Tools of the Trade:

Planned Missing Data Designs for Research in Cognitive Development

Mijke Rhemtulla and Todd Little

Center for Research Methods and Data Analysis

University of Kansas

Author Note

This work was supported by a Banting postdoctoral fellowship to M. Rhemtulla and an NSF grant (NSF0066969; T.D. Little & W. Wu, co-PIs).

We thank members of the University of Kansas Center for Research Methods and Data Analysis (Todd Little, Director) for feedback on a previous draft of this paper.

Correspondence may be addressed to M. Rhemtulla, mijke@ku.edu, or T. Little, yhat@ku.edu.

Abstract

Data collection can be the most time- and cost-intensive part of developmental research. This article describes some long-proposed but little-used research designs that have the potential to maximize data quality (reliability and validity) while minimizing research cost. In *planned missing data designs*, missing data are used strategically to improve the validity of data collection in one of two ways. Multi-form designs allow one to increase the number of measures assessed on each participant without increasing each participant's burden. Two-method measurement designs allow one to reap the benefits of a cost-intensive gold-standard measure, using a larger sample size made possible by a rougher, cheaper measure. We explain each method using examples relevant to cognitive development research. With the use of analysis methods that produce unbiased results, planned missing data designs are an efficient way to manage cost, improve data quality, and reduce participant fatigue and practice effects.

Keywords: missing data, multi-form designs, two-method measurement, intentionally missing data

Data collection can be the most time- and cost-intensive part of developmental research. Developmental scientists are seldom able to avail themselves of the conveniences of a ready subject pool, online survey methodology, or even self-administered measures. Instead, behavioral measures (e.g., habituation, head-turn, pointing, and reaching tasks), individual cognitive assessments, laboriously transcribed videotapes of interactions, and imaging techniques (e.g., optical imaging, Hespos, 2010; electroencephalography, Csibra, Kushnerenko, & Grossman, 2008; fMRI, Redcay et al., 2010) are commonly necessary if we are to measure the developing understanding of our populations. Gaining access to participants, and finding the right time, situation, and measures to assess them, places strong limits on the amount of data that can be gathered. Not only is a goal to maximize the number of participants and the amount of information per participant, but also to maximize the reliability and validity of each participant's data. Achieving this goal frequently involves trade-offs: Sometimes, a rougher measure may be used to cut costs and allow a larger sample size; a paper and pencil IQ test rather than administering the WISC, for example, or a teacher questionnaire about children's reading ability rather than a reading assessment. In other cases, a short form of an assessment may be used to avoid participant fatigue and gather data from multiple assessments in one sitting.

We describe a tool to help optimize this trade-off. In planned missing data designs, missing data are used strategically to improve the validity of data collection in one of two ways. Multi-form designs (Graham, Hofer, & Piccinin, 1994; Graham, Taylor, Olchowski, & Cumsille, 2006) allow one to increase the number of measures assessed on each participant without increasing each participant's burden. Two-method measurement (Graham & Collins, 1991; Graham et al., 2006; Palmer, Graham, Taylor, & Tatterson, 2002) allows one to reap the benefits of a cost-intensive gold-standard measure, using a larger sample size made possible by a rougher,

cheaper measure. Before detailing these designs, we briefly overview modern approaches to analyzing missing data.

Missing Data

Missing data arise for many reasons. These reasons, or *missingness mechanisms*, are classified into three categories. The cause of missingness is either completely unrelated to any of the observed or missing variables (missing completely at random; MCAR), related to the observed variables but not the missing values themselves (missing at random; MAR), or directly related to the values that are missing (missing not at random; MNAR). MCAR missingness is ideal because any modern missing data estimation procedure will provide unbiased estimates and recover much of the power lost due to data attrition. Unfortunately, *unplanned* MCAR is probably the least common missing data process.

