For Cell A, complete cases $N = 740$ yields the same standard error found in the best missing data design. We refer to this new *N* as the *effective N*. That is, this missing data design behaves, with respect to the key standard error and statistical power, as if the sample size were $N = 740$. Further, we find it useful to calculate the *Effective N Increase Factor*, which is simply the effective *N* divided by the nominal complete cases *N* (625 in our example). The Effective N Increase Factors for the first four cells of Simulation 1 appear in Figure 3. This factor was 1.18 for Cell A and 1.57 for Cell B (note that this factor is not defined for Cell E because there was no clear minimum standard error for this cell).

A fourth interesting finding from Simulation 1 is that the curve for Cell E was monotonically decreasing. This occurred because in Cell E, the cheap measure did not contain bias in the population and bias was not modeled (although the same result occurred whether the bias was modeled or not). Increases in the sample size for the self-report measures yielded the customary decreases in the standard error, and because there was no bias, there was no corresponding decrease in the efficiency of bias estimation as the sample size for the expensive measures decreased.

Finally, comparing Cells A and D in Figure 3 shows that the effect of lowering the (valid) reliability of the expensive measures changes the overall level of the standard error (and statistical power); all of the standard errors for Cell D were higher than those for Cell A. However, other aspects of the two sets of standard errors were essentially the same, implying that changes in (valid) reliability of the expensive measure (sometimes missing) had minimal impact on the value of using a missing data design. Comparing Cells B and C leads us to the same conclusion.

Simulation 2: Less Extreme, More Extreme Cost Differential

The simulation results depicted in Figure 3 were based on actual research costs from one study. However, studies will vary considerably in the cost differential between cheap and expensive measures. Simulation 2 expands on the results of the first simulation by examining the effects of less and more extreme cost differential between these measures. These results are presented in Figure 4. The results shown in Figure 4 pertain to Cell A circumstances only (expensive measure better in the sense of having better valid reliability and less bias).

For all of the scenarios shown in Figure 4, the cheap measure was \$7.30 per person. The per-person cost for the

