Published in final edited form as:

Eval Rev. 2004 October ; 28(5): 434–464. doi:10.1177/0193841X04264662.

Alternative Methods for Handling Attrition:

An Illustration Using Data From the Fast Track Evaluation

E. Michael Foster,
Pennsylvania State University

Grace Y. Fang, and
Pennsylvania State University

Conduct Problems Prevention Research Group

Abstract

Using data from the evaluation of the Fast Track intervention, this article illustrates three methods for handling attrition. Multiple imputation and ignorable maximum likelihood estimation produce estimates that are similar to those based on listwise-deleted data. A panel selection model that allows for selective dropout reveals that highly aggressive boys accumulate in the treatment group over time and produces a larger estimate of treatment effect. In contrast, this model produces a smaller treatment effect for girls. The article's conclusion discusses the strengths and weaknesses of the alternative approaches and outlines ways in which researchers might improve their handling of attrition.

Keywords

attrition; imputation; selection models; nonresponse

When asked, many program evaluators likely will name attrition as one of the greater threats to evaluation.¹ The deleterious effects are several. First, even under the best of conditions, attrition likely reduces the statistical power of the analyses of the available data (Orr 1999).² In addition, attrition may compromise the external validity of the study, especially in those circumstances in which the likelihood of response is related to observed characteristics, such as race or place of residence. In those instances, the resulting parameter estimates may generalize only to a subset of the population of interest (Orr 1999). Suppose, for example, that all African American participants drop out of an evaluation. In that instance, the study's findings likely generalize only to the remaining racial and ethnic groups in the study. This example is extreme; one generally does not lose all individuals of a given race or ethnicity (or of any subgroup, unless it is very narrowly defined). Rather, attrition is often linked to a range of

Correspondence concerning this article should be addressed to E. Michael Foster, Pennsylvania State University, 159 Henderson Building South, University Park, PA 16802-6500. For additional information concerning Fast Track, see <http://www.fasttrackproject.org>.

Conduct Problems Prevention Research Group members include, in alphabetical order, Karen L. Bierman, Department of Psychology, Pennsylvania State University; John D. Coie, Department of Psychology, Duke University; Kenneth A. Dodge, Center for Child and Family Policy, Duke University; E. Michael Foster, Department of Health Policy and Administration, Pennsylvania State University; Mark T. Greenberg, Department of Human Development and Family Studies, Pennsylvania State University; John E. Lochman, Department of Psychology, University of Alabama; Robert J. McMahon, Department of Psychology, University of Washington; and Ellen E. Pinderhughes, Department of Child Development, Tufts University

¹Note that this article addresses dropout from the evaluation rather than from the intervention itself. The empirical example represents an intention-to-treat analysis.

²Power is often reduced but not in every case; power might be improved if the most extreme (and least predictable) observations leave the study.

characteristics in a complex manner. As a result, the population to which study findings generalize can be difficult to identify or describe.

Most serious, certain forms of attrition also compromise the internal validity of a study (Little and Rubin 1987, 2002). Particularly problematic are those instances in which the likelihood of response is related to the values of the variable for which values are only partially observed. This sort of attrition seems possible in many cases. Consider, for example, an intervention targeting aggressive adolescents. If individuals who are incarcerated are less likely to participate in follow-up interviews, then aggression likely varies between those who do and do not participate. Especially worrisome are treatment-control differences in the direction or strength of these relationships. Suppose an intervention was effective in keeping highly aggressive youth out of jail; in that case, these more aggressive youth would be more likely to participate in the study if they were in the treatment group. Such an imbalance would make the intervention appear less effective.

The problems posed by attrition often are difficult to correct or even diagnose in many instances. Compounding matters is that until relatively recently, common practices such as mean imputation made implausible assumptions about the underlying processes determining response (i.e., the missing data mechanism) and potentially made attrition problems worse. Fortunately, attrition (and nonresponse more generally) is an active area of research in statistics. Recent developments include multiple imputation (MI), ignorable maximum likelihood methods, pattern-mixture models, and extensions of econometric selection models (Little and Rubin 1987, 2002; Schafer and Graham 2002). Growth in this area has been fueled not only by theoretical developments but also by advances in statistical computing. The latter have made available to nonmethodologists a wide array of tools previously only accessible to the methodological elite.

The multitude of new methods does create a problem: Applied researchers are left with a somewhat overwhelming array of alternatives for handling nonresponse in general and attrition in particular. Choosing among the alternatives can be difficult. As discussed below, the various methods often make different assumptions about the missing data mechanism. In many real-world instances, the researcher knows relatively little about the mechanisms shaping attrition and other forms of nonresponse.

In this article, we apply three of the newer methods (ignorable maximum likelihood methods, MI, and panel selection econometric methods) to data from the evaluation of the Fast Track (FT) intervention. FT is an ongoing, multisite, randomized trial designed to prevent the onset of serious conduct disorder and its concomitants (such as delinquency and substance abuse) in adolescence. FT targets young children with emergent behavior problems who are at greater risk for a lifetime of problems. Using a common measure of parent-reported aggression for children in Grades 3, 4, and 5, we estimate a simple longitudinal (mixed-effects) model using different procedures for handling attrition. We compare alternative parameter estimates, highlighting not the superiority of any single method over the other but the fact that each makes its own assumptions.

This article has four sections. First, we review alternative methods for handling attrition. The second section describes the FT study and the data used in these analyses. The third presents the results of our analyses, and the article concludes with a discussion of the study's implications.

Method

A major goal of any evaluation is to produce an estimate of program impact that is correct on average (i.e., unbiased) and as precise as possible (i.e., efficient). As sample size increases,

that estimate should be consistent: The confidence interval for the estimate should collapse to the true program effect. One determinant of whether a given analysis produces such an estimate is whether the way in which missing data are handled accurately reflects the nature of the missing data mechanism.

In this section, we review the different types of missing data mechanisms and then consider alternative methods for handling attrition (Allison 2002; Hall et al. 2001; Little and Rubin 1987, 2002; Schafer and Graham 2002; Verbeke and Molenberghs 2000). Our discussion focuses on the newer methods applied in our empirical example. Our review here is nontechnical. For a full review, see the references cited above and in the sections that follow.

Types of Missing Data Mechanisms

In the typical evaluation, a researcher may be interested in a range of variables. This range includes variables such as baseline characteristics that may be available for all cases (or nearly so). The impact of the evaluation may be tracked over time, and key outcomes may be available for a shrinking subset of participants. The missing data mechanism describes the process that generates the missing data. Methodologists classify missing data mechanisms into one of three categories: Data are said to be either “missing completely at random” (MCAR), “missing at random” (MAR), or “missing not at random” (MNAR) (Little and Rubin 1987, 2002; Schafer and Graham 2002). MCAR refers to a situation in which the likelihood of response depends on neither observed nor unobserved values of variables included in the analysis. This form of missing data represents a special case of simple random sampling (Allison 2002).

MAR refers to a situation in which the likelihood of response depends on observed values of the completely and partially observed variables.³ MAR is less restrictive than MCAR is. MCAR requires that the observed data represent a random sample of the complete data but only within subclasses defined by the observed values (Schafer 1997). MAR does not preclude the possibility that the likelihood of missing data is related to the partially observed variable; that dependency, however, must be explained entirely by the relationship between response and the observed data. For example, individuals lost to follow-up may have higher incomes than do those who are not. For the data to be MAR, however, that tendency must be explained by the relationship between other variables (such as race and education) that are observed. MAR requires that—for a specific level of education and racial group in this example—the observed data on income must be a representative sample of all data (Schafer 1997).

