Statistics, Politics, and Policy

Volume 3, *Issue* 1 2012 *Article* 6

Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies

Christopher H. Rhoads, *University of Connecticut*

Recommended Citation:

Rhoads, Christopher H. (2012) "Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies," *Statistics, Politics, and Policy*: Vol. 3: Iss. 1, Article 6.

DOI: 10.1515/2151-7509.1012

Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies

Christopher H. Rhoads

Abstract

Policy analysts involved in quantitative research have many options for handling missing data. The method chosen will often greatly influence the substantive policy conclusions that will be drawn from the data. The most frequent methods for handling missing data assume that the data are missing at random (MAR). The current paper notes that an omnibus, nonparametric test of the MAR assumption is impossible using the observed data alone. Nonetheless various purported tests of the missingness mechanism (including tests of MAR) appear in the literature. The current paper clarifies that all of these tests rely on some assumption that cannot be tested from the data. The paper notes that tests of the missingness mechanism are frequently misinterpreted and it clarifies the appropriate interpretation of such tests. Policy analysts are encouraged not to develop the false impression that modern procedures for handling missing data in conjunction with tests of the missingness mechanism provide protection against the ill effects of missing data. Any justification for a particular approach to handling missing data must be come from substantive knowledge of the missingness process, not from the data.

KEYWORDS: missing data, tests of the missingness mechanism, sensitivity analysis

Author Notes: The author would like to thank Larry Hedges, Chuck Manski, Dan McCaffrey, David Banks and two anonymous reviewers for many useful comments.

1. INTRODUCTION

Data which is collected and analyzed with an eye towards informing policy decisions will frequently be subject to substantial non-sampling error in the form of missing data. It is the norm, rather than the exception, that policy researchers will be unable to obtain values on all variables of interest for all subjects in a study.

The importance of using appropriate procedures to handle missing data is underscored by the number of publications dealing with the topic. A June 8, 2010 search on Google Scholar for publications since 2007 containing the words "missing data" in the title returns over 1,000 results. In the last decade, review articles on the topic of missing data have been published in journals as diverse as Drug Information Journal (Myers, 2000), Obesity Reviews (Gadbury, Coffey and Allison, 2003), Nutrition (Fitzmaurice, 2008), Circulation: Cardiovascular Quality and Outcomes (He, 2010), Review of Educational Research (Peugh and Enders, 2004) and Annual Review of Psychology (Graham, 2009), among many others. The Institute of Education Sciences, established by the U.S. federal government in 2002 to help inform education policy by facilitating rigorous research about what works in educational practice, has recently issued a 100 plus page report on the topic (Puma, Olsen, Bell and Price, 2009).

The manner in which missing data is handled often receives scant attention in a research report. Generally a sentence or two is devoted to the attrition rates in treatment and control groups. Another few sentences will be devoted to the way in which the researchers have chosen to handle missing data in their analysis. However, faced with an inability to eradicate all missing data, researchers have a number of possible options for handling missing data at the analysis phase of a project. Which option is chosen can substantially impact the conclusions that are drawn from the analysis. Therefore, understanding how the authors have handled missing data, and thinking seriously about the plausibility of the assumptions necessary to justify a particular missing data model, are crucial to understanding the policy implications of a study. In particular, the choice of an appropriate method may be greatly influenced by the individual data structure in question. There are certain methods that are most appropriate for handling missing data due to dropout (as often occurs in longitudinal studies). Other (perhaps less complex) methods are better suited to cross-sectional data. In multi-site studies it may well be appropriate to assume different missing data processes for the different sites.

The validity of a particular approach to handling missing data will depend on the nature of the process that generates the missing data, often referred to as the *missing data mechanism*. Rubin (1976) elaborated three classifications for missingness mechanisms, a taxonomy that continues to define modern discussions of missing data procedures

The data are referred to as *missing completely at random* (MCAR) if the probability that a data point is missing does not depend on any variables, either

observed or unobserved. Historically, most procedures for handling missing data require that the data be MCAR in order for unbiased inferences to be possible. For instance *complete case analysis* requires MCAR data in order to produce unbiased estimates of parameters of interest. An example of this approach was given in Gibb, Beevers, Andover and Holleran (2006).

Gibb, et al. collected data in a 6-week longitudinal study of university undergraduates. The study's goal was to test the "vulnerability-stress" hypothesis, which holds that individuals who tend to attribute negative events to stable and global causes will tend to develop symptoms of depression in the presence, but not the absence, of negative life events. The study's research questions were explored by fitting hierarchical linear models to the subset of respondents for whom there was complete data over the course of the six week study. Other procedures which either require MCAR to be valid (or may not be valid under any conditions) are available-case analysis, mean imputation, and (for longitudinal studies) last-value-carried-forward (LVCF).