When missing data are MAR, missingness is predictable from other variables in the observed data (e.g., if girls tended to skip a particular question but boys did not, then the missingness would be predicted by sex). Even in the MAR case, modern missing data estimation procedures recover much of the lost information and power, to the extent that the missing values are predictable by the remaining observed variables. When missing data are MNAR, then the missingness is predicted by the missing variables (e.g., if children with poor math skills tended to skip the math questions on an assessment, and there were no other observed indicators of math ability). In the MNAR case, there is no way to recover the missing information. A critical design feature, therefore, is the collection of data on variables that are likely to predict the missingness. The degree to which modern approaches are able to recover the missing data process depends on (a) whether correlates of the missing data mechanism are measured and (b) the strength of the linear relationship of this known process with the missing data.

There are two recommended modern missing data estimation techniques (see Enders, 2010, for a detailed and approachable resource). The first is full information maximum likelihood (FIML). FIML uses all of the observed data to produce parameter estimates (e.g., regression parameters, means, or variance estimates) that maximize the probability of the observed data having come from the population implied by those estimates. FIML is implemented in most structural equation modeling software, but it can be used for a wide range of simple and complex analyses including multiple regression. The second is Multiple imputation (MI), which is a "data-based" method to achieve the same results as FIML. In MI, *m* (the number of imputations, which should be 20 or more) values are generated for each missing value in the dataset, resulting in *m* complete datasets. Then the analysis is carried out on each dataset, and the results are combined across datasets to achieve parameter estimates and standard errors that are as accurate as those produced by FIML.

With both methods, it is possible to include key auxiliary variables in the estimation procedure that may or may not be relevant to the theoretical model but are likely related to the missingness (e.g., SES, sex). In MI, this is done by including auxiliary variables in the imputation model but not in the analysis model. In FIML, this process involves including the auxiliary variables as "saturated correlates" in the model, allowing them to influence estimation without affecting the interpretation of model parameters (Graham, 2003); this procedure has been automated in M*plus* as of version 5.1 (Asparouhov & Muthén, 2008). Measuring and including the right auxiliary variables can turn unplanned MNAR missingness into MAR, by turning unpredictable missingness into predictable missingness. Planned missing designs can reduce both MAR and MNAR missingness by reducing the testing burden on participants; that

is, planned MCAR missingness can result in less unplanned missingness[1].

In the following sections we review two planned missing designs that have been proposed (e.g., Graham, Taylor, Olchowski, & Cumsille, 2006; Graham, Taylor, & Cumsille, 2001), and give examples of how they might be used and adapted for developmental research.

Multi-Form Design

Planned missing solutions for various types of data and design have been proposed in the psychology literature, including designs for multi-trait multi-method research (Bunting, Adamson, Mulhall, 2002), longitudinal growth curves (Graham, Taylor, & Cumsille, 2001; McArdle & Woodcock, 1997), and cross-sectional survey research (Graham et al., 2006). One of the simplest and most versatile of these designs is the multi-form measurement design (Graham et al., 2006), where each participant is randomly assigned to receive only a subset of the items on a longer measure. The multi-form design can be used in a cross-sectional study and easily extended to multiple time points to lower the burden of longitudinal data collection.

The multi-form design is most useful when one wants to collect data using a large number of items or responses but is faced with time constraints or concerns about respondent burden and fatigue. The most popular multi-form design is the three-form design (Hansen et al., 1988; Graham et al., 2006; Graham, Hofer, & MacKinnon, 1996; Graham et al. 1994), in which a large number of items (e.g., survey questions) are divided into 4 subsets including a common block (X) and three partial blocks (A, B, and C). The items assigned to X are those that are central to the study's hypotheses or are indicative of potential MAR missing processes; this

[1] Researchers are encouraged to examine and report their unplanned missingness, whether or not a planned missing design is used. Many software packages contain tools to help visualize and quantify the amount of missing data. With planned missing designs, it may be helpful to use a single-imputation technique (e.g., EM imputation) to create a complete dataset, and then re-delete unplanned missing values before carrying out analyses to characterize the remaining missingness (e.g., tests of MCAR missingness, sensitivity analyses for MNAR missingness; see Enders, 2010). A single EM imputation should not be used to carry out the substantive analyses of interest.

block is administered to all participants and is usually administered first. The partial blocks of items in A, B, and C form three combinations: AB, BC and AC. Each combination is administered to one third of the participants (see Figure 1A). As a result, one third of the participants do not answer the questions in sets A, B and C. This shortened survey length results in about one third more items in a survey than in a complete data design, given the same time constraints. Or, maintaining the same survey size as the complete data design, each participant will answer fewer questions, reducing the effect of fatigue and minimizing drop-out. Most importantly, researchers randomly control which items are missing. If order effects are a concern, additional forms with different block orders can be created.