Many of the forms of missingness that concern evaluators are MAR (Foster and Bickman 1996). For example, researchers often consider whether the follow-up rate differs between the treatment and control groups or whether it depends on key outcome measures at baseline. Because treatment status and baseline scores are generally observed completely, even when these relationships exist, they represent forms of MAR. As discussed below, there are a variety of methods for handling missing data that are MAR.

Statisticians refer to the first two types of nonresponse as “ignorable.”⁴ In this context, ignorable means that likelihood-based inferences about the parameters of interest can proceed

³There are finer grained distinctions among these categories than are presented here. For example, one can distinguish missing at random (MAR) models according to whether the likelihood of nonresponse depends on past values of the outcome of interest or just on other covariates (such as race or age). This distinction matters for some statistical methods, such as generalized estimating equations (Hall et al. 2001).

⁴The assumption of ignorability also requires that the parameters of the model predicting nonresponse not include any of those in the model predicting the outcome of interest (Schafer and Graham 2002). This “distinctness” condition generally holds but would need to be evaluated in any particular application. In any case, of the two conditions for ignorability, MAR is the more important—if the distinctness condition does not hold, inference based on the ignorable likelihood is still valid but not fully efficient. See Little and Rubin (2002, 120).

without regard to the missing data mechanism. In particular, one can analyze the likelihood function for the available data as if it were the complete-data likelihood (i.e., the likelihood function that describes the likelihood of response as well as missing and observed values of the variables of interest). The resulting parameter estimates pertain to the population as a whole, including the cases for which data are missing (Schafer 1997).

Under the third type of missing data mechanism, MNAR, the likelihood of response can depend on both observed and unobserved values of the outcome. In that case—if complete data were available—the values of the partially observed variable would differ between the cases for which data are and are not missing. In other words, even conditioning on observed characteristics, the cases for which the partially observed variable is available are not a representative sample of all observations. Individuals lost to follow-up may have higher incomes (were it observed) than do those who are not, even allowing for differences between the groups in race and education.

Different analytical methods make different assumptions about the missing data mechanism. Perhaps most common is listwise deletion in which the analyst drops incomplete observations from the analysis. For simple analyses (such as descriptive statistics), listwise deletion assumes the data are MCAR. For multivariate analyses (such as regression), listwise deletion assumes the data are MAR—in particular, that all relevant, systematic determinants of missingness are included as covariates. If these assumptions are not correct, then analyses of the listwise-deleted data may produce misleading estimates of treatment impact.

In any given situation, the actual missing data mechanism is unknown. However, as discussed below, the evaluator can assess the plausibility of the alternative assumptions based on what he or she knows about the evaluation and the population included and what they reveal about how the missing data were generated. An understanding of the types of missing data mechanisms is also essential for understanding the applicability of the particular analytical method used.

Recent Improvements in Methods for Handling Attrition

Recent methodological developments improve on prior methods (such as listwise deletion) in two ways. First, these methods often make better use of all available data. Even if the MAR assumption is correct, listwise deletion may discard large amounts of data, often reducing the precision with which key parameters are estimated. The first two methods employed in the empirical example below—maximum likelihood estimation involving the ignorable likelihood function (IML) and MI—represent alternative means of incorporating incomplete cases more fully into analyses. Second, recently developed methods may be valid under a broader range of missing data mechanisms, such as MNAR. The third method employed below—the panel selection model—relaxes the MAR assumption or, more precisely, replaces that assumption with other assumptions. In this subsection, we briefly describe each of these methods.

IML—The principle underlying the first method—IML⁵—is simple: In calculating the likelihood function used to estimate model parameters, each observation contributes all available information (Enders 2001).⁶ For example, in the case of panel data, an individual might contribute those waves for which he or she participates but add nothing for those waves

⁵These methods are described using various terms. Schafer (1997) described them as maximizing the “observed data likelihood” (p. 12).

⁶The intuition behind the maximum likelihood estimation involving the ignorable likelihood function (IML) is like that behind the pairwise-deletion option in structural equations software in which each element of the covariate matrix is calculated using all available observations for a given pair of variables. Pairwise deletion still assumes the data are missing completely at random and often results in computation problems, especially in small samples (Allison 2002; Enders 2001). Under IML, these problems are possible but less likely.

for which he or she is absent from the study. Under the MAR assumption, the parameter estimates have good statistical properties.

The advantage of this approach is that it is largely transparent to the user: Changes in the way the likelihood function is calculated are performed behind the scenes and require no additional steps by the analyst. Furthermore, in many cases, this method is used by default (e.g., “Proc Mixed” in the Statistical Analysis System [SAS]). However, because various software packages implement the estimation procedure differently, the user must carefully read the manual for the particular software used. For example, Mplus (Version 2) can handle missing covariates, whereas Proc Mixed cannot (Muthen and Muthen 2001; Littell et al. 1999; Verbeke and Molenberghs 1997). This discussion highlights a limitation of IML: The analyst is limited to those models (and forms of missing data) for which the necessary likelihood functions are programmed. Mplus, for example, includes ordinary multiple regression but not logistic regression (Muthen and Muthen 2001).

MI—One traditional option for dealing with missing data is to replace each missing value with a plausible estimate. The appeal of this method is obvious: The researcher then can analyze the filled-in data using any complete-data method. This strategy includes mean imputation, in which the missing values are replaced with the mean value for the observed cases. This practice, however, does not preserve the relationship between the filled-in variable and the other variables in the analysis. In the case of an evaluation, substituting the mean of the outcome of interest for all missing cases—regardless of their treatment status—would obviously bias estimates of treatment impact toward zero.

For that reason, various forms of conditional mean imputation have been developed. For example, the missing value of a partially observed variable for a given individual could be replaced with the mean for observed cases matched on observed characteristics, such as race, education, or treatment status. One means of implementing this strategy would involve a regression estimated using the complete cases. The missing data would be replaced by an appropriate predicted value using the estimated regression coefficients. Although an improvement over mean imputation, this method still suffers from a major shortcoming: The imputed value does not reflect the residual variation in the partially observed variable. As a result, the variance in the filled-in variable would be too small. Furthermore, the covariance with the other variables may be distorted as well. In particular, among the imputed cases, the variables used as explanatory variables in the imputation regression would predict the imputed values perfectly. As a result, the strength of the relationship between those variables would be exaggerated.

For these reasons, researchers have attempted to improve conditional mean imputation by adding a random draw from an appropriate error distribution to the predicted values. These draws may be taken from a normal distribution based on the estimated variance of the regression residuals or from the empirical distribution of the actual residuals. Adding residual variation in some form preserves the variance of the filled-in variable as well as its relationship to other variables. This practice still suffers from limitations, however. Analyses of the completed data do not reflect the fact that the filled-in data were estimates.

MI represents a response to this problem. Under MI, a researcher generates several filled-in data sets and analyzes each separately. The researcher then combines the results of the analyses. In particular, point estimates such as regression coefficients are calculated as the average of the parameter estimates generated using each imputed data set. The variance of these estimates is calculated as a weighted sum of the average of the variances from each imputed file and the variation in the parameter estimates across imputations. (No covariance terms are included because the imputed values are effectively independent across imputations.) The between-

imputation variation is critical: It captures the uncertainty created by the fact that the filled-in data were originally missing (Schafer 1997, 1999; Schafer and Olsen 1998).

In most instances, only a relatively few imputations (5 to 10) are required (Allison 2002; Schafer 1997). The methods for combining the parameter estimates and for generating the estimates of their variance are widely available (e.g., StataCorp 2001; SAS Institute 2001a, 2001b).

MI can incorporate any of several imputation strategies, including the regression-based method described above. The actual method employed depends on the pattern of missing data. For complex patterns of missing data, one might use an iterative series of regressions. One might use a regression to impute one missing variable using all available information and then use that imputed variable in a regression imputing other missing variables.⁷ One might cycle through a series of regressions, iteratively filling in and updating estimates of the missing data.