The most commonly recommended procedures for handling missing data in the present day require a weaker condition than MCAR to produce valid inferences. These procedures require that the missingness mechanism be *missing at random* (MAR). A missingness process is MAR if the probability of missingness depends on fully observed values in the data set, however it does not depend on the unobserved missing values. It is worth noting that all MCAR missing data processes are also MAR, but the converse is not true. The two general procedures most commonly thought of as "state of the art" (Peugh and Enders, 2004; Schafer and Graham, 2002) require that the missing data be MAR. The first of these procedures is *multiple imputation* (MI). An example of this sort of approach is given by Simmons, Hairell, Edmonds, Vaughn, Larsen, Willson, Rupley and Byrns (2010). Simmons, et al. describe a study examining two different approaches for helping teachers to facilitate content comprehension and content vocabulary knowledge in fourth grade social studies students. Missing data was handled by using multiple imputation to compute parameter estimates and their standard errors.

The second "state of the art" procedure is maximization of the *observed data likelihood* (the complete data likelihood with the missing data integrated out). This is sometimes called the "ignorable maximum likelihood" (IML) method. This approach was taken by Grella, Scott, Foss and Dennis (2008) in a six year longitudinal study of individuals with substance abuse addictions. Interest was in modeling the factors impacting transition between each of four states: recovery, treatment, incarcerated and using. Mixed-effects multinomial logistic regression models were fit by maximizing the full-information likelihood function using the MIXNO software (Hedeker, 1999).

The fact that MI and ML procedures require only the MAR assumption provides a great opportunity for the researcher. She may gather information on auxiliary variables that would not otherwise be of interest but which are believed to be correlated with the cause of missingness. By including these variables in the model used to analyze the data a missingness process that would otherwise be MNAR (see next paragraph) may be transformed to a MAR missingness process. In MI applications, these auxiliary variables can be included in the imputation model and a more parsimonious model can still be utilized to answer the substantive question of interest.

Finally, a missingness process is *missing not at random* (MNAR) if it is not MAR, that is if the probability of missingness depends on the underlying value of the missing data. Methods for handling MNAR data are of two types. The first type is *selection models*. Selection models specify an explicit, parametric model which relates unobserved values to the probability of missingness.

The other main way of handling MNAR data is via a *pattern mixture model*. Unlike selection models, pattern mixture models do not explicitly posit a model relating unobserved values to the probability of response. Instead, the data are stratified by missing data pattern and separate models are fit within each strata. The stratum-specific estimates may then be combined via a researcher posited weighting function.

I note that sometimes aspects of more than one approach to missing data are found in the same study. Consider the case of Flook, Repetti and Ullman (2005). Flook, et al. describe a study exploring the relation between social acceptance in the classroom and academic performance. The conceptual model that the authors wanted to test implied that a lack of peer acceptance impacted academic performance through the mediators of "Academic Self-Concept" and "Internalizing Symptoms" (which includes experiencing negative emotional states such as sadness and anxiety). The authors conducted a path analysis to test their theory of mediation. For variables with less than 8% data mising, missing values were imputed from the EM estimate of the covariance matrix. For variables with more than 8% data missing, only cases with complete data on these variables were retained.

Unlike methods meant to deal with MNAR missing data, methods that require MAR or MCAR missing data do not posit a specific model for the missing data mechanism. For this reason MAR and MCAR missing data are often together called "ignorable" missing data types. Because ignorable missing data methods do not specify a particular model for the missing data some analysts may have the mistaken impression that these methods require weaker assumptions than MNAR methods. This is not the case. The MAR assumption is a very strong assumption, picking out a single model from the infinite dimensional space of possible models for the missingness mechanism.

Given that all approaches to the missing data problem rely on (different) strong assumptions, statistical tests that use the observed data to help the researcher identify the correct missing data mechanism would be quite useful. In fact, a number of tests of the missing data mechanism have appeared in the literature. The goal of this paper is to clarify the proper interpretation of the results of these tests and to note that tests based on the observed data can never distinguish between a missingness process that is MAR and one that is MNAR. Since MAR makes assumptions about the distribution of the missing data after conditioning on all observed data, the observed data provides no evidence either for or against the hypothesis. Procedures that purport to test MAR are in fact testing something else altogether, generally either (a) a necessary (but not sufficient) condition for MCAR or (b) the parametric modeling assumptions.