Depending on qualities of the survey such as the number of items and the presence of subscales, the protocol may be divided into more than the 4 parts (X, A, B, C) and more than three forms may be needed. As the number of item sets grows, each form will typically be missing more than one item set, though this is entirely at the researcher's discretion. For example, a six-form design can be constructed with five item sets (X, A-D; Figure 1B), where each form is missing two item sets, or a ten-form design can be constructed with six item sets (X, A-E), where each form is missing three item sets (Graham et al., 2006).

To get the most power out of this design, items between sets should be maximally correlated (Raghunathan & Grizzle, 1995). Higher correlations between item sets provide more information for the missing data estimation procedures, so statistics will be estimated with greater certainty (i.e., more precision) than when the correlations among item sets are low. The lower the correlation between forms the more power will be lost. Even if the correlations are very high, some power will be lost (see Enders, 2010). To maximize the degree of correlation among the item sets, we recommend breaking up any subscales and including some items from

each subscale in each item set. For example, a 9-item measure of mathematics skill would have 3 items assigned to block A, 3 items to block B, and 3 items to block C. Any one participant would respond to 6 of the nine items and the correlations of these 6 responded-to items would be optimized with regard to the 3 items that were randomly assigned to be missing. If important items exist that are expected to be unrelated to most other items in the survey, they should be placed in the X set.

When to use multi-form designs. Multi-form designs are very versatile designs, but they may not always be ideal. Experimental studies where very few behaviors are assessed, for example, are not good candidates for these designs. Properties of research that are best suited to multi-form designs include:

1) *The ideal battery is too long*. Length can be an issue because of time constraints (where demands on the time of the participants or the experimenters requires a brief interview), or attention constraints (where participants could not reasonably be expected to give their full attention for long enough to administer all the measures of interest).
2) *It is possible to increase the sample size.* If the predicted effects are expected to be small, the correlations between items are weak, or the analysis tool requires larger sample sizes, a multi-form design may require a larger sample size than the original complete data design. As the correlation between the forms of a multi-form design become smaller, more power will be lost compared to a design where every participant responds to every measure. Even if the correlations are very high, some power will be lost.
3) *The research focus is at the group level*. If the research goal is individual assessment rather than estimating the relations between variables, planned missing designs are not appropriate. Modern missing data estimation procedures enable unbiased estimation of

population parameters, but they cannot be used to recover an individual's unique score on a missing variable.

4) *Sample size is sufficiently large*. Modern missing data estimation requires that the sample be large enough to estimate the covariances among variables. Covariance estimates stabilize at around $N = 125$ complete cases (Little, in press), so a conservative estimate would be three times this number ($N = 375$), to allow for sufficient overlapping complete cases to estimate cross-set covariances.

Multi-form design extensions. Variations of the multi-form design are limited only by the practical necessities of the research. More than three forms might be warranted by the number and size of scales or measures to be included; for example, if an assessment consists of a 4 subscales containing 10 items each, one representative item from each can be placed in the X-set, and 3 items each in each of 3 forms, or two representative items from each can be placed in the X-set, and 2 items in each of 4 forms. If different subscales have a different number of items, another number of forms might be ideal. If a subscale cannot be broken up into forms it can be included in the X-set. If items across forms are expected to be highly correlated with each other, then a large amount of planned missing data may be appropriate, whereas less missing data would be warranted when correlations are expected to be low.