An alternative (used in the analyses below) involves the estimation of a multivariate model using Markov Chain Monte Carlo (MCMC). The particular form of MCMC used is known as data augmentation (DA). DA involves iterating back and forth between random draws of the model's parameters from their posterior distribution conditional on the observed and imputed data and random draws of the missing data conditional on estimates of model parameters. Because it involves convergence to a joint statistical (posterior) distribution of missing values and model parameters rather than parameter estimates, DA raises a set of rather technical issues. (For more details, see Schafer 1997.) An added advantage of both the sequential generalized regression and the MCMC approach is that the resulting standard errors of the parameters of the analytical model reflect the uncertainty associated with the parameters of the imputation model.

One advantage of MI is that the imputation model can be quite general. The only restriction on the choice of variables used in the imputation stage is that the analytical model not be more general than the imputation model. For example, one cannot add variables to the analysis that were not included in the imputation model. If a variable is added at the analysis stage, the relationships between that variable and the others are not preserved in the imputation stage, and estimates based on that relationship (e.g., regression coefficients) in the imputed data are potentially biased. As an example, this problem might occur when the analysis model includes interactions between two variables when those interactions were not included in the imputation stage. For this reason, researchers interested in treatment moderation should impute data separately for treatment and control groups and combine the imputed data. Doing so preserves all interactions between treatment and other variables in the model.

On the other hand, the imputation model may be more general than the analysis model—the researcher intentionally may include variables in the imputation model that are excluded from the analysis model (Collins, Schafer, and Kam 2001).⁸ For example, time of study entry might be related to severity and so might be useful for imputing data. However, the researcher might have little interest in the impact of that variable on the outcome and so may exclude it from the analytical model. By separating the analysis and the imputation models, MI makes it easy to include variables in the latter but not the former.

⁷This method illustrated a category of methods known as sequential generalized regression models. The software MICE (multivariate imputation by chained equations) represents one means of implementing this method.

⁸For example, in an analysis of child behavior problems, Foster (2002) included information on the use of services from community service providers in the imputation model. This information was excluded from the analysis model because service use is potentially endogenous and because including it would change the interpretation of the coefficient of interest (the impact of an innovation in service delivery) from total to direct. Recently, structural equation modeling has been extended to allow one to incorporate missing-data-relevant variables (Graham 2003).

Like IML, MI assumes the data are MAR. Methodologists disagree as to how plausible this assumption is, and this issue is difficult to resolve in general. Whether MAR holds may depend on the population being studied and other characteristics of the evaluation. We return to this issue below.

Panel selection models—A second area of methodological innovation involves efforts to relax the MAR assumption (Allison 2002; Little and Rubin 1987; Schafer and Olsen 1998; Verbeke and Molenberghs 2000). These methods build on selection models common in econometrics and involve specifying a model for the outcome of interest and one for participation in the research study (Heckman 1974, 1976, 1979).⁹ These models have been extended to longitudinal applications (Verbeke and Nijman 1992; Lillard and Panis 1998) and involve maximum likelihood estimation in which every individual contributes the probability of participating in the study (or not, as appropriate) for a given wave of data collection to the likelihood function. Individuals who participate contribute the probability density function for the observed value of the outcome. As spelled out in more detail below, the key feature of the model is that the likelihood of participation depends on unobserved determinants of the outcome. Such a relationship violates the MAR assumption (Allison 2002; Verbeke and Molenberghs 2000).

Typical is the following model derived from Lillard and Panis (1998):

$$Y_{i,t} = \beta_0 + \beta_1 T_i + B'X_{i,t} + \mu_i + \varepsilon_{i,t}, \quad (1)$$

where i indexes individuals and t indexes time. Equation 1 is a standard random-effects (or mixed-effects) longitudinal model (Davidson and MacKinnon 1993). Y is the outcome of interest, which varies over time (e.g., aggression in the example below). μ and ε represent time-invariant and time-varying unobserved or unmeasured characteristics that affect the outcome of interest, respectively. Both are assumed to be normally distributed.

Equation 2 models the likelihood of participation and represents the standard probit specification (Agresti 2002; Greene 1993; Wooldridge 2002).

$$R^*_{i,t} = \gamma_0 + \gamma_1 T_i + \Gamma Z_{i,t} + (\lambda^C + T_i \lambda^T) \mu_i + \delta_{i,t}, \quad (2A)$$

$$P(R_{i,t}=1) = P(R^*_{i,t} > 0), \quad (2B)$$

$$P(R_{i,t}=1) = 1 - \Phi(-\gamma_0 - \gamma_1 T_i - \Gamma Z_{i,t} - (\lambda^C + T_i \lambda^T) \mu_i), \quad (2C)$$

where R indicates participation at a point in time (1 = participation; 0 = not) and is assumed to be a function of an unobserved, continuous variable R^* . R^* is a function of covariates and a standard normal error term (δ). (See Equation 2A.) Individuals participate when R^* is positive and do not otherwise (Equation 2B). Given the distribution of δ and Equation 2B, the probability of participation can be specified as in Equation 2C.

⁹Another method is pattern-mixture models, which we discuss briefly.

In our example below, Equations 1 and 2 have the same explanatory variables—treatment status, race, grade, and cohort. Our focus here is on the first of these, and so we represent treatment status with a separate variable. The other covariates are represented by the vectors X and Z . Ideally, some variables would be included in Z that are not included in X . In their original article, Lillard and Panis (1998) included the number of phone calls required to interview a subject at the preceding wave as a (unique) determinant of nonresponse. The methodological literature indicates that identification can be a problem in cross-sectional models when X and Z comprise the same variables (Allison 2002; Little and Rubin 1987, 2002; Stolzenberg and Relles 1990, 1997). In that case, identification is achieved only through the functional form or distributional assumptions. For that reason, analyses of the robustness of model parameters to distributional assumptions are particularly important. We provide such analyses below.

As discussed above, the key feature of the model is that Equations 1 and 2C are interrelated: μ , the time-invariant unobservable, affects the likelihood of participation. This feature of the model allows individuals with high or low scores on the outcome measure to be more or less likely to participate in the study. This effect is captured by the λ parameters. Two λ s are included because we have allowed the effect of μ to vary with treatment status.¹⁰ λ^C represents the effect of μ in the control group; λ^T is the interaction between μ and treatment status (T).

Two additional features of this model should be mentioned. First, the model allows for MNAR but not without replacing the MAR assumption with others. In particular, as with all maximum likelihood methods, distributional and functional form assumptions are critical to producing parameter estimates. The model assumes that the outcome measure of interest is normally distributed, and parameter estimates may be sensitive to this assumption (Little and Rubin 1987, 2002).

Second, this model specifies one form of nonignorable nonresponse, but others are possible. For example, one might include a random component to the trend over time and allow that unobservable to affect participation. Furthermore, one might assume a different structure altogether. For example, Diggle and Kenward (1994) specified a model in which the current value of the partially observed outcome variable affects the likelihood of response; that is, $Y_{i,t}$ appears on the right-hand side of an equation like Equation 2.

Although these models are computationally intensive, software is increasingly available that can be used for estimation. The aML package, for example, can be used to estimate the model described above (Lillard and Panis 2000).

Data: The FT and Evaluation

In the next section, we compare and contrast results from analyses that use each of these three methods (MI, IML, and panel selection methods). Before presenting those results, we briefly describe the FT study as well as the outcome measure that is the focus of these analyses. We also describe other variables used in the analyses as well as the level and nature of attrition in the study.