A further point of this paper is to note that since MAR is a necessary condition for MCAR and MAR is untestable, tests of the MCAR assumption are in fact testing only a necessary condition for MCAR. This point seems to be little recognized in applications of the tests. Research reports far too frequently take failure to reject the null hypothesis of a test of MCAR as the sole justification for an analysis that assumes MCAR. This is a mistaken application of these tests. Tests of the MCAR assumption should be used only if the MAR assumption is deemed likely to hold a priori, since all tests of this sort can only test MCAR relative to a MAR alternative hypothesis. The ultimate justification for choosing a particular analytic technique for handling missing data should come from substantive knowledge of the particular problem at hand, not from any features of the observed data.

The rest of this paper proceeds as follows. Section 2 defines some terminology and gives a broad outline of the reason why the observed data can provide no evidence either for or against MAR. Section 3 considers published tests of the MCAR and MAR assumptions, critiques their use by applied researchers, and notes that procedures that claim to test the MAR or "ignorability" condition are in fact testing something different. The final section discusses some alternatives to assuming an ignorable missing data mechanism. These alternatives can be grouped into three categories: modeling the MNAR missingness mechanism, computing nonparametric bounds for the partially identified quantities of interest, and sensitivity analysis.

2. Missing data mechanisms and the impossibility of testing MAR

Rubin (1976) introduced a taxonomy for mechanisms that generate missing data that has become the foundation of virtually any modern discussion of missing data methods. I summarize that taxonomy in the following, using the notation of Little and Rubin (2002).

Let $\mathbf{Y}$ denote the complete data matrix. That is, $\mathbf{Y}$ is the data that would have been observed had there been no missing data. $\mathbf{Y}$ is subdivided by writing

$\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, where $\mathbf{Y}_{obs}$ = observed part of $\mathbf{Y}$ and $\mathbf{Y}_{mis}$ = missing part of $\mathbf{Y}$. I suppose that the distribution of $\mathbf{Y}$ may be characterized by a finite dimensional parameter vector θ. Let $\mathbf{R}$ represent a missing data indicator matrix indexing which elements of $\mathbf{Y}$ are missing. The random matrix $\mathbf{R}$ is characterized by its conditional distribution given the complete data and is indexed by a parameter vector ϕ. Thus, we characterize types of missing data by reference to $P(\mathbf{R}|\mathbf{Y}, \phi)$, where $P(.)$ denotes a generic probability distribution over its argument(s). Then the data are *missing completely at random* (MCAR) if

$$P(\mathbf{R}|\mathbf{Y}, \phi) = P(\mathbf{R}|\phi) \text{ for all } \mathbf{Y},\ \phi. \tag{1}$$

The MCAR assumption is a special case of the less restrictive assumption that the data are *missing at random* (MAR). This assumption implies that

$$P(\mathbf{R}|\mathbf{Y}, \phi) = P(\mathbf{R}|\mathbf{Y}_{obs}, \phi) \text{ for all } \mathbf{Y}_{mis},\ \phi. \tag{2}$$

In other words, here the probability of missingness is allowed to depend on any fully observed values of the matrix $\mathbf{Y}$. Any data for which the MAR assumption does not hold is termed *missing not at random* (MNAR).

Rubin (1976) explored conditions under which the missing data mechanism is "ignorable". He notes that the complete data likelihood may be written as $P_\theta(\mathbf{Y})P_\phi(\mathbf{R}\ |\mathbf{Y})$. Ignorability means that inferences about θ may be made by (a) fixing the random variable $\mathbf{R}$ at the observed pattern of missing data $\widetilde{\mathbf{r}}$, and (b) assuming that the values of $\mathbf{Y}_{obs}$ come from the marginal density

$$\int P_\theta(\mathbf{Y}) d\mathbf{Y}_{mis}. \tag{3}$$

The conditions necessary for inferences that ignore the missing data mechanism to be valid will vary with the type of inferential procedure used. In all cases the so-called *parameter distinctness (PD)* condition is necessary. This condition states that the parameter space for the combined parameter (θ, ϕ) is the Cartesian product of the individual parameter spaces. Rubin (1976) shows that, in addition to *PD*, inferences based on the sampling distributions of observed statistics require that the missing data mechanism be MCAR in order to be correct. However, Bayesian inference or direct-likelihood inference (inference that results solely from comparing the ratio of the likelihood function for different values of the parameter) only require the weaker MAR assumption.

As noted in the introductory section, methods that assume an ignorable missing data mechanism are the most frequently used. Little and Rubin's standard treatment of missing data makes this point as follows:

> Essentially all the literature on multivariate incomplete data assumes that the data are MAR, and much of it also assumes that the data are MCAR. Chapter 15 deals explicitly with the case when the data are not MAR and models are needed for the missing-data mechanism.