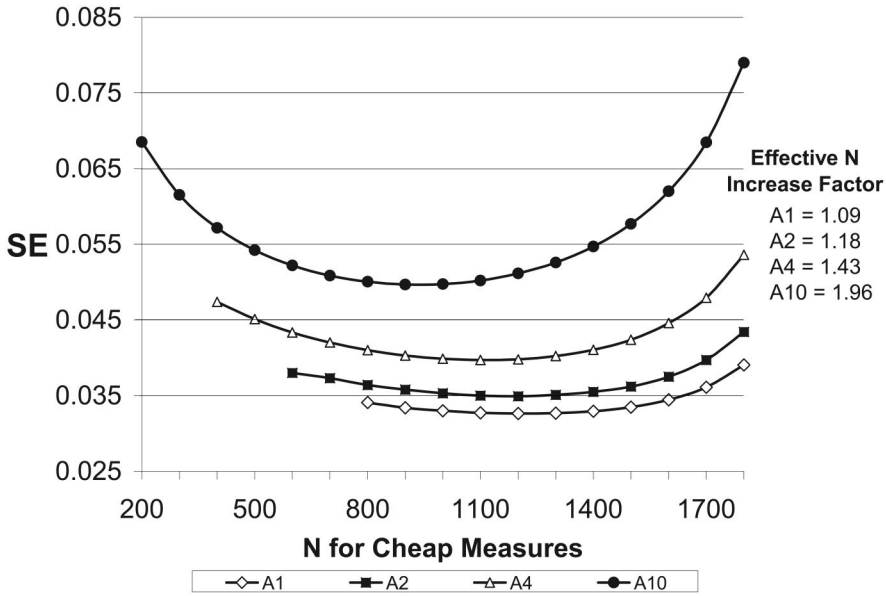

Figure 4. Simulation 2 standard errors for the regression coefficient of the Smoking factor predicting the Health factor. *N*s displayed along the *x*-axis refer to number of study subjects presented with the cheaper, less valid measures. Corresponding *N*s for those also receiving the expensive measures are shown as follows (*N* for cheap measures/*N* for expensive measures): Cell A1 = (1,100/583), (1,400/401), (1,700/219); Cell A2 = (800/549), (1,100/418), (1,400/288), (1,700/157); Cell A4 = (500/380), (800/307), (1,100/234), (1,400/161), (1,700/88); and Cell A10 = (500/156), (800/126), (1,100/96), (1,400/66), (1,700/36). All configurations cost approximately $15,050 in total. All standard errors were based on analyses of the population covariance matrix under the various missing data conditions of Simulation 2. Effective *N* Increase Factors were calculated by dividing the effective *N* by the complete cases *N*. All Simulation 2 analyses are based on Cell A conditions from Simulation 1 (factor loadings = .50, .70 [cheap, expensive] + self-report bias). The per-person cost differential between cheap and expensive measures was varied as follows. Cell A1 = expensive ($12.04), cheap ($7.30), 1.6:1 cost ratio; Cell A2 = expensive ($16.78), cheap ($7.30), 2.3:1 cost ratio; Cell A4 = expensive ($30.00), cheap ($7.30), 4.1:1 cost ratio; and Cell A10 = expensive ($73.00), cheap ($7.30), 10:1 cost ratio.

expensive measures varied for the four curves. Each curve is labeled with the approximate ratio between the expensive and cheap measures. The bottom curve (A1) shows results based on $12.04 per-person for the expensive measure (a 1.65:1 cost ratio). The second curve from the bottom (A2) is the same as Cell A in Figure 3 ($16.78 per person; a 2.3:1 cost ratio). The second curve from the top (A4) shows results based on somewhat higher expensive measure costs ($30.00 per person; a 4.1:1 cost ratio). The top curve in Figure 4 (A10) shows results based on even wider, but still realistic, cost differential ($73.00 per person; a 10:1 cost ratio).

For Simulation 2, the total cost for measuring the smoking variables remained approximately $15,050 (this total varied slightly for different cost ratios). Thus, the number of complete cases, cases with cheap measures, and cases with expensive measures all varied depending on the cost ratio. For Cell A1, $N = 778$ complete cases cost $15,047, overall. For each increase of 100 cheap measures there must be 60.6 fewer expensive measures in order to keep overall costs the same. Thus, for $N = 1{,}100$, $N = 1{,}400$, and $N = 1{,}700$ cheap measures, there were, respectively, $N = 583$, $N = 401$, and $N = 219$ expensive measures. For Cell A2, $N = 625$ complete cases cost $15,050, overall (as described in Simulation 1). For each increase of 100 cheap measures, there must be 43.5 fewer expensive measures in order to keep overall costs the same. Thus, for $N = 800$, $N = 1{,}100$, $N = 1{,}400$, and $N = 1{,}700$ cheap measures, there were, respectively, $N = 549$, $N = 418$, $N = 288$, and $N = 157$ expensive measures. For Cell A4, $N = 403$ complete cases cost $15,032, overall. For each increase of 100 cheap measures there must be 24.33 fewer expensive measures in order to keep overall costs the same. Thus, for $N = 500$, $N = 800$, $N = 1{,}100$, $N = 1{,}400$, and $N = 1{,}700$ cheap measures, there were, respectively, $N = 380$, $N = 307$, $N = 234$, $N = 161$, and $N = 88$ expensive measures. Finally, for Cell A10, $N = 187$ complete cases cost $15,016, overall. For each in-

crease of 100 cheap measures there must be 10 fewer expensive measures in order to keep overall costs the same. Thus, for $N = 500$, $N = 800$, $N = 1{,}100$, $N = 1{,}400$, and $N = 1{,}700$ cheap measures, there were, respectively, $N = 156$, $N = 126$, $N = 96$, $N = 66$, and $N = 36$ expensive measures.

The results shown in Figure 4 have the same basic pattern as before. Even with the less extreme cost estimates shown in the bottom curve (A1), the complete cases design was not the best alternative. The effective *N* for curve A1 was 9% higher than the nominal complete cases *N*. When the cost differential was more extreme (Curve A4), the effective *N* was 43% higher than the nominal complete cases *N*. For the even wider cost differential shown in Curve A10, the effective *N* was nearly double the nominal complete cases *N*.