Overview of FT Intervention

The FT intervention is the focus of an ongoing, 15-year evaluation being conducted at four sites—Nashville, Tennessee; rural Pennsylvania; Seattle, Washington; and Durham, North Carolina. The evaluation focuses on the experiences of 891 high-risk children living in high-poverty areas. Because the study involved a classroom intervention, entire elementary schools

¹⁰In actual estimation, we estimated the model as a two-group model, allowing for heteroscedasticity in the variances of ε and μ .

($N = 54$) were assigned to either the intervention or the control conditions. To improve the balance of key characteristics across the treatment and control groups, data on the demographics for each school were obtained (e.g., size, percentage of students who received free or reduced-price lunch, ethnic composition, achievement scores), and within each site the participating school sites were divided into matched sets. These sets were randomly assigned to intervention and control conditions (Conduct Problems Prevention Research Group [CPPRG] 2002b; Lochman and CPPRG 1995).

In intervention schools, teachers were trained to present a classroom curriculum promoting emotion recognition, social understanding, and self-control, and they were provided with weekly consultation about classroom management. In addition, using information from teachers and parents, high-risk children were identified for participation in targeted intervention components. These included parent training and home visits as well as social skill training, tutoring, and friendship enhancement in the classroom. Those components were targeted to children according to their strengths and problems, and components such as mentoring were added or removed as participants aged. The analyses presented here focus on the experiences of these high-risk children.

Evidence to date suggests that FT has produced real benefits for its participants. By Grade 3, intervention children were 17% less likely to demonstrate serious conduct dysfunction (66% vs. 55% for the control and treatment groups, respectively). Teacher ratings of conduct problems and official records of special education suggested that the intervention was preventing problem behavior at school. Parent ratings revealed reductions in conduct problems at home. Intervention effects also were apparent for parenting behavior and children's social cognitive skills, variables that prior analyses identified as mediators of treatment impact (CPPRG 1999a, 1999b, 2002a, 2002b, 2002c).¹¹

One of the key outcomes in cross-sectional analyses has been the Parent Daily Report (PDR), which we describe in the next subsection along with the data collection more generally.

Data Collection

As part of the ongoing evaluation, FT collects information from several sources each year. Interviews are conducted each summer with study children and their families. These interviews provide information on a wide range of topics, including family demographics, socioeconomic status, and family functioning. The summer interview includes a common measure employed in prevention research involving aggression in young children, the PDR. The measure asks parents about their child's behavior over the past 24 hours (Chamberlain and Reid 1987). This information is typically collected by phone on multiple occasions. The behaviors involved cover a broad range of behaviors; for the purposes of these analyses, we focused on 22 items related to aggressive and oppositional behavior. A total score was calculated by summing these items and averaging across three administrations (the summer in-person interview and two subsequent phone administrations).¹² For the analyses below, we examined PDR data from Grades 3, 4, and 5.

¹¹The estimate of intervention effects presented in this article is based on a single measure. Analyses of a fuller range of outcome measures can be found in the articles cited. It is important to note that, analytically, children were nested within classrooms. Because observations on children in the same classroom may not be independent, conventional standard errors may be incorrect. We have explored this possibility in great detail, but the effect here is likely very small. If one thinks of the nesting as involving the original kindergarten classrooms, then by the time children reach Grade 5, the intraclass correlation is very small. On the other hand, if one thinks of the clustering as involving the Grade 5 classroom, then the intraclass correlation is larger. However, it is important to note that by that point, the participants are dispersed across more than 500 classrooms. As a result, the "misspecification effect" (or MEFT) is very small (Scott and Holt 1982). (The MEFT is the ratio of the true [cluster-adjusted] standard error to one calculated under the assumption that observations nested in the same classrooms are independent.) For example, the MEFT for the Grade 5 Parent Daily Report is 1.00, suggesting that the standard errors calculated under an assumption of independence are essentially correct.

As discussed above, information from additional sources can be used in the imputation model, and FT data are a rich source of such information. This includes teacher reports of the child's behavior at school, such as the Teacher Report Form (TRF), a teacher version of the Child Behavior Checklist (Achenbach 1991); the child's own report of delinquent and aggressive behavior, such as the "Things You Have Done" (TYD) measure (Elliott, Ageton, and Huizinga 1985; Elliott, Huizinga, and Menard 1989); and peer reports of the child's behavior. The last evaluates peers' perceptions of children in their classrooms across a variety of dimensions (Coie, Dodge, and Coppotelli 1982; Asher and Dodge 1986; Terry and Coie 1991). Our analyses incorporate these measures as well as supplemental information on special education obtained from school records.

Levels and Patterns of Attrition

Table 1 describes levels of missing data and the relationship of missing data to treatment status. One can see that rates of attrition were generally low. By Grade 5, between 85% and 90% of participants were still in the study. Furthermore, PDR data were available for an even higher percentage of children for at least one of the three grades—no more than 9.3% of individuals were missing data at all three grades. At the same time, a substantial percentage were missing at some point. That percentage is highest for girls in the control group—fully one quarter of this group (24.5%) were missing PDR data for at least one grade.

Comparisons of the treatment and control groups revealed striking gender differences. For boys, the two groups had similar rates of attrition. For girls, however, the rate of attrition was substantially higher in the control group: The percentage of girls missing all three waves of data was almost 4 times that for the treatment group (9.3% vs. 2.4%, respectively).

Table 2 compares individuals with and without the PDR at Grade 5.¹³ Given the gender differences in Table 1, these breakdowns are provided for girls and boys separately and disaggregated by treatment status. For both treatment and control boys, one can see that African Americans were overrepresented among those for whom PDR data were available. This somewhat surprising finding largely reflected site variation in attrition (African American participants were concentrated at Nashville and Durham, where the response rate was higher). One can see that the baseline PDR was not related to attrition.

The figures for girls highlight further gender differences. First, one notes that the relationship between race and attrition was no longer apparent. Furthermore, although the differences were not statistically significant, girls who were lost to follow-up showed lower levels of aggression at baseline; this tendency was especially pronounced in the control group.

As noted above, a strength of the MI approach is that one can include additional variables (such as information from alternative sources) in the imputation model. Table 3 describes the availability of these data in this study. Taking the TRF as an example, one can see considerable overlap in data sources. Children with PDR data were considerably more likely to have TRF data. Still, it is striking that TRF data were available for a substantial percentage of children without PDR data. For boys, that percentage was roughly half and did not depend on treatment status (48.8% and 51.2% for the treatment and control groups, respectively). Other gender differences were apparent. Nearly 7 in 10 treatment girls lacking PDR data at Grade 5 had TRF

¹²Three administrations were collected for most children (98% in Grade 5, for example). In a small minority of cases, the average score was calculated across available administrations. Reports at different administrations are highly—but not perfectly—correlated. For example, for Grade 5 reports included in the analysis below, reports at different administrations are correlated in the neighborhood of .6.

¹³Tables for Grades 3 and 4 are similar and are available from the authors.

data. For control girls, that figure was only 22%. Similar differences were found for school record data (used to determine whether the child was enrolled in special education).

In the next section, we present parameter estimates for our outcome model: the mixed model specified in Equation 1 above. Given the observed patterns in missingness, we estimated the model separately for boys and girls.

Results

Table 4 presents the estimated treatment effect (β_1) for alternative models. The first row in the table is provided for comparison purposes; it involves analyses of listwise-deleted data (i.e., of individuals who were present for all three waves). In that model, for both boys and girls, the estimated impact of the intervention was negative (i.e., reduces aggression). The treatment effect for girls was substantially greater than that for boys (in absolute terms). The decision to estimate the model separately for boys and girls was straightforward in light of the patterns of missing data described above. However, we also conducted a Chow test to determine whether the data should be pooled. The null hypothesis that the data should be pooled across genders (relative to a common model with gender included only as a covariate) was rejected at $p < .01$.