> Since it is rarely feasible to estimate the mechanism with any degree of confidence, the main thrust of these methods is to conduct sensitivity analyses to assess the effect of alternative assumptions about the missing-data mechanism (Little and Rubin, 2002, p. 22).

2.1. **The possibility of tests of the MCAR and MAR assumptions.** The necessity of assuming MAR or MCAR to justify most commonly used missing data procedures has understandably led to methodological research that attempts to devise "diagnostics" for the missing data mechanism and test procedures that attempt to use the observable data to guide researchers as to when it is reasonable to maintain the MAR or the MCAR assumption. Unfortunately, the empirical data can never provide evidence either for or against the MAR assumption, and the MCAR assumption can be falsified by the data, but it can never be proven true.

The basis of the problem is as follows. The MCAR assumption implies that $P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R})$. However, since the values of $\mathbf{Y}_{mis}$ are unknown the empirical evidence gives equal support to the following two hypotheses:

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R}) \tag{4}$$

$$P(\mathbf{R}\,|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R}\,|\mathbf{Y}_{mis}). \tag{5}$$

Similarly, the MAR assumption implies that $P(\mathbf{R}|\mathbf{Y}) = P(\mathbf{R}|\mathbf{Y_{obs}})$. But the hypotheses

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R}|\mathbf{Y}_{obs}) \tag{6}$$

$$P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) \tag{7}$$

are both equally plausible as far as the empirical data are concerned.

2.2. **In what sense is MCAR testable?** While we can never hope to distinguish between the situations described by equations (4) and (5) using only the data, it is possible to show from the data alone that both equation (4) and (5) are very likely to be false. This is because, while MAR implies that missingness may depend on observables but cannot depend on unobservables, MCAR implies that missingness cannot depend on either observables or unobservables. While the data tells us nothing about whether missingness depends on unobservables, it certainly can tell us whether missingness depends on observables.

It will be useful in thinking about the logic behind proposed tests of the missingness mechanism to have on hand the following characterizations of the MCAR and MAR assumptions. These follow immediately from the application of Bayes' rule to equations (4) and (6).

$$P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R})P(\mathbf{Y}_{obs}|\mathbf{R}) = P(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) \Leftrightarrow \text{(MCAR)} \tag{8}$$

$$P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}) = P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}) \Leftrightarrow \text{(MAR)}. \tag{9}$$

Note that equality in the equation characterizing MCAR will hold if $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}) = P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ and

$$P(\mathbf{Y}_{obs}|\mathbf{R}) = P(\mathbf{Y}_{obs}). \qquad \text{(OAR)} \tag{10}$$

So MCAR can be thought of as the conjunction of condition MAR, which cannot be tested, and the condition given in equation (10), which can be tested. The condition given in (10) is an implication of Rubin's (1976) *observed at random* condition. While this characterization of MCAR as the conjunction of MAR and OAR is not often remarked upon in the literature, it is important for understanding the nature of tests of the MCAR assumption. This characterization makes clear that tests of MCAR are in fact testing OAR, that is, they are testing a necessary condition for MCAR. These tests provide no evidence either for or against the MAR assumption. This fact is made explicit in the discussion that follows.

3. Tests of the Missingness Mechanism

Tests of the missingness mechanism can be grouped into three main types: (a) Tests of MCAR, (b) tests of MAR against a particular MNAR alternative, and (c) tests that claim to evaluate MAR or ignorability, but in fact test a weaker condition. This section evaluates examples of each of these types of tests in some detail.

3.1. **Tests of MCAR.**

3.1.1. *A simple example.* Suppose that we have the following simple data structure. Two variables are measured for each subject. Values of the first variable are always observed but values of the second variable are sometimes missing. Assume that the complete data follows a bivariate normal distribution. Then a simple test of OAR could be constructed as follows. First, group cases on the basis of whether or not they have missing data on the second variable. Then use an independent groups t test to determine whether the means on the fully observed variable are significantly different for the two groups. If they are, then the OAR hypothesis is rejected (and hence, so is the MCAR hypothesis). If they are not, then there is insufficient evidence to determine whether or not OAR holds. Notice that, regardless of the outcome of the test, no evidence is adduced either for or against MAR. In principal other features (besides the means) of the distribution of the data in the two groups could be compared (such as the variances). However, most tests of MCAR focus on comparing multivariate means and so are extensions of the simple t test described above.