Simulation 3: Different Effect Sizes

Simulation 3 examined two cells from Simulation 2: Cell A2 (Reliability Scenario A with a 2.3:1 cost ratio) and Cell A10 (Reliability Scenario A with a 10:1 cost ratio). We examined the effect size used in Simulations 1 and 2 ($\rho = .40$ between the two latent variables smoking and health) and a smaller effect size ($\rho = .10$; a small effect size in terms of Cohen, 1977). The results of Simulation 3 appear in Figure 5. The curves $A2_{(.40)}$ and $A10_{(.40)}$ are the same as the A2 and A10 curves from Figure 4. Curve $A2_{(.10)}$ is the same cost scenario as Curve $A2_{(.40)}$ (2.3:1 cost ratio), but with the lower effect size. Curve $A10_{(.10)}$ is the same cost scenario as Curve $A10_{(.40)}$ (10:1 cost ratio), but with the lower effect size. The sample sizes for complete cases and for cheap and expensive measures for Cells $A2_{(.40)}$ and $A2_{(.10)}$ were the same as those shown in Simulation 2 for Cell A2. The sample sizes for Cells $A10_{(.40)}$ and $A10_{(.10)}$ were the same as those shown in Simulation 2 for Cell A10.

The main difference between the two effect sizes is that the curves for the smaller effect size are rather deeper than those for the higher effect size. For Cell $A2_{(.10)}$, the effective *N* was 35% higher than the nominal complete cases *N*.