The multi-form design is typically described in the context of cross-sectional studies, but it is easily carried into a longitudinal framework by simply using the same multi-form design at each measurement occasion. One of two strategies can be used for assigning participants to forms across measurement occasions. If measuring the relations between single variables across measurement occasions is of greatest interest, giving each participant the same form across all measurement occasions is warranted (Max Occasion Overlap; see Table 1). This strategy ensures

the maximum amount of overlap of each variable with itself across time, because every participant responds to the same items at every measurement occasion. If minimizing re-test effects and testing the relations among latent variables is of greater interest, then using a different form at each measurement occasion is preferred (Max Item Overlap; see Table 1). This strategy ensures that every participant contributes information to the measurement of the relations between every pair of variables, because each participant will respond to every item at least once. To balance these approaches and to simplify survey administration, one can randomly assign participants to forms at each measurement occasion, independent of what form they received previously.

Two-Method Measurement Design

A second type of planned missing design involves using two different measures of a single construct (Graham et al., 2006). This design relies on the availability of more than one instrument or method by which a construct of interest might be measured. Often some methods are easy and inexpensive but may be biased (e.g., teacher reports of children's cognitive abilities) and others are expensive but more accurate (e.g., direct assessments of children's cognitive abilities). For this design to work, one measure must be both inexpensive and known to contain bias, and a second measure must be both more expensive or time-consuming and known to be unbiased (valid). If an ideal measure exists that is both inexpensive and valid, then of course the most efficient design would use this measure alone. The two-method measurement design uses a simple structural equation model to estimate the unbiased relations among constructs of interest.

Inexpensive biased measures, if used alone, can be administered to a large number of participants, yielding larger sample sizes; however, research using biased measures alone always suffers from degraded validity (e.g., Ready & Wright, 2011). In contrast, research using

expensive unbiased measures alone frequently limits the practical sample size (unless research funds and access to participants are unlimited), leading to lower statistical power and weaker statistical inference techniques. The two-method design borrows the advantages from both types of measures by using both in a single study. Every participant in a relatively large sample is administered the inexpensive (biased) measure. In addition, a random subset of participants is administered the expensive (unbiased) measure[2]. By measuring a random subsample with the unbiased measure, the degree of bias in the inexpensive measure can be estimated and removed. The result is maximum cost efficiency: an optimally large sample size (which provides sufficient power) coupled with unbiased measurement of the construct (which optimizes validity).

The optimal ratio of total sample size to expensive-measure sample size depends on 3 factors, including the cost ratio (where a larger cost ratio requires a larger sample size ratio), the effect size (where a smaller effect size requires a larger sample size ratio), and the reliability of each indicator (where a more reliable inexpensive measure requires a larger size ratio). Graham et al. (2006) display each of these trends in figures, along with estimates of "effective N" that can be used in power calculations.

As with multi-form designs, certain criteria must be met for the two-method measurement design to be appropriate.

1) *The inexpensive measure is systematically biased*. A systematically biased instrument measures something else in addition to the construct it is designed to measure; for example, a parent-report measure of child ability might measure social desirability bias in

[2] If the full sample has a complex structure, including stratified sampling characteristics, nesting, and/or subgroups with small N, it will be important to ensure that the subsample administered the expensive measure is broad enough to capture this diversity. It may be necessary to choose a larger random sample for the expensive measure, and/or assign participants with particular characteristics (e.g., minorities) to receive both measures. If the latter approach is taken, the variables used to select participants must be included in the missing data estimation procedure.

addition to the abilities of their child.

2) *The expensive measure is unbiased*. The expensive measure should be considered a "gold standard" for measuring a construct[3]. If it is also biased but to a much smaller degree, then the two-method measurement design can be appropriate; in this case, it will only be possible to attenuate the bias of the cheap measure to the degree that the expensive measure is also biased. A bifactor model that removes different types of bias from two measures, however, may also be possible (Little, in press).
3) *Both measures access the same construct*. The two measures may be on different scales (e.g., parent report of child's intelligence and WISC-R scores) but the underlying construct that is being assessed must be the same.
4) *The research focus is at the group level*. The two-method measurement approach does not involve computing a correction to the cheap measure that can be applied to individual scores. Instead, the correction applies to the latent variable measuring the construct.
5) *Sample size is sufficiently large*. No research has examined how big the total sample size must be to ensure accurate estimates in this model. In general, a minimum sample size of $N = 125$ *complete cases* will result in covariances estimates that are stable enough to allow estimation of most structural equation models (Little, in press). The total sample, then, should be at least 125 times the chosen ratio of inexpensive to expensive sample sizes (see above).