3.1.2. *Little's test.* Probably the most well known test of the MCAR assumption was proposed by Little (1988). This test can be regarded as an extension of the simple test described above to the case where there are multiple variables measured for each individual and many different patterns of missing data. Rather than perform separate t tests for each variable and each pattern of missing data, Little provides

an omnibus test. The test is a likelihood ratio test comparing the likelihood under MAR to the likelihood under MCAR. Specifically, first the data set is stratified according to patterns of missing data. Let $j = 1, \ldots, J$ index the distinct patterns of missingness. The likelihood under MAR assumes that the observed data mean in strata j, $\bar{\mathbf{Y}}_{\mathbf{obs,j}}$, follows a multivariate normal distribution with mean $\nu_{\mathbf{j}}$. The likelihood under MCAR is the same as under MAR, with the additional restriction that the $\nu_{\mathbf{j}}$ parameters must be the same across all missing data patterns. Since the test is a likelihood ratio test of MAR vs. MCAR, it is clear that it is best regarded as a test of OAR. It provides no evidence for or against MAR.

3.1.3. *Other likelihood ratios tests of MCAR vs. MAR.* There are other proposed tests of the MCAR assumption in the literature (Lipsitz, Laird and Harrington, 1994; Park and Lee, 1997; Chen and Little, 1999). Of these, the Lipsitz, Laird and Harringrton (1994) test is quite explicitly a likelihood ratio test of a MCAR model versus a MAR alternative. Hence, the test is useful only if one is willing to accept the MAR assumption *a priori*. The Park and Lee (1997) test is a variant of the Park and Davis (1993) test described in great detail in the Appendix. Strata are defined based on patterns of missing data and parameter estimates are compared across strata. If these estimates are sufficiently similar across strata, then the null hypothesis that the data are MCAR is not rejected. The Chen and Little (1999) test is essentially a non-parametric extension of Little's (1988) test, intended for use with generalized estimating equation (GEE) methods. Like Little's test it is useful only in distinguishing MCAR and MAR missing data processes. It has no value for distinguishing MCAR from MNAR missing data processes.

The problem with tests of the MCAR assumption is that those who propose these tests often fail to emphasize that the condition being tested is only necessary and not sufficient for MCAR. This makes it likely that applied researchers will use the tests without appropriate caution.

Much of the writing on missing data contributes to the misperception that MCAR can be tested against MNAR alternatives. While acknowledging that MAR is not testable, the literature tends to imply that MCAR is testable. Thus we find statements like, "It is certainly possible to test whether the data are missing completely at random...However, it is impossible to test whether the data are missing at random (Lewis-Beck, Bryman and Liao, 2004)." Similarly, Myers (2000) notes that "MAR is inherently untestable" but claims "The MCAR assumption can be assessed by comparing the distribution of observed variables between dropouts and non-dropouts. If no significant differences are found with respect to the variables, then there is no apparent evidence that the data from the clinical trial are representative only of completers." Peugh and Enders (2004) say, "It is important to note that MCAR is the only mechanism that can be empirically tested from a set of data." Fitzmaurice (2008), in discussing the MAR assumption, states, "Unlike an MCAR

mechanism it is not possible to verify this assumption from the data at hand." Finally, Van Ness, Murphy, Araujo, Pisani and Allore (2007) seem to imply that both MAR and MCAR are testable when they say, "before imputing or weighting, however, one should confirm that missing values are MAR or MCAR."

Almost certainly most authors cited above are aware that MCAR can only be tested in the sense of being "ruled out" by the data. However, the examples presented below seem to show that many data analysts have the mistaken impression that the MCAR assumption can be entirely justified by the failure of a given hypothesis test to reject MCAR.

A quick computer search turned up a number of papers citing non-significant results from Little's MCAR test as the sole justification for an analytic procedure that requires the MCAR assumption (Steginga, Pinnock, Gardner, Gardiner and Dunn, 2005; Gibb, Beevers, Andover and Holleran, 2006; Kunina, Wilhelm, Formazin, Jonkmann and Schroeders, 2007; Buckler and Unnever, 2008; Salmon, Holcombe, Clark, Krespi, Fisher and Hill, 2007). Typical comments were, "The MCAR Little-Test (Little, 1988) was not significant, implying that the data is (sic) missing completely at random and that its absence is not a function of other observed or unobserved variables" (Kunina et al., 2007), and "the value of Little's MCAR test was not significant, indicating that the data were missing completely at random" (Steginga et al., 2005).

Another set of papers reported results from Little's test, but used an analytic method that required only the MAR assumption. No justification for maintaining the MAR assumption was given (Sundin and Ogren, 2006; Gonzalez, Desai, Sofuoglu, Poling, Oliverto, Gonsai and Kosten, 2007; Puustinen, Lyyra, Metsapelto and Pulkkinen, 2008). Another paper interpreted a significant p value for Little's test as evidence that missing data was "non-ignorable" (Lang, Baltes and Wagner, 2007).