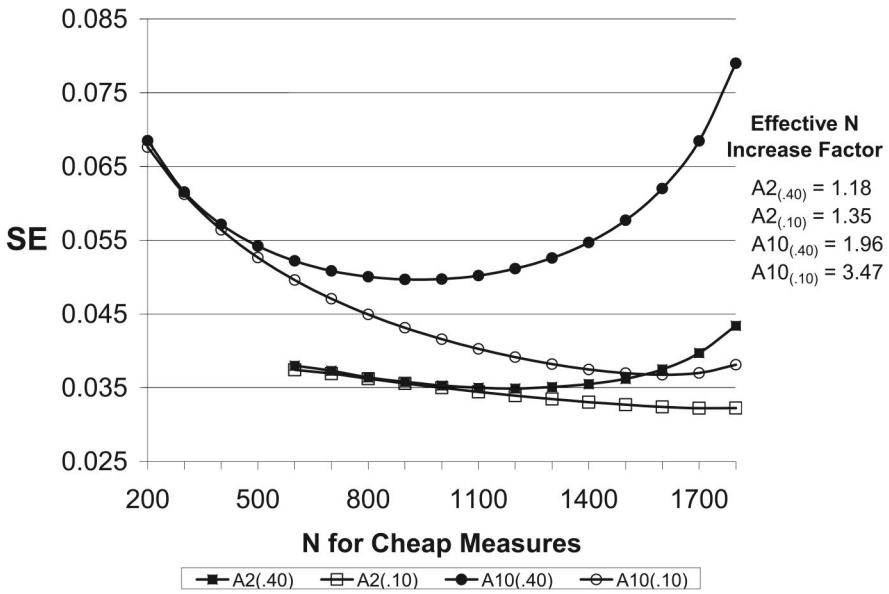

Figure 5. Simulation 3 standard errors for the regression coefficient of the Smoking factor predicting the Health factor. Sample sizes displayed along the *x*-axis refer to the number of study subjects presented with the cheaper, less valid measures. Corresponding *N*s for those also receiving the expensive measures are shown as follows (*N* for cheap measures/*N* for expensive measures): Cells $A2_{(.40)}$ and $A2_{(.10)}$ = (800/549), (1,100/418), (1,400/288), (1,700/157) and Cells $A10_{(.40)}$ and $A10_{(.10)}$ = (500/156), (800/126), (1,100/96), (1,400/66), (1,700/36). All configurations cost approximately $15,050 in total. All standard errors were based on analyses of the population covariance matrix under the various missing data conditions of Simulation 3. Effective *N* Increase Factors were calculated by dividing the effective *N* by the complete cases *N*. Curves $A2_{(.40)}$, $A2_{(.10)}$, $A10_{(.40)}$, and $A10_{(.10)}$ correspond to Simulation Cells $A2_{(.40)}$, $A2_{(.10)}$, $A10_{(.40)}$, and $A10_{(.10)}$, respectively. The cheap measure to expensive measure cost ratio and the effect size of the regression coefficient of interest (Smoking predicting Health) were varied as follows: Cell $A2_{(.40)}$ = 2.3:1 cost ratio, $\rho = .40$; Cell $A2_{(.10)}$ = 2.3:1 cost ratio, $\rho = .10$; Cell $A10_{(.40)}$ = 10:1 cost ratio, $\rho = .40$; and Cell $A10_{(.10)}$ = 10:1 cost ratio, $\rho = .10$.

For Cell $A10_{(.10)}$, the effective *N* was 3.45 times higher than the nominal complete cases *N*. Furthermore, for Cell $A2_{(.10)}$, the best design involved collecting cheap measures for $N = 1{,}700$ cases but expensive measures for just $N = 157$ cases. For Cell $A10_{(.10)}$, the best design involved collecting cheap measures for $N = 1{,}600$ cases but expensive measures for just $N = 46$ cases.

Discussion of the Two-Method Measurement Design

In this section, we discuss the two-method measurement design in light of the knowledge gained from the artificial data illustrations. We discuss power implications and strategies for making use of these designs in empirical research. We also discuss assumptions underlying the design, trade-offs, and limitations of the design.

Implications for Statistical Power

The results in Figures 3–5 show that under a variety of design circumstances planned missing designs yield smaller standard errors and larger effective *N*s for key parameters than do the corresponding complete cases designs. But what implications do these results have for the statistical power for testing key substantive hypotheses? For Cells $A2_{(.40)}$ and $A10_{(.40)}$ in Figure 5, the difference in power was very small between the best and worst designs shown (power $> .995$ in all cases). However, for Cells $A2_{(.10)}$ and $A10_{(.10)}$ in Figure 5, where the effect size was small, there was a clear power differential between the best design and the corresponding complete cases design. For Cell $A2_{(.10)}$, the complete cases design had power = .47 for finding a significant regression coefficient for Smoking predicting Health. With the best missing data design, power = .61. For Cell $A10_{(.10)}$, the complete cases design had power = .18; the best missing data design had power = .55.

Generalizability of Effective N Increase Factor

Of course, these power figures depended on the specific sample sizes under study. Had the sample sizes shown in Figure 5 been much smaller, we might have seen differences in power between complete cases and missing data designs, even for Cells $A2_{(.40)}$ and $A10_{(.40)}$. This illustrates the fact that power, per se, does not generalize to studies with different sample sizes. However, the Effective *N* Increase Factor does generalize well across studies with different sample sizes. For example, any study with the $A10_{(.10)}$ scenario will have a 3.47 Effective *N* Increase Factor, regardless of the sample size. The Effective *N* Increase Factor may thus be used as an aid in calculating power for any study.

Using the Two-Method Design in Practice: Researcher Has Fixed Budget

In this article, we have focused mainly on the situation in which the researcher has a fixed data collection budget. From our artificial data examples, it is clear that for every research scenario, there is a single most efficient design. The problem is that in practice one typically does not know in advance precisely which design will be best. Fortunately, one of the most relevant factors, the cost ratio between cheap and expensive measures, is generally known in advance.

Another relevant factor, effect size, is also easy to deal with. Here we capitalize on the fact that researchers commonly posit a range of plausible effect sizes for their most important effects. We argue that one will always be best off choosing a missing data design that is optimal for the smallest plausible effect size: If the effect turns out to be that small, one has the optimal design. If the effect turns out to be stronger than expected, then the benefit of the larger effect size more than offsets the degree to which the design is not optimal for the larger effect size.