Example. Several researchers have examined the question of whether children's early attentiveness in the classroom predicts later cognitive achievement through elementary school (Duncan et al., 2007; Fuchs et al., 2005; Fuchs et al., 2010). These studies typically use a

[3] We assume that the degree of bias in each of these measures is roughly known from prior literature, including explicit validation attempts.

teacher-report or maternal-report measure of attention such as the SWAN (Swanson et al., 2004) where a child's attention is rated on 18 Likert-scaled behavior items derived from the DSM-IV criteria for attention disorders (American Psychological Association, 2000). Because the SWAN is inexpensive and easy to administer, it is attractive to use. Teacher-reports of a child's classroom attention, however, are invalid to the extent that teachers display report bias such as being affected by the child's socioeconomic status and ethnicity (Ready & Wright, 2011; Stevens, 1980), level of oppositional behaviors (Stevens, Quittner, & Abikoff, 1998), and child gender (Sciutto, Nolfi, & Bluhm, 2004). One measure of child attention that is not prone to these biases may be a classroom observation system such as the Classroom Observations of Conduct and Attention Deficit Disorders (COCADD; Atkins, Pelham, & Licht, 1985) where an independent observer examines many instances of each child's behavior, coding each according to a list of attentive and inattentive behaviors. Direct observations such as COCADD are time- and cost-intensive, and infeasible for very large samples that are frequently used in this type of research.

Suppose that direct observation of each child requires $40 on average (i.e., the cost of training, and the time spent observing and coding). In addition, suppose that teacher ratings cost approximately $10 each to administer and collect. For a fixed budget of $10,000, then, it would be possible to obtain teacher ratings of 1000 students, or direct observations of 250 students. Such a consideration would induce many researchers to eschew direct observation entirely. *Both* measures, however, can be included in a research design to gain the advantages of both methods. Using the two-method measurement design, for the same budget, 500 children could be assessed with teacher ratings, and 125 of those 500 children (25% of the total sample) could additionally be assessed using direct observations.

Analysis. The analysis model is shown in Figure 2. The analysis of the two-method measurement approach requires the use of a straightforward structural equation model (see Klein, 2010, for an accessible introduction to structural equation modeling). Attention is modeled as a latent variable with the teacher report and direct assessment measures as indicators. The latent variable Attention reflects the variance shared by all of its indicators; that is, only the reliable variance that is captured by both teacher report and direct assessment is captured by the latent variable. The remaining variance (that which is not shared among all measures) makes up each indicator's *residual* variance. In Figure 2, teacher report is divided into two indicators, where each is the average of half the items on the scale[4]. The reason for using more than one indicator per measure is that it allows the model to separate reliable from unreliable variance (for a primer on the logic of latent variable models, see Kline, 2010). The residuals of these two indicators are allowed to load on a secondary construct; this residual information reflects the amount of variance that is shared by the teacher-report measures but not with the direct assessment measures. This secondary construct is termed *Teacher Report Bias* because the common information among teacher-report measures that is not related to the direct assessment measures is the amount of teacher-report bias. This construct may or may not be of interest, but it must be estimated to ensure that none of the teacher bias is allowed to affect the latent Attention variable. This model does not include a Bias factor for the direct assessment measures, because they are assumed not to contain bias.

In turn, Attention may be used as a dependent variable or as a predictor of other variables of interest (e.g., third grade math ability), and the resulting regression coefficients will be free of

[4] Each measure should be composed of at least two variables (as shown in Figure 2). We used indicators that represent the two halves of each scale for simplicity, but such parceling is not necessary. For example, each scale item could be included as a separate indicator of Attention. For more information about parceling options, see Little, Rhemtulla, Gibson, & Schoemann (in press).

bias, to the extent that both the inexpensive and expensive measures are appropriate.