The makers of the popular statistical software SPSS have perhaps contributed to the confusion about the testability of MAR and MCAR hypotheses. The help file associated with its Missing Value Analysis procedure states that one of the functions of the procedure is to answer the question, "Are values missing randomly?" (SPSS, 2008). Neither the help file nor the output from the procedure itself makes explicit how SPSS intends to determine whether values are missing randomly, although there are options to obtain a p-value from Little's MCAR test, as well as the possibility to perform t tests for a common mean across patterns of missing data.

3.2. **Tests of MAR against a MNAR alternative.** As noted above, the MAR assumption cannot be tested against a general MNAR alternative. However, one way to get around this fact is to impose *a priori* assumptions about what the model for the missing data mechanism might be. For example, Zhou et al. (1999) propose a model-based likelihood ratio test of the MAR assumption versus a non-ignorable alternative. Such tests will depend on the non-ignorable likelihood being correctly

specified. However, these tests cannot provide any evidence for or against MAR in a model-free context. This point was made explicit in a recent paper by Molenberghs, Beunckens, Sotto and Kenward (2008). This paper demonstrates that every MNAR model has a MAR model with equivalent fit to the observed data. Hence, unless one has a strong *a priori* belief in the MNAR model, the results of likelihood ratio tests such as those of Zhou et al. cannot be interpreted as presenting evidence against the MAR hypothesis.

3.3. **Purported tests of MAR or "ignorability" that in fact test a weaker condition.** The methods described in this subsection have been presented as procedures for helping to determine if a missing data process either satisfies the MAR condition or is ignorable. Below, I demonstrate why these procedures cannot be regarded as testing the MAR condition (nor can they be regarded as testing ignorability).

3.3.1. *Katz.* Katz (2006) equates data that are missing randomly with "ignorable" missing data and describes a procedure for testing if data are "missing randomly". Katz describes the procedure as follows:

> To test whether observations are missing randomly assign each subject a value of 0 or 1 depending on whether or not the subject has one or more missing observations on the outcome. Using multivariable logistic regression, test the association between this variable and the independent variables and the prior values of the outcome. If the data are missing randomly, there should be no association between the independent variables and the prior values of the outcome, and whether or not the subject has missing values (Katz, 2006, p. 165).

It should be clear that this procedure is testing a condition which is equivalent to (OAR), namely $P(\mathbf{R}|\mathbf{y}_{obs}) = P(\mathbf{R})$. Depending on the analytic procedure, this condition may or may not be necessary for ignorability, however it certainly is not sufficient.

Nonetheless, it appears that some applied researchers are utilizing this procedure and interpreting it as a test of MAR, rather than of OAR. For instance, see the recent paper by Neron et al. (2007).

3.3.2. *Sherman.* Sherman (2000) presents tests of certain types of MAR and MCAR assumptions versus a generic MNAR hypothesis for the case of data which can be summarized by a *m* x *n* contingency table. However, Sherman's understanding of MAR differs from the usual understanding. Sherman defines MAR as the situation where, "the conditional distribution of the response variable given the explanatory variables for the complete data is the same as the corresponding conditional distribution for the full data (p. 365)." This definition is different from the one given in equation (2), and Sherman's test is therefore not a test of MAR as commonly understood. Furthermore, much like all other tests of MCAR, Sherman's test of

MCAR is in fact testing the OAR condition. A full discussion of Sherman's paper can be found in Rhoads (2009).

3.4. **Park and Davis.** Park and Davis (1993) propose a test meant to determine if the missing data mechanism is "ignorable" or not. Park and Davis are careful to distinguish their test from tests that merely compare a null hypothesis of MCAR against an alternative hypothesis of MAR, stating that "the alternative hypothesis for our methodology is the more general one that the missing data process is non-ignorable" (p. 637). The support for this claim seems to come from the fact that a missing data process can be consistent with Park and Davis' null hypothesis without being MCAR. Since the MCAR condition does not distinguish between response variables and explanatory variables, tests of MCAR will reject if the distribution of the covariate varies with the pattern of missingness. The null hypothesis for the Park and Davis test is that the data are "observed at random" conditional on a fully observed vector of covariates. Thus the Park and Davis null allows the covariate distribution to vary with pattern of missingess. Nonetheless, the alternative hypothesis space for the Park and Davis test includes MAR mechanisms. Thus, the Park and Davis test is exactly like other tests of the MCAR assumption, only with everything conditional on the fully observed covariate vector. A full description of the Park and Davis test, as well as an illustration of a non-ignorable missing data process that is consistent with Park and Davis' null hypothesis, is provided in the appendix.