It is also possible to deal with the third relevant factor, factor loadings on the valid common factor and the bias factor. The best way to figure out what to do in this regard is to collect some data using the measures in question prior to start of the study. Many researchers will have collected at least some pilot data for the two kinds of measures. One can then test a two-factor model like the left side of Figure 2 (Smoking and Bias factors). If the factor loadings (on the valid common factor and bias factor) are similar to what we have shown in this article, then one of our designs will likely be very close to optimal, and the researcher can use our figures to make design decisions. If the factor loadings are markedly different from what we have examined here, then one could conduct a brief simulation, as we have done here, with those values. Using the LISREL code we have provided in the online supplemental data to this article, one can write out the population covariance matrix corresponding to those factor loadings. Then, using the provided LISREL code, one could conduct a series of analyses using as input the implied covariance matrix from the previous analysis. In this way, one could tailor a simulation to data one would actually be using.

Assumptions and Trade-Offs

Most important in this regard is the assumption that the lack of construct validity, or bias, in the cheap measures can be modeled. It is easy to model this bias if it can be assumed that the correlation between the bias and the outcome variable (r_{BY}) is negligible. This assumption, which will be met under many research circumstances, was the assumption made in the present study. When the assumption that $r_{BY} = 0$ is not viable, other models are possible that would estimate that correlation (or regression) as well (e.g., see Gra-

ham & Collins, 1991). Although some such models do tend to be a bit unstable because of model underidentification, we have found that many of these bias correction models are stable in this context (e.g., Palmer et al., 2002). In addition, the Bayesian approach (e.g., see Taub et al., 2005) holds much promise in this context.

As with the 3-form design and other designs in that family, there are trade-offs with the two-method measurement design. The advantage of using this kind of design has been well described in the previous sections. General linear model analyses involving the construct of interest may be tested with more power. The trade-off, however, is that other hypotheses not involving the general linear model may be tested with less power. This loss of power would likely occur for analyses for which the researcher prefers to use the more valid, expensive measure alone. In addition, the advantages we have described for two-method measurement are less likely to be realized if the SEM approach we described here is not an option. On the other hand, as Bayesian approaches are developed, such as that described recently by Taub et al. (2005), they may be good alternatives that can be used in this context when the SEM approach cannot.

Future Research Directions and Open Questions

We have described two kinds of measurement design involving planned missing data. Although these designs may not be suited for every circumstance, they have been and will continue to be an enormous benefit to researchers in a wide variety of psychological research settings.

The 3-Form Design and Other Members of This Family of Planned Missing Data Designs

The 3-form design has already proven to be highly useful. But it is important to remember the basic condition underlying this type of design: The 3-form design (and other designs in this family) will be useful whenever one has the time and other resources to ask some number of questions of one's respondents but would like to ask more questions. If this basic condition is met, then one of these designs will be useful.

In this article, we focused mainly on the 3-form design. This design is particularly well suited to the situation in which one wishes to leverage one's resources to a modest degree; the 3-form design involves asking approximately 33% more questions than can be answered reasonably by any one subject. But what if one wants to leverage one's resources to an even greater extent? What if it is important to the researcher to be able to collect data on even more variables than would be possible with the 3-form design? One good option would be to choose another member of this family of designs (e.g., the SQSD; Raghunathan & Grizzle, 1995). We suspect that in order for these larger designs to be viable, one must have a rather large sample size, for example, $N = 3{,}000$ or larger. The practical limits with respect to sample size for the various designs in this family of design will be addressed in future research.

Small samples. We have provided power information for a study with $N = 300$. On the basis of the information we have provided (e.g., see Table 5), we believe the 3-form design can be useful with samples as small as this. But what if one's study sample size is even smaller? Or what if one's sample size is $N = 300$, but one is uncomfortable using the 3-form design as described? Under these circumstances, it is very possible to adjust the 3-form design so that the leveraging of one's resources is even more modest than described here. And with reduced leveraging comes less risk.