Summary

Planned missing data designs provide a rare opportunity to maximize data efficiency. On the basis of unequivocally sound statistical theory, statisticians have long championed these methods. Recent advances in software and computational capacity now make implementing them efficient and easy. In this paper, we reviewed two cross-sectional planned missing designs: the multi-form design, which can reduce participant burden and fatigue effects, and the two-method measurement design, which is a low-cost method to attenuate the bias in low-quality survey data. Although these methods are both cross-sectional, they can easily and effectively be extended to longitudinal designs. Both designs provide high-quality, highly valid data, at relatively low cost.

Additional Resources

Researchers may be interested in computing scale reliability (e.g., Cronbach's alpha). When using MI, this simply requires computing validity in each imputed dataset and combining the estimates across sets. When using FIML, Raykov (2009) describes a model-based method to compute scale validity that works with missing data.

Enders' (2010) textbook contains a very good chapter on carrying out power calculations with planned missing data designs, as well as many other relevant applied topics.

The Center for Research Methods and Data Analysis at the University of Kansas maintains a series of how-to guides on a range of quantitative methods topics (accessible at crmda.ku.edu/kuantguides). These currently include guides on multiple imputation, auxiliary variables in FIML, and creating a 3-form planned missing design.

References

American Psychiatric Association (2000). Diagnostic and Statistical Manual of Mental Disorders (4th ed., text revision). Washington, DC: Author.

Atkins, M. S., Pelham, W. E., & Licht, M. H. (1985). A comparison of objective classroom measures and teacher ratings of attention deficit disorder. *Journal of Abnormal Child Psychology, 13,* 155-167. doi: 10.1007/BF00918379

Asparouhov, T., & Muthén, B. O. (2008). Auxiliary variables predicting missing data. Technical report. Available at www.statmodel.com/download/AuxM2.pdf

Bunting, B. P., Adamson, G., & Mulhall, P. K. (2002). A Monte Carlo Examination of an MTMM Model With Planned Incomplete Data Structures. *Structural Equation Modeling, 9,* 369 - 389. doi: 10.1207/S15328007SEM0903_4

Csibra, G., Kushnerenko, E., & Grossman, T. (2008). Electrophysiological methods in studying infant cognitive development. In C. A. Nelson & M. Luciana (Eds.), Handbook of developmental cognitive neuroscience (2nd ed., pp. 247–262). Cambridge, MA: The MIT Press.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R. (2010). fMRI item analysis in a theory of mind task. *NeuroImage, 55,* 705-712. doi: 10.1016/j.neuroimage.2010.12.040

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A., Klebanov, P., Pagani, L. S., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428-1446. doi: 10.1037/0012-1649.43.6.1428

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

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of*

Educational Research, *102*, 453-462. doi: 10.1037/0022-0663.97.3.493

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, *97*, 493-513. doi: 10.1037/0022-0663.97.3.493

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., & Hamlett, C. L. (2010). The contributions of numerosity and domain-general abilities for school readiness. *Child Development, 81*, 1520-1533. doi: 10.1111/j.1467-8624.2010.01489.x

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10,* 80-100. doi: 10.1207/S15328007SEM1001_4

Graham, J. W., & Collins, N. L. (1991). Controlling correlational bias via confirmatory factor analysis of MTMM data. *Multivariate Behavioral Research, 26*, 607-629. doi: 10.1207/s15327906mbr2604_3

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.), National Institute on Drug Abuse Research Monograph Series (pp. 13-62). Washington, DC: National Institute on Drug Abuse.

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

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

Graham, J. W., Taylor, B. J.,& Cumsille, P. E. (2001). Planned missing data designs in the

analysis of change. In L. M. Collins &A.G. Sayer (Eds.), New methods for the analysis of change (pp. 335−353). Washington, D.C.: American Psychological Association. doi: 10.1037/10409-011

Hansen, W. B., Graham, J. W., Wolkenstein, B. H., Lundy, B. Z., Pearson, J. L., Flay, B. R., & Johnson, C. A. (1988). Convergent and discriminant validity for assessment of skill in resisting a role play alcohol offer. *Behavioral Assessment, 11*, 353-379.