Note that although the table presents only the estimated treatment impact, the vector X included grade, race, three dummy variables representing study site, and two dummy variables representing the three cohorts. (The full set of parameter estimates is available from the first author.) Although we do not dwell on their impact, the inclusion of these covariates is important, especially given their relationship to the patterns of missing data.

Estimates from IML

Row 2 presents the results from an IML model estimated using maximum likelihood estimation as implemented in SAS Proc Mixed. That model takes advantage of all available waves of data collection. One can see that those results were very similar to the listwise-deletion results. As one might expect, the larger sample size lowered the standard error a bit, but the reduction was very modest.

Estimates from MI

Rows 3 to 5 present the MI results. We conducted three sets of analyses. Each was based on 10 imputations. For the first, the only variables included in the imputation were the covariates in the analytical model (Equation 1): grade, race, site, and cohort. In the second, we added PDR scores for kindergarten through Grade 2. In the third and final model, we included the additional variables described above (TRF, peer ratings, TYD, and special education). We proceed in steps to illustrate the value of the data from supplemental sources in imputing the missing data.

Looking first at row 3, one can see that the point estimates (and standard errors) generated by MI were very similar to those in the first two rows. Somewhat surprisingly, adding the PDR from earlier waves and then data from other respondents made very little difference (rows 4 and 5, respectively). If anything, there was a very slight tendency for the gap between genders to narrow. The effect of the intervention for both boys and girls was marginally significant ($p = .07$ and $.08$, respectively).

Estimates from Panel Selection Model

Row 6 in each panel of Table 4 presents the results of the panel selection models (Equations 1-2). (The appropriate aML code is included in the appendix.) For these models, we present three parameters: the effect of treatment (β_1), the selection effect for the control group (λ^C), and the treatment-selection interaction (λ^T). As with the models above, the other covariates

were included in the analyses, and the corresponding parameter estimates are available from the first author.

Looking first at the lambda parameters, one can see that the likelihood of response did depend on unobserved (time-invariant) determinants of the outcome. The nature of the selection mechanism, however, differed dramatically by gender. Looking first at boys, one can see that control boys with higher levels of aggression were less likely to participate in the study ($\lambda^C < 0$). For girls, however, the effect was reversed: More aggressive girls were more likely to participate ($\lambda^C > 0$). Looking at the treatment by selection interaction, one can see that the treatment group was more representative of the original samples for both boys and girls. For girls, the net effect for the treatment group was still positive ($\lambda^C + \lambda^T = .39 - .21 = .18$). Aggressive girls were still overrepresented in the treatment group, but this tendency was muted relative to that for the control group. For boys, the selection effect was actually reversed in the treatment group (relative to control boys): Aggressive boys were actually overrepresented ($\lambda^C + \lambda^T = -.19 + .65 = .46$).

This model produced a rather different estimate of the impact of treatment, especially for boys. Adjusting for the apparent overrepresentation of the most aggressive girls in the control group reduced the estimated impact of treatment ($\beta_1 = -.34$). On the other hand, allowing for the apparent overrepresentation of the most aggressive boys in the treatment group increased the estimated effect of treatment for boys ($\beta_1 = -1.04$). The effect for boys represents a modest effect—roughly one third of a standard deviation.

It is worth noting that for both boys and girls, the estimate in row 6 differed from that estimated under a model that ignores the selection mechanism. Using the appropriate likelihood ratio test, we rejected the null hypotheses that λ^C and λ^T were 0 at $p < .001$ and $p < .08$ for boys and girls, respectively. Of course, this test is specific to the functional form and distributional assumptions embedded in the model. Furthermore, this finding does not imply rejection of the MI estimates. That model is not nested within the panel econometric model; the probit specification for participation is not required in the MI framework.

Sensitivity Analyses

Like the MAR approaches, selection modeling depends on assumptions, and critics of this approach highlight the importance of those assumptions (Laird 1994; Little and Rubin 2002; Schafer and Graham 2002). An example of this sensitivity can be found in Kenward's (1998) reanalysis of the original Diggle and Kenward (1994) data. In the original analysis, the authors rejected MAR—they found that unobserved values of the dependent variable affect the likelihood of response. Kenward, however, determined that the findings were quite sensitive to distributional assumptions and to the handling of two outlier cases. In particular, if the two cases were removed or an alternative distribution was used for the outcome of interest, a MAR model fit the data as well as the MNAR model fit.

As a result, we performed two sensitivity analyses.¹⁴ First, we identified a number of observations that the model fit poorly: These were 16 cases in which the model grossly underestimated the child's actual level of aggression. Dropping these cases had little effect on the parameters of the MNAR model.

¹⁴Of course, this list of sensitivity analyses is far from exhaustive. For example, one might simulate the data that result from alternative selection mechanisms (Rosenbaum 2002). It is also important to keep in mind that the analyses rest on more assumptions than just those involving the process that shapes attrition. These include the single-unit value treatment assumption, other aspects of functional form, as well as the distribution of the random effects. A full set of sensitivity analyses might consider the robustness of the results to these assumptions as well. For more discussion of these issues, see Cook and Weisberg (1999).

A second set of sensitivity analyses involved the model's distributional assumptions. Although unimodal, the PDR data were far from normally distributed: The most common value was 0, with higher levels of aggression becoming increasingly less common. That the assumption of normality would fit the data so poorly was a cause for concern, especially given the potential importance of distributional assumptions to the stability of the selection model (Allison 2002; Little and Rubin 1987, 2002; Stolzenberg and Relles 1990, 1997).

For this reason, we considered an alternative model in which the outcome was modeled using the binomial distribution (Wooldridge 2002). This model was appropriate given that the number of items potentially endorsed by parents was fixed. The results of those analyses are available from the first author and were similar to those reported above. The effect of treatment was significant for boys but not for girls; although both effects were larger in absolute terms, the effect for boys was roughly double that for girls. (In terms of effect size, the effect for boys represented roughly 0.8 standard deviations.) The selection term for the control group, λ^C , was statistically significant for both boys and girls. We also found that the interaction between treatment status and the selection mechanism (λ^T) was statistically significant in the model for girls.

The results of these supplemental analyses differed from those reported in Table 4 in two ways. First, although the results still suggest that aggressive boys were underrepresented in the control group, we can no longer reject the null hypothesis that the selection mechanism is the same for the treatment and control groups. A second difference is that the estimated treatment effect was much less sensitive to the modeling of nonresponse. In particular, the effect of treatment in the MNAR model was similar to that in the model that restricts the two lambda parameters to 0. The effect of nonignorable non-response on the estimated treatment effect was much smaller in the binomial model.

Conclusion

Using data from the FT evaluation, this article examined the effects of that intervention under alternative methods for handling attrition. We found that—in the context of the outcome examined and the statistical models employed—the estimate of treatment impact varies across the missing data methods considered. In particular, the results of the panel selection model stand in fairly sharp contrast to the IML and MI results. Estimates of treatment impact under MAR and MNAR differ according to both statistical and practical criteria.

The results for girls are quite striking and suggest that the missing data mechanism differs by gender. Whether such variation is common or an idiosyncratic feature of these data is hard to assess. Other studies typically do not describe analyses of missing data in detail sufficient to determine whether the nature of attrition varies by gender.

The variation in estimated treatment effect across methods for handling attrition is somewhat surprising in light of the high follow-up rates maintained by the FT evaluation. Using the criteria by which attrition problems are often assessed (Foster and Bickman 1996; Orr 1999), FT earns a relatively clean bill of health. It seems likely that key parameter estimates may even be more sensitive to the handling of missing data in studies in which attrition or other forms of nonresponse are greater (Schafer 1997). These findings—and the reading of the methodological literature more generally—suggest that program evaluators should remain vigilant in minimizing attrition.