Actual power with the 3-form design. In the present article, we calculated power (e.g., in Table 5) on the basis of simple correlation coefficients. For these calculations, we used the nominal sample size for various parts of the design. However, the true power for testing effects with the 3-form design may be different from what we have shown. In fact, the power values shown in Table 5 relating to the 3-form design are a lower bound on the power to test these various effects. It has been shown that the power to test a particular effect is enhanced in the missing data case to the extent that other variables are available (auxiliary variables) that are highly correlated with the variables containing missing data (e.g., see Collins et al., 2001). Calculating the increase in power under various conditions is beyond the scope of this article, but suffice it to say, one's actual power with the 3-form design will generally be at least a little better than shown in Table 5, especially if one has longitudinal data. These issues will be pursued in future research.

Two-Method Measurement

The two-method measurement design shows enormous promise. Like the 3-form design, two-method measurement allows researchers to collect high quality data in a cost effective manner. By no means did our illustrations cover all possible research scenarios. Given the scenarios that we did examine, we showed that there was always a benefit of using a two-method measurement design and that the benefit ranged from a modest 10% increase in effective sample size to as high as tripling the effective sample size compared with the corresponding complete cases design.

Future research relating to this design will explore research domains in which the design may be applied. Our examples mainly involved measurement of cigarette smoking. But we believe that two-method measurement designs will prove to be useful in numerous other domains, including studies relating to nutrition intake, adiposity measurement, physical activity, physical conditioning, and blood vessel health. In fact, this design should prove to be useful

in virtually any research domain in which highly valid, but expensive, measures are sought.

Future research will also explore extensions of the two-method measurement design itself. What other configurations of the design will prove to be useful? To what extent is it possible to have just one cheap measure, just one expensive measure, or just one of each? To what extent is it possible to make use of more standard statistical methods (i.e., not SEM) to achieve some or all of the benefits we have described? To what extent will other approaches, such as Bayesian approaches (e.g., Taub et al., 2005) prove to be useful?

MAR designs: Two-phase designs. Most of what we present in this article relates to what might be thought of as MCAR designs. Cases are randomly assigned to receive both the cheap and expensive measures or just the cheap measure. But what about the two-phase designs in which the cheap and expensive measures are given in sequence (e.g., Deming, 1977; Shrout & Newman, 1989; Wittchen, Kessler, Zhao, & Abelson, 1995)? In theory, there is no reason why this sort of design would be any different from those we have described. The fact that the missingness on the expensive measure is dependent entirely on the first (cheap) measure makes this a case of MAR missingness. One assumption in making this sort of design work is that the two measures can reasonably be thought of as measuring the same construct. Another assumption is that it must be reasonable to think of the expensive measure as being a gold standard (i.e., completely valid) measure of the construct in question. If these two assumptions are met, the two-phase measurement design fits neatly into the two-method measurement framework, and all the benefits we have discussed apply.

Small sample size. One of the biggest strengths of the two-method measurement design is that the benefit can be assessed using the Effective *N* Increase Factor, which shows the proportion by which the effective sample size is increased using this design. The result of the two-method measurement is typically to increase the effective sample size without increasing costs. This will be a benefit in most research scenarios regardless of sample size.

References

Adams, L. M., & Darwin, G. (1982). Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education, 17,* 231–240.

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2,* 20–33.

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 71–103). San Francisco: Jossey Bass.

Arbuckle, J. L. (1995). *Amos users' guide.* Chicago: SmallWaters.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide.* Chicago: SmallWaters.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs.* Oxford, England: Clarendon Press.

Avesani, C. M., Draibe, S. A., Kamimura, M. A., Cendoroglo, M., Pedrosa, A., Castro, M. L., & Cuppari, L. (2004). Assessment of body composition by dual energy X-ray absorptiometry, skinfold thickness and creatinine kinetics in chronic kidney disease patients. *Nephrology Dialysis Transplantation, 19,* 2289–2295.

Berman, J. S., & Kenny, D. A. (1976). Correlational bias in observer ratings. *Journal of Personality and Social Psychology, 34,* 263–273.