3.4.1. *Donaldson and Moinpour.* Donaldson and Moinpour (2005) (hereafter DM) advance a graphical method for using the empirical data to determine the applicability of the MAR assumption. DM consider the case of longitudinal studies concerned with the measurement of quality of life (QoL) outcomes in cancer patients. For simplicity assume, as DM do, that only QoL outcome data are missing and that missingness is solely due to attrition over time. Thus, let $R_i = k$ indicate that subject i has completed the first k assessments (and hence is missing assessments at times $k+1, \ldots, K$). Let Y_{it} represent the latent QoL score of the i^{th} subject at the t^{th} time period and Q_i represent the baseline QoL quartile. DM propose plotting observed mean QoL scores at each time point, stratified by the observed value of R_i. If the observed means appear similar, DM argue that it is appropriate to proceed with an analysis assuming MAR. Formally, the DM procedure can be summarized as follows. The missingness process can be assumed MAR if and only if we have

$$E(\mathbf{Y}_{obs}|\mathbf{R}, \mathbf{Q}) \approx E(\mathbf{Y}_{obs}|\mathbf{Q}). \tag{11}$$

The $\approx$ symbol is used rather than an $=$ sign because the graphical nature of the DM method allows for some leeway in what would otherwise be an equality. However, notice that the condition given in equation (11) is the OAR assumption conditional on baseline QoL. So, much like the Park and Davis test, the DM procedure is not

testing MAR, but instead is testing for OAR, conditional on baseline QoL. Thus, the condition given in equation (11) is neither necessary nor sufficient for the missingness process to be MAR.

Other methods purporting to test MAR have been developed, and very likely will continue to be developed in the future. While the critiques presented here apply only to the particular procedures considered, the reader should keep in mind that similar problems must exist in any procedure claiming to test MAR or ignorability. There is simply no way the observed data can help to distinguish between a MAR and a MNAR missing data process.

4. ALTERNATIVES TO ASSUMING AN IGNORABLE MISSING DATA MECHANISM

If we cannot formulate a test to see if MAR is a plausible assumption and we are worried that the MAR assumption may not hold, what can we do? This section will explore three possible alternatives: modeling the MNAR missingness mechanism, computing non-parametric bounds for quantities of interest and sensitivity analysis. None of them is entirely satisfying, however each has its place in the diaspora of missing data methods.

4.1. **MNAR models.**

The first alternative has already been discussed briefly in the introductory section and involves formulating an explicit model for the missing data mechanism when the mechanism is believed to be MNAR. Modeling the missing data mechanism may be done in one of two ways, corresponding to two different factorizations of the complete-data likelihood. These two methods are *selection models* and *pattern-mixture models*.

4.1.1. *Selection models.*

Selection models factor the complete data likelihood in the following manner

$$P(\mathbf{Y}, \mathbf{R}) = P(\mathbf{Y}|\theta)P(\mathbf{R}|\mathbf{Y}, \phi). \tag{12}$$

Notice that the MAR and MCAR assumptions place restrictions on the form of the second factor in equation (12). Thus, methods that assume an ignorable missing data mechanism can be viewed as special cases of selection models. If the missing data mechanism is ignorable (MAR) then the last component of the above factorization can be ignored and maximum likelihood estimation can proceed by integrating the missing values out of the likelihood. If one is unwilling to maintain the MAR assumption, then a specific model must be postulated for the missing data mechanism. Identification is achieved because the $P(\mathbf{R}|\mathbf{Y}, \phi)$ term will elaborate a model that specifies a particular relation between the missing and observed data. A well known example of such a model is the "Heckman Selection Model" (Heckman, 1979). A monotonicity assumption is made so that it is assumed that observations above (below) a certain threshold are censored. If we assume that the probability of being observed looks like the normal cumulative distribution function (as in a

probit regression) then we can identify the model and obtain consistent estimates of complete data parameters of interest. However, inferences can be highly sensitive to the underlying normality assumption (Winship and Mare, 1992). Since this is an assumption about the parametric form of the missing data, it cannot be tested. Sensitivity to parametric form can be reduced if one makes an exclusion restriction (assuming that some variables predict missingness but are otherwise unrelated to the outcome). However, note that this exclusion restriction is also not testable from the data. It is generally true that estimates obtained from selection models are highly dependent on the correct specification of the model for missingness and are very sensitive to changes in parametric modeling assumptions (Schafer and Graham, 2002; Little and Rubin, 2002; Allison, 2002). Thus, just like models that assume an ignorable missing data mechanism, untestable assumptions are needed for inference to proceed.