Does the variation across methods presented above mean that one method has triumphed over the others? In particular, do these findings imply that MAR is implausible in general or even for these data? The answer is no. The competing models really represent a choice among alternative assumptions. Some readers may find MAR implausible, but it is important to note

that the panel selection model depends on key assumptions. Foremost among these is normality (Little and Rubin 1987, 2002). This assumption is somewhat troubling in the context of the outcome measures used in prevention research, which often have skewed or otherwise strange distributions. In these particular analyses, the parameter estimates did not appear especially sensitive to the assumption of normality, but this sensitivity would have to be assessed on a study-by-study basis.

How, then, does a researcher pick a method for handling attrition? Understandably, the literature does not identify a method that is the “best” for all data, analytical techniques, or research questions. The only way to determine whether a model is superior to another is to know the underlying missing data mechanism. Outside of Monte Carlo studies, the actual mechanism is unknown, and so in studies of actual data, the researcher is left to rely on other criteria, such as disciplinary preference. Those preferences involve tolerance for different types of assumptions or model complexity and, more generally, reflect how a discipline views the world as working. Economists, for example, are suspicious of the MAR assumption, and this suspicion reflects their view of human behavior. In a world in which agents act rationally and are often better informed than are researchers, the processes governing attrition and the outcomes of interest seem likely intertwined. For example, decisions to participate in a research study may reflect the value a potential participant places on his or her time: Persons with less free time may be less likely to participate. As a result, participation is tied to a whole range of choices surrounding time use, such as employment and child care. Attrition is then linked to other individual characteristics, such as education or even mental health. In intervention research (in which participants may perceive a relationship between research participation and receipt of services), such relationships may be even more complex.

At the same time, other methodologists find the assumptions necessary for MNAR models harder to believe or even assess. For that reason, many statisticians have eschewed these models, fearing that this putative cure for attrition problems may be worse than the illness (Little and Rubin 1987, 2002). When dealing with nonignorable nonresponse, these methodologists often prefer pattern-mixture models, which involve estimating separate models for subgroups defined by varying patterns of missing data (Little and Rubin 1987, 2002). These methods also rely on largely untestable assumptions. In particular, some elements of the model are clearly not identified: The data, for example, provide no information about trends over time for the subgroup of individuals who participated in only a single wave of data collection. Some assumption has to be made about the trend over time for that group and how it relates to the time trends for other groups (e.g., that the time trend is proportional to the number of waves in which an individual participates). These assumptions are no less arbitrary than are those embedded in the selection models and may heavily influence estimates of key model parameters. Still, according to its advocates, the pattern-mixture approach is preferred because the assumptions on which it rests are more transparent than are the distributional assumptions or exclusion restrictions on which the selection models rely.

If methodologists cannot give the applied researcher definitive guidance, can he or she determine the right answer based on the data involved? Unfortunately, there is no single statistical test that can identify the best method for handling attrition. Indeed, one can always find aMAR model that fits the data as well as one that employs other assumptions (Verbeke and Molenberghs 2000). The best the applied researcher can do, therefore, is to consider the robustness of his or her findings. As illustrated in this article, one can proceed in two steps. The first is to examine the variability of key findings across alternative models. If the key findings vary little, the researcher can relax and turn to another analysis or project.

However, if as illustrated above, the different approaches produce different results, then the analyst must pick the best model based on other information. The validity of the MAR

assumption will depend on key features of the study. Relevant considerations include the population being studied, the research design, and what they suggest about the missing data mechanism. For example, how likely are potential participants to be incarcerated or otherwise institutionalized, and are interviews conducted in those facilities?

The assessment of MAR could be improved if researchers made a better effort to document the circumstances surrounding attrition. For families unwilling to participate, address information could be used to determine whether and where they have moved. Such information would be useful in determining whether nonresponse is related to residential mobility and thus perhaps to the outcomes of interest. Other information might involve the characteristics of prior interviews, such as their length. Information of this type might be incorporated in the analyses themselves. For example, in a selection model, interview characteristics could be used to predict subsequent nonresponse (Lillard and Panis 1998). As noted above, these models perform best when the variables predicting the outcome of interest and nonresponse do not overlap completely. Information of this type can improve MI as well. Schafer and Graham (2002), for example, suggested that researchers ask respondents how likely they were to participate in future interviews. Such information could be included in the imputation model and could increase the plausibility of the MAR assumption.

In conclusion, the problem of attrition remains a difficult one. Although this article does not identify the single best way to handle attrition, it does illustrate alternative ways in which analysts can consider and evaluate the sensitivity of their key findings to the handling of nonresponse. The methods considered here represent only a partial list of available alternatives; as discussed above, for example, pattern-mixture models represent an added way to relax the MAR assumption. The inventory of possible methods will continue to grow as methodology improves. Our hope is that evaluators will continue to expand the set of methods they use and that they will report the results of alternative analyses. As a result, our knowledge of the sensitivity of key findings to the handling of attrition will grow as well. Over time, evaluators may develop a more general sense of when and how the handling of attrition matters and when key results are likely insensitive.

Acknowledgments

This work was supported by National Institute of Mental Health (NIMH) Grants R18 MH48043, R18 MH50951, R18 MH50952, and R18 MH50953. The Center for Substance Abuse Prevention and the National Institute on Drug Abuse also have provided support for Fast-Track through a memorandum of agreement with the NIMH. This work was also supported in part by Department of Education Grant S184U30002 and NIMH Grants K05MH00797 and K05MH01027.

We are grateful for the close collaboration of the Durham public schools, the metropolitan Nashville public schools, the Bellefonte area schools, the Tyrone area schools, the Mifflin County schools, the Highline public schools, and the Seattle public schools. We greatly appreciate the hard work and dedication of the many staff members who implemented the project, collected the evaluation data, and assisted with data management and analyses. The authors acknowledge the helpful comments of Elizabeth Gifford, Eric Ford, Joe Schafer, Damon Jones, John Graham, Allison Olchowski, David MacKinnon, Mark Courtney, and the Methodology Workshop at Pennsylvania State University. The authors are responsible for any remaining errors.

Appendix: Specification of aML Code

aML Code	Function	Comments
<code>dsn = pdrcov . dat ;</code>	Reads data formatted for aML	The steps for creating data in the appropriate format are described in the aML manual
<code>define regressor set Beta ; var = 1</code> <code>(site==2) (site==3) (site==4) race</code> <code>(cohort==2) (cohort==3) (year - 4) ;</code>	Specifies a set of regression coefficients corresponding to each variable listed	There are two vectors of coefficients corresponding to Equations 1 and 2 in the text aML uses the convention "(X==x)" to create a dummy variable for each value of the variable X