Biglan, A., Gallison, C., Ary, D., & Thompson, R. (1985). Expired air carbon monoxide and saliva thiocyanate: Relationships to self-reports of marijuana and cigarette smoking. *Addictive Behaviors, 10,* 137–144.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building.* New York: Wiley.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Collins, L. M., Hyatt, S. L., & Graham, J. W. (2000). LTA as a way of testing models of stage-sequential change. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 147–161). Hillsdale, NJ: Erlbaum.

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine, 30,* 65–73.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330–351.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Cox, C. P. (1958). Experiments with two treatments per experimental unit in the presence of an individual covariate. *Biometrics, 14,* 499–512.

Davis, H., & Gergen, P. J. (1994). The weights and heights of Mexican-American adolescents: The accuracy of self-reports. *American Journal of Public Health, 84,* 459–462.

Deming, W. E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *International Statistical Review, 45,* 29–37.

du Toit, M., & du Toit, S. (2001). *Interactive LISREL: User's guide.* Lincolnwood, IL: Scientific Software International.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65,* 241–261.

Fisher, J. O., Johnson, R. K., Lindquist, C., Birch, L. L., & Goran, M. I. (2000). Influence of body composition on the accuracy of

reported energy intake in children. *Obesity Research, 8,* 597–603.

Flay, B. R., Miller, T. Q., Hedeker, D., Siddiqui, O., Britton, C. F., Brannon, B. R., et al. (1995). The Television, School, and Family Smoking Prevention and Cessation Project: VIII. Student outcomes and mediating variables. *Preventive Medicine, 24,* 29–40.

Gelman, A. (2000). Should we take measurements at an intermediate design point? *Biostatistics, 1,* 27–34.

Going, S. B. (1996). Densitometry. In A. F. Roche, S. B. Heymsfield, & T. G. Lohman (Eds.), *Human body composition* (pp. 3–24). Champaign, IL: Human Kinetics Books.

Goodman, E., Hinden, B. R., & Khandelwal, S. (2000). Accuracy of teen and parental reports of obesity and body mass index. *Pediatrics, 106,* 52–58.

Graham, J. W., & Collins, N. L. (1991). Controlling correlational bias via confirmatory factor analysis of MTMM data. *Multivariate Behavioral Research, 26,* 607–629.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research methods in psychology: Vol. 2 Handbook of psychology* (pp. 87–114). New York: Wiley.

Graham, J. W., Flay, B. R., Johnson, C. A., Hansen, W. B., Grossman, L. M., & Sobel, J. L. (1984). Reliability of self-report measures of drug use in prevention research: Evaluation of the Project SMART questionnaire via the test-retest reliability matrix. *Journal of Drug Education, 14,* 175–193.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 201–218). Hillsdale, NJ: Erlbaum.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31,* 197–218.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. Seitz (Eds.), *Advances in data analysis for prevention intervention research: National Institute on Drug Abuse Research monograph series* (No. 142, pp. 13–63). Washington, DC: National Institute on Drug Abuse.

Graham, J. W., Johnson, C. A., Hansen, W. B., Flay, B. R., & Gee, M. (1990). Drug use prevention programs, gender, and ethnicity: Evaluation of three seventh-grade Project SMART cohorts. *Preventive Medicine, 19,* 305–313.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in analysis of change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association.

Hansen, W. B., & Graham, J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine, 20,* 414–430.

Hansen, W. B., Johnson, C. A., Flay, B. R., Graham, J. W., & Sobel, J. L. (1988). Affective and social influences approaches to the prevention of multiple substance abuse among seventh grade students: Results from Project SMART. *Preventive Medicine, 17,* 135–154.

Hawkins, J. D., Graham, J. W., Maguin, E., Abbott, R., Hill, K. G., & Catalano, R. F. (1997). Exploring the effects of age of alcohol use initiation and psychosocial risk factors on subsequent alcohol misuse. *Journal of Studies on Alcohol, 58,* 280–290.