Hespos, S. J. (2010). What is optical imaging? *Journal of Cognition and Development, 11*, 1-13. doi: 10.1080/15248370903453642

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd Ed.). New York: Guilford.

Little, T. D. (in press). Longitudinal structural equation modeling. New York: Guilford.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (in press). Why the items versus parcels controversy needn't be one. *Psychological Methods*.

Martinez, J. F., Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: Evidence from the ECLS. *Educational Assessment, 14*, 78-102. doi: 10.1080/10627190903039429

McArdle, J. J., & Woodcock, R. W. (1997). Expanding test–retest designs to include developmental time-lag components. *Psychological Methods, 2*, 403–435. doi: 10.1037/1082-989X.2.4.403

Minshew, N. J., Turner, C. A., & Goldstein, G. (2005). The application of short forms of the Wechsler Intelligence Scales in adults and children with high functioning autism. Journal of Autism and Developmental Disorders, 35, 45-52. doi: 10.1007/s10803-004-1030-x

Palmer, R. F., Graham, J. W., Taylor, B., & Tatterson, J. (2002). Construct validity in health

behavior research: Interpreting latent variable models involving self-report and objective measures. *Journal of Behavioral Medicine, 25*, 525-550. doi: 10.1023/A:1020689316518

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

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement and Evaluation in Counseling and Development, 42*, 223-232. doi: 10.1177/0748175609344096

Ready, D. D., & Wright, D. W. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*, 335-360. doi: 10.3102/0002831210374874

Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D. E., & Saxe, R. (2010). Live face-to-face interaction during fMRI: A new tool for social cognitive neuroscience. *NeuroImage, 50*, 1639-1647. doi: 10.1016/j.neuroimage.2010.01.052

Savalei, V., & Rhemtulla, M. (in press). On obtaining estimates of the fraction of missing information from FIML. *Structural Equation Modeling.*

Sciutto, M. J., Nolfi, C. J., & Bluhm, C. (2004). Effects of child gender and symptom type on referrals for ADHD by elementary school teachers. *Journal of Emotional and Behavioral Disorders, 12*, 247-253. doi: 10.1177/10634266040120040501

Stevens, J., Quittner, A. L., & Abikoff, H. (1998). Factors influencing elementary school teachers' ratings of ADHD and ODD behaviors. *Journal of Clinical Child Psychology, 27*, 406-414. doi: 10.1207/s15374424jccp2704_4

Swanson, J.M., Schuck, S., Mann, M., Carlson, C., Hartman, K., Sergeant, J.A., Clevinger, W.,

Wasdell, M., & McCleary, R. (2006). Categorical and dimensional definitions and evaluations of symptoms of ADHD: The SNAP and SWAN Rating Scales. Retrieved from http://www.ADHD.net.

Table 1

Longitudinal Three-Form Design

		Max Occasion Overlap			Max Question Overlap		
Group	Prop.	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
1	1/3	Form 1	Form 1	Form 1	Form 1	Form 2	Form 3
2	1/3	Form 2	Form 2	Form 2	Form 2	Form 3	Form 1
3	1/3	Form 3	Form 3	Form 3	Form 3	Form 1	Form 2

A. 3-form design				
Prop.	*Set X*	*Set A*	*Set B*	*Set C*
1/3	**X**	**X**	**X**	**O**
1/3	**X**	**X**	**O**	**X**
1/3	**X**	**O**	**X**	**X**

B. 6-form design					
Prop.	*Set X*	*Set A*	*Set B*	*Set C*	*Set D*
1/6	**X**	**X**	**X**	**O**	**O**
1/6	**X**	**X**	**O**	**X**	**O**
1/6	**X**	**O**	**X**	**X**	**O**
1/6	**X**	**X**	**O**	**O**	**X**
1/6	**X**	**O**	**X**	**O**	**X**
1/6	**X**	**O**	**O**	**X**	**X**

Figure 1. **X** = item set is administered; **O** = item set is omitted. Participants are randomly assigned to planned missing condition at the study onset.