aML Code	Function	Comments
<pre> define regressor set Gamma ; var = 1 . . . (year - 4) ; define regressor set Beta1 ; var = 1 ; define regressor set Gamma1 ; var = 1 ; define normal distribution ; dim=1 ; name=epsC ; define normal distribution ; dim=1 ; name=epsT ; define normal distribution ; dim= 1 ; number of integration points=10 ; name=muC ; define normal distribution ; dim= 1 ; number of integration points=10 ; name=muT ; define parameter lambdaC ; define parameter lambdaT ; continuous model ; keep if (male==0) and (miss==0) and (trt==0) ; outcome = pdr ; model = regset Beta + intres (draw=_id, ref=muC) + res (draw=_iid, ref=epsC) ; </pre>	<p>Specifies coefficients for treatment status</p> <p>Specifies the error distributions used in Equations 1 and 2</p> <p>Specifies the selection parameters</p> <p>Specifies Equation 1 for the control group</p> <p>The “keep” command subsets the data</p>	<p>These are not included in the vectors beta and gamma for clarity (separate parameters are also required for the multigroup framework as described below)</p> <p>Epsilon and mu are specified twice, once each for the treatment and control groups— doing so allows for heteroscedasticity between the two groups</p> <p>The “dim” option specifies the dimension of the distribution; the errors are all univariate</p> <p>This model is for girls (male==0)</p> <p>Only those cases with nonmissing data (miss=0) contribute to the likelihood function for the outcome variable</p> <p>This portion of the likelihood pertains to the comparison group (trt==0)</p> <p>The “draw” parameter specifies that muC and epsC are person-specific and time-varying errors, respectively</p>
<pre> probit model ; keep if (male==0) and (trt== 0) ; outcome=(miss==0) ; model = regset Gamma + par lambdaC * intres (draw=_id, ref= muC) ; </pre>	<p>Specifies Equation 2 for the control group</p>	<p>The model is a probit as described in the text; the outcome is participation (miss==0)</p> <p>One can see that participation depends on muC as described in the text; because this model is for the control group, only one of the lambda terms appears</p> <p>The N(0, 1) error is implicit and included by aML without explicit specification</p>
<pre> continuous model ; keep if (male==0) and (miss==0) and (trt==1) ; outcome = pdr ; model = regset Beta + regset Beta1 + intres (draw=_id, ref=muT) + res (draw=_iid, ref=epsT) ; probit model ; keep if (male==0) and (trt==1) ; outcome=(miss==0) ; model = regset Gamma + regset Gamma1 + par lambdaC * intres (draw=_id, ref=muT)+ par lambdaT * intres (draw=_id, ref= muT) ; starting values ; ; </pre>	<p>Specifies Equations 1 and 2 for the treatment group (trt==1)</p>	<p>One can see that both lambda terms appear in the second equation</p> <p>Beta1 and Gamma1 appear only in the equations for the treatment group; they capture the impact of treatment in the outcome and participation equations (in effect, these coefficients represent constants in the multigroup model and so capture the impact of group membership)</p> <p>Specifies starting values</p>

NOTE: The program code is specified in three sections. After reading in the data, the first specifies the model components (regression coefficients, error distributions, and any additional parameters). The second specifies the equations used to calculate the likelihood function. The third section specifies starting values.

Biographies

E. Michael Foster is a professor of health policy and administration and of demography at Pennsylvania State University. His research focuses on the evaluation of services and interventions targeted to youth with emotional and behavioral problems or who are at risk for those problems. For more information, see www.personal.psu.edu/emf10/

Grace Y. Fang is a statistician in the Fast Track project at Pennsylvania State University

References

- Achenbach, TM. Manual for the Child Behavior Checklist 4-18 and 1991 profile. Burlington: Department of Psychiatry, University of Vermont; 1991.
- Agresti, A. Categorical data analysis. Vol. 2nd. Hoboken, NJ: John Wiley; 2002.
- Allison, PD. Missing data. Thousand Oaks, CA: Sage; 2002.
- Asher SR, Dodge KA. Identifying children who are rejected by their peers. *Developmental Psychology* 1986;22:444–49.
- Chamberlain P, Reid JB. Parent observation and report of child symptoms. *Behavioral Assessment* 1987;9:97–109.
- Coie JD, Dodge KA, Coppotelli HA. Dimensions and types of social status: A cross-age perspective. *Developmental Psychology* 1982;18:557–69.
- Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods* 2001;6:330–51. [PubMed: 11778676]
- Conduct Problems Prevention Research Group. Initial impact of the Fast Track prevention trial for conduct problems, I: The high-risk sample. *Journal of Consulting and Clinical Psychology* 1999a; 67:631–47.
- Conduct Problems Prevention Research Group. Initial impact of the Fast Track prevention trial for conduct problems, II: Classroom effects. *Journal of Consulting and Clinical Psychology* 1999b; 67:648–57.
- Conduct Problems Prevention Research Group. Evaluation of the first three years of the Fast Track prevention trial with children at high risk for adolescent conduct problems. *Journal of Abnormal Child Psychology* 2002a;30:19–35.
- Conduct Problems Prevention Research Group. The implementation of the Fast Track program: An example of a large-scale prevention science efficacy trial. *Journal of Abnormal Child Psychology* 2002b;30:1–17.
- Conduct Problems Prevention Research Group. Predictor variables associated with positive Fast Track outcomes at the end of third grade. *Journal of Abnormal Child Psychology* 2002c;30:37–52.
- Cook, RD.; Weisberg, S. Applied regression including computing and graphics. New York: John Wiley; 1999.
- Davidson, R.; MacKinnon, JG. Estimation and inference in econometrics. New York: Oxford University Press; 1993.
- Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *Applied Statistics* 1994;43 (1):49–93.
- Elliott, DS.; Ageton, SS.; Huizinga, D. Explaining delinquency and drug use. Beverly Hills, CA: Siegel; 1985.
- Elliott, DS.; Huizinga, D.; Menard, S. Multiple problem youth: Delinquency, substance use, and mental health. New York: Springer-Verlag; 1989.
- Enders C. A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling* 2001;8(1):128–41.
- Foster, EM. Using services data to adjust for attrition in outcome analyses: An application of multiple imputation. Paper presented at “A System of Care for Children's Mental Health: Expanding the Research Base, 15th Annual Research Conference”; Tampa, FL. 2002.
- Foster EM, Bickman L. An evaluator's guide to detecting attrition problems. *Evaluation Review* 1996;20 (6):695–723.
- Graham JW. Adding missing-data-relevant variables to FIML-based structural equations models. *Structural Equation Modeling* 2003;10(1):80–100.
- Greene, WH. Econometric analysis. Vol. 2nd. New York: Macmillan; 1993.
- Hall SM, Delucchi KL, Velicer WF, Kahler CW, Ranger-Moore J, Hedeker D, Tsoh JY, Niaura R. Statistical analysis of randomized trials in tobacco treatment: Longitudinal designs with dichotomous outcome. *Nicotine and Tobacco Research* 2001;3:193–202. [PubMed: 11506764]
- Heckman JJ. Shadow wages, market prices and labor supply. *Econometrica* 1974;42:679–94.

- Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 1976;5:475–92.
- Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979;47:153–61.
- Kenward MG. Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine* 1998;17:2723–32. [PubMed: 9881418]
- Laird NM. Discussion of Diggle PJ, Kenward MG: Informative dropout in longitudinal data analysis. *Applied Statistics* 1994;43:84.
- Lillard LA, Panis CWA. Panel attrition from the Panel Study of Income Dynamics. *Journal of Human Resources* 1998;33:437–57.
- Lillard, LA.; Panis, CWA. aML multi-level multiprocess statistical software. Version 1.0. Los Angeles: EconWare; 2000.
- Littell, RC.; Milliken, GA.; Stroup, WW.; Wolfinger, RD. SAS system for mixed models. Cary, NC: SAS Institute Inc.; 1999.
- Little, RJA.; Rubin, DB. Statistical analysis with missing data. New York: John Wiley; 1987.
- Little, RJA.; Rubin, DB. Statistical analysis with missing data. Vol. 2nd. Hoboken, NJ: John Wiley; 2002.
- Lochman JE, Conduct Problems Prevention Research Group. Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology* 1995;63:549–59. [PubMed: 7673532]
- Muthen, LK.; Muthen, BO. Mplus user's guide. Los Angeles: Muthen & Muthen; 2001.
- Orr, LL. Social experiments: Evaluating public programs with experimental methods. Thousand Oaks, CA: Sage; 1999.
- Rosenbaum, PR. Observational studies. New York: Springer-Verlag; 2002.
- SAS Institute. The MI procedure. Cary, NC: SAS Institute; 2001a.
- SAS Institute. The MIAnalyze procedure. Cary, NC: SAS Institute; 2001b.
- Schafer, JL. Analysis of incomplete multivariate data. London: Chapman and Hall; 1997.
- Schafer JL. Multiple imputation: A primer. *Statistical Methods in Medical Research* 1999;8:3–15. [PubMed: 10347857]
- Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychological Methods* 2002;7(2):147–77. [PubMed: 12090408]
- Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems. *Multivariate Behavioral Research* 1998;33(4):545–71.
- Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* 1982;77:848–54.
- StataCorp. Stata statistical software: Release 7.0. College Station, TX: Stata Corporation; 2001.
- Stolzenberg RM, Rells DA. Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods and Research* 1990;18:395–415.
- Stolzenberg RM, Rells DA. Tools for intuition about sample selection bias and its correction. *American Sociological Review* 1997;62:494–507.
- Terry R, Coie JD. A comparison of methods for defining sociometric status among children. *Developmental Psychology* 1991;27:867–80.
- Verbeek M, Nijman T. Testing for selectivity bias in panel data models. *International Economic Review* 1992;33:681–703.
- Verbeke, G.; Molenberghs, G. Linear mixed models in practice. New York: Springer-Verlag; 1997.
- Verbeke, G.; Molenberghs, G. Linear mixed models for longitudinal data. New York: Springer-Verlag; 2000.
- Wooldridge, JM. Econometric analysis of cross section and panel data. Cambridge, MA: MIT Press; 2002.