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., Rickert, W., & Robinson, J. (1989). Measuring the heaviness of smoking: Using self-reported time to the first cigarette of the day and number of cigarettes smoked per day. *British Journal of Addiction, 84,* 791–800.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5,* 64–86.

Hyatt, S. L., & Collins, L. M. (2000). Using latent transition analysis to examine the relationship between parental permissiveness and the onset of substance use. In J. Rose, L. Chassin, C. Presson, & S. Sherman (Eds.), *Multivariate applications in substance use research* (pp. 259–288). Hillsdale, NJ: Erlbaum.

Johnson, E. G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14,* 303–334.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 95–110.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide.* Moorseville, IN: Scientific Software.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 69–86.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247–252.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95,* 49–69.

Kozlowski, L. T., & Herling, S. (1988). The use of objective measures in smoking treatment. In A. Marlatt & D. Donovan (Eds.), *Assessment of addictive behaviors* (pp. 214–235). New York: Guilford Press.

Lanza, S. T., Collins, L. M., Schafer, J. L., & Flaherty, B. P.

(2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychological Methods, 10,* 84–100.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lohman, T. G. (1996). Dual energy x-ray absorptiometry. In A. F. Roche, S. B. Heymsfield, & T. G. Lohman (Eds.), *Human body composition* (pp. 63–78). Champaign, IL: Human Kinetics Books.

Lord, F. M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement, 22,* 259–267.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

Munger, G. F., & Loyd, B. H. (1988). The use of multiple matrix sampling for survey research. *Journal of Experimental Education, 56,* 187–191.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide.* Los Angeles, CA: Muthén and Muthén.

Neale, M. C. (1991). Mx: Statistical modeling [Computer software]. Richmond: Virginia Commonwealth University, Department of Human Genetics.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). Mx: Statistical modeling (5th ed.) [Computer software]. Available from http://www.vcu.edu/mx/

Palmer, R. F., Graham, J. W., Taylor, B. J., & Tatterson, J. W. (2002). Construct validity in health behavior research: Interpreting latent variable models involving self-report and objective measures. *Journal of Behavioral Medicine, 25,* 525–550.

Pechacek, T. F., Murray, D. M., Luepker, R. V., Mittelmark, M. B., & Johnson, C. A. (1984). Measurement of adolescent smoking behavior: Rationale and methods. *Journal of Behavioral Medicine, 7,* 123–140.

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

Raghunathan, T. E., Lepkowski, J. M., VanHoewyk, J., & Solenberger, J. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27,* 85–95.

Roubenoff, R., Kehayias, J. J., Dawson-Hughes, B., & Heymsfield, S. B. (1993). Use of dual-energy x-ray absorptiometry in body-composition studies: Not yet a "gold standard." *American Journal of Clinical Nutrition, 58,* 589–591.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* New York: Chapman and Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling.* Cambridge, MA: Ballinger Publishing.

Shrout, P. E., & Newman, S. C. (1989). Design of two-phase prevalence surveys of rare disorders. *Biometrics, 45,* 549–555.

Taub, N. A., Morgan, Z., Brugha, T. S., Lambert, P. C., Bebbington, P. E., Jenkins, R., et al. (2005). Recalibration methods to enhance information on prevalence rates from large mental health surveys. *International Journal of Methods in Psychiatric Research, 14,* 3–13.

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.

Thompson, S. K. (2002). *Sampling* (2nd ed.). New York: Wiley.

Thompson, S. K., & Collins, L. M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence, 68,* S57–S67.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28,* 317–331.

West, S. G., & Aiken, L. S. (1997). Toward understanding individual effects in multicomponent prevention programs: Design and analysis strategies. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 167–209). Washington, DC: American Psychological Association.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Williet, W. W. (1990). *Nutritional epidemiology.* New York. Oxford University Press.

Wittchen, H. U., Kessler, R. C., Zhao, S., & Abelson, J. (1995). Reliability and clinical validity of UM-CIDI *DSM-III-R* generalized anxiety disorder. *Journal of Psychiatric Research, 29,* 95–110.

Received November 16, 2004
Revision received August 17, 2006
Accepted August 24, 2006 ■