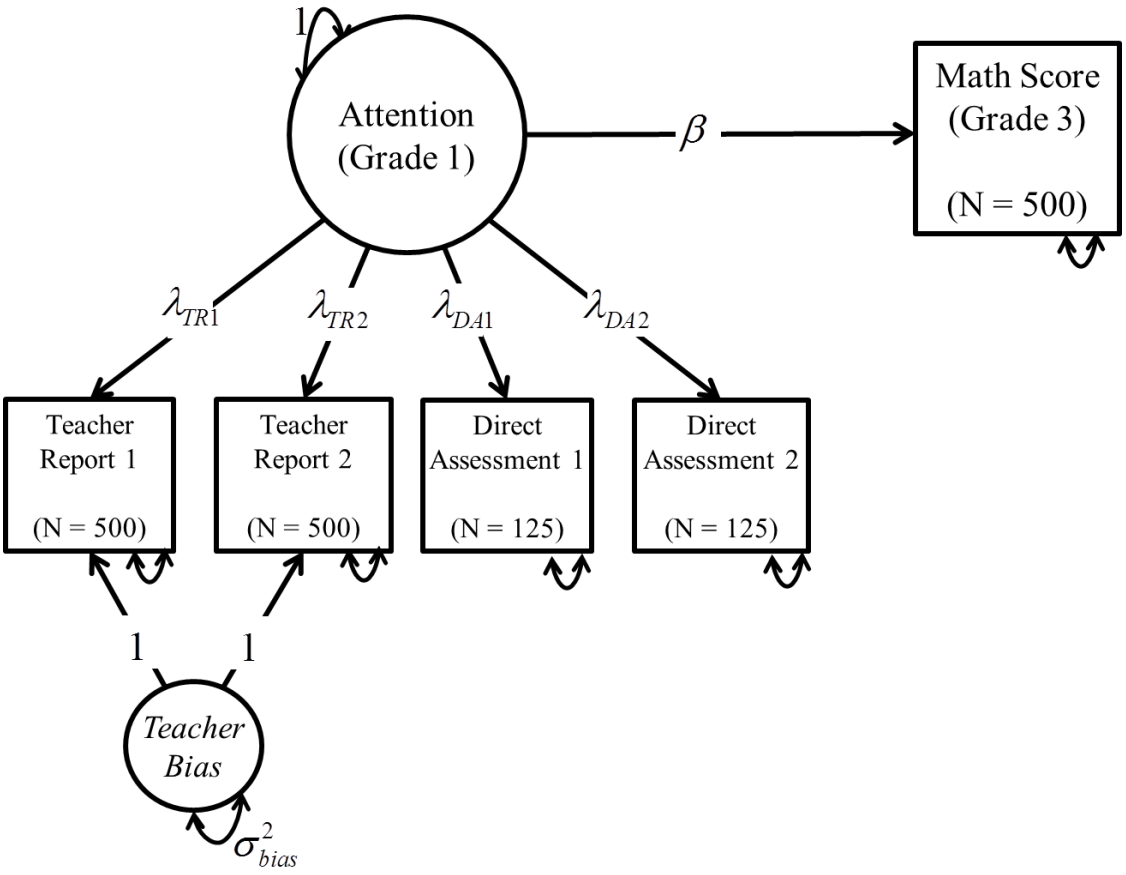

Figure 2. Analysis model for two-method measurement. Teacher reports 1 and 2 are two halves of the teacher report measure (e.g., 9 items each), and Direct Assessment 1 and 2 are two halves of the direct assessment measure (e.g., half of the observations each). The reason for splitting each assessment into two parts is to distinguish the reliable variance in each (i.e., that variance which is correlated across halves) from the unreliable variance. It is equally possible to use more than 2 indicators, and researchers may even choose to use as many indicators as there are items or observations. For practical reasons, it may be prudent to limit the number of indicators. To estimate bias, the two factor loadings of teacher rating on the latent construct *Bias* are fixed at 1 and its variance is freely estimated. This is equivalent to simply allowing the residual variance of the two teacher report items to correlate. In this design, complete data on all four observed variables are collected from 125 participants; the remaining 375 only provide teacher report data.