TABLE 1
Percentage of Observations Missing PDR at Grades 3 to 5, by Gender and Treatment Status

	Boys (n = 615)		Girls (n = 276)		p ^a
	Treatment	Control	Treatment	Control	
Grade 3	8.4	10.9	5.6	17.9	<.01
Grade 4	10.0	12.2	5.6	13.9	.02
Grade 5	13.4	13.9	10.4	15.2	.24
Missing PDR at all three waves	5.6	6.4	2.4	9.3	.02
Missing PDR at any of the three waves	17.2	21.4	14.4	24.5	.04

NOTE: PDR = Parent Daily Report.

^a p refers to significance level of treatment-control difference.

TABLE 2
Comparison of Observations With and Without PDR at Grade 5, by Gender and Treatment Status

	Boys						Girls					
	Treatment Group: PDR at Grade 5			Control Group: PDR at Grade 5			Treatment Group: PDR at Grade 5			Control Group: PDR at Grade 5		
	Missing	Present	p ^a	Missing	Present	p ^a	Missing	Present	p ^a	Missing	Present	p ^a
Sample size	43	277	—	41	254	—	13	112	—	23	128	—
Percentage African American	39.5	57.4	.03	29.3	51.6	.01	38.5	48.2	.50	47.8	49.2	.90
Site (%) ^b			<.01			<.01			.17			.64
Durham	11.6	30.0		2.4	30.7		0	19.6		13.0	21.1	
Nashville	18.6	28.2		29.3	23.6		15.4	23.2		34.8	28.1	
Pennsylvania	20.9	24.9		22.0	24.4		30.8	27.7		21.7	28.1	
Seattle	48.8	17.0		46.3	21.3		53.9	29.5		30.4	22.7	
Cohort (%) ^b			.31			.27			.67			.07
1	32.6	37.2		26.8	39.8		23.1	31.3		13.0	31.3	
2	46.5	34.7		39.0	34.3		46.2	33.9		34.8	39.1	
3	20.9	28.2		34.2	26.0		30.8	34.8		52.2	29.7	
Baseline PDR	0.24	0.23	.74	0.25	0.23	.47	0.21	0.23	.66	0.21	0.25	.23

NOTE: PDR = Parent Daily Report.

^a *p* refers to significance level of difference between individuals with and without PDR at Grade 5.

^b Entries sum to 100%.

TABLE 3
Availability of Supplemental Data, by Treatment, Gender, and Availability of PDR

	Boys						Girls					
	Treatment Group: PDR at Grade 5			Control Group: PDR at Grade 5			Treatment Group: PDR at Grade 5			Control PDR Group: at Grade 5		
	Missing	Present	p ^a	Missing	Present	p ^a	Missing	Present	p ^a	Missing	Present	p ^a
Sample size	43	277	—	41	254	—	13	112	—	23	128	—
TRF												
Percentage with measure	48.8	90.3	<.01	51.2	88.2	<.01	69.2	91.1	.02	21.7	88.3	<.01
Aggressive behavior	20.43	17.42	.28	19.48	16.63	.33	11.56	13.01	.72	7.00	11.65	.41
Externalizing behavior	25.10	21.12	.24	23.91	20.46	.33	14.00	15.88	.69	8.60	14.26	.41
Self-reported aggression												
Percentage with measure	20.9	96.8	<.01	2.4	97.2	<.01	0	96.4	<.01	0	98.4	<.01
Average score	12.67	5.75	.11	2.00 ^b	5.18	—	NA	3.19	—	NA	3.20	—
Peer ratings												
Percentage with measure	46.5	66.4	.01	51.2	62.6	.17	84.6	63.4	.13	43.5	62.5	.09
Aggressive behavior	4.95	5.11	.88	5.76	5.26	.65	1.09	1.92	.15	2.80	1.88	.36
Hyperactivity	4.45	5.71	.21	5.71	5.25	.66	1.27	1.70	.43	1.7	2.26	.57
IEP												
Percentage with measure	58.1	98.9	<.01	56.1	97.6	<.01	61.5	98.2	<.01	34.8	99.2	<.01
Percentage having IEP	44.0	33.2	.28	26.1	33.9	.45	25.0	22.7	.73	12.5	31.5	.24

NOTE: PDR = Parent Daily Report; TRF = Teacher Report Form; IEP = individualized Education Plan; NA = not available.

^a *p* refers to significance level of difference between individuals with and without PDR.

^b This entry represents a single case, so a test of statistical significance was not possible.

TABLE 4

Estimated Treatment Effect Under Alternative Models, by Gender

Model	Boys: Selection Parameters						Girls: Selection Parameters					
	Estimated Treatment Effect (β_1)			Control Group (α_C)			Estimated Treatment Effect (β_1)			Control Group (α_C)		
	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p	Estimate	SE	p
1. Listwise deletion	-0.47	0.25	.06	NA	NA	NA	-0.75	0.38	.05	NA	NA	NA
2. ignorable maximum likelihood	-0.40	0.23	.08	NA	NA	NA	-0.75	0.35	.03	NA	NA	NA
3. Standard MI ^a	-0.37	0.23	.10	NA	NA	NA	-0.73	0.35	.03	NA	NA	NA
4. Enhanced MI (1) ^b	-0.44	0.23	.05	NA	NA	NA	-0.57	0.34	.09	NA	NA	NA
5. Enhanced MI (2) ^c	-0.43	0.23	.07	NA	NA	NA	-0.60	0.34	.08	NA	NA	NA
6. Panel selection model	-1.04	0.28	<.01	-0.19	0.02	<.01	-0.34	0.42	.42	0.39	0.03	<.01

NOTE: Samples sizes are 276 girls and 615 boys for all analyses except listwise deletion. For those analyses, the sample sizes are 221 and 497, respectively. For each model, the outcome variable is the Parent Daily Report averaged across three administrations for a given wave of interviews. For all models, race, cohort, and site were also included as predictors of both participation and the outcome. The full set of parameters is available from the first author. MI = multiple imputation; NA = not available.

^aStandard MI refers to MI model in which the imputation and analysis models include the same variables.

^bEnhanced MI (1) refers to the first enhanced imputation model, which included the outcome measured for the baseline, Grade 1, and Grade 2 interviews.

^cEnhanced MI (2) includes the same variables as Model 1 but also includes the Teacher Report Form, the Things You Have Done measure, peer reports, and special education during Grades 4 and 5. Also included was an array of demographic variables from the baseline interview.