4.2. **Pattern-Mixture Models.** In *pattern-mixture models* the complete data likelihood is factored as follows:

$$P(\mathbf{Y}, \mathbf{R}) = P(\mathbf{Y}|\mathbf{R}, \theta)P(\mathbf{R}|\phi). \tag{13}$$

Thus, the sample data are stratified by pattern of missing data and a separate model is specified for each pattern. This factorization of the likelihood is conceptually less appealing than the factorization used in selection models. However, one advantage of pattern-mixture models is that it is not necessary to specify a specific form for the missing data mechanism. Unfortunately, pattern-mixture models are in general not identified without the use of restrictions that relate parameters describing the incomplete cases to parameters describing the complete cases. To see this more clearly, it is useful to perform an additional factorization to equation (13) to obtain:

$$P(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{R}) = P(\mathbf{Y}_{obs}|\mathbf{R}, \theta, \phi)P(\mathbf{R}|\phi, \theta)P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \mathbf{R}, \theta, \phi). \tag{14}$$

The final factor in equation (14) is not identified from observed data. The presentation in (14) makes clear that one way identification can be achieved is by assuming that the complete data and missing data models share certain parameters.

4.3. **Nonparametric Bounds.** Manski (1995, 2003, 2007) proposes a fundamentally different approach to inference in the face of missing data. Rather than making untestable identifying assumptions, Manski proposes that we begin our inquiry by asking what we can learn from the observed data alone, without any additional assumptions. This approach involves computing nonparametric "worst-case" bounds on population quantities of interest, rather than point estimates of these quantities.

The idea of computing bounds in this context has existed at least since Cochran, Mosteller and Tukey (1954, pp. 274-282). The idea was discussed in the context of estimating population proportions in Cochran (1977, p. 361-3). However, the idea received little attention until Manski (1989).

Computing worst-case bounds is remarkably simple, and is most easily illustrated with reference to a binary outcome variable, where "1 indicates a "success" and "0" indicates a "failure". The lower bound on the proportion of successes is computed by simply assuming all of the missing data consists of "0s". The upper bound assumes that all of the missing data consists of "1s".

A common objection to the reporting of nonparametric bounds is that the bounds are so wide as to be uninformative (Cochran, 1977; Raghunathan, 2000). However, Manski (2007) has argued that even when these bounds are wide there is value in reporting them. First, while all may agree that the nonparametric bounds are based on overly conservative, "extreme", assumptions about the missing data process, there may be considerable disagreement about how much less conservative we ought to be. In a context such as this the nonparametric bounds serve to establish a "domain of consensus" (Manski, 2007) among researchers. Second, reporting bounds allows consumers of quantitative policy analysis to understand which conclusions are justified on the basis of the observed data alone and how those conclusions might be strengthened by making additional, untestable (but nonetheless possibly credible) assumptions.

Manski (2003) shows that the interval defined by the worst-case bounds can be narrowed by imposing certain monotonicity assumptions that are much weaker than the MAR assumption, but may also be correspondingly more credible. Additionally, recent applied work shows that, in certain situations, analysis in terms of nonparametric bounds can provide useful answers to important policy questions. For instance, Pepper (2000) uses this method to examine the impact that growing up in a family that receives Aid to Families with Dependent Children (AFDC) has on likelihood of future welfare receipt, and Gonzalez (2006) uses the approach to estimate the impact of limited English proficiency on employment opportunities for Hispanic workers in the United States.

While the bounds approach may be appropriate for some problems, there are limitations. The bounds approach is most easily applied to the analysis of binary outcome data and other situations where there is a natural bound for the variable of interest. Applications to ordinal outcomes have been developed, but the procedures involved are quite complex (Scharfstein, Manski and Anthony, 2003). While theoretical results exist for the case of conditional prediction with continuous outcomes, multiple conditioning variables and a general missing data pattern (Manski, 2003, p. 49), it is virtually impossible to formulate these results in a form that is usable for applications.

4.4. **Sensitivity Analysis.** Sensitivity analysis may provide a useful middle road between the strong and untestable identifying assumptions associated with MAR and MNAR models and the extremely conservative bounding approach which imposes no assumptions whatsoever on the distribution of the missing data. When

conducting a sensitivity analysis, different choices are made for the values of non-identified parameters and different ignorable and non-ignorable missing data models are fit. At the extreme, where all possible models consistent with the observed data are considered, the sensitivity analysis simply replicates the non-parametric bounds approach described previously.

However, more frequently a sensitivity analysis entertains a limited number of alternative models and/or a limited number of values for non-identified parameters. This approach is not objectionable provided researchers are circumspect about the claims that they make at the end of the day. Too frequently a very limited number of models are fit and the consistency of estimates across these models is taken as evidence that the estimates must be accurate. For instance, Fairclough et al. (1998) fit two different MAR models and two different non-ignorable models to QoL data from a breast cancer study. The estimates obtained from the different models were similar. Fairclough et al.(1998) use this fact as the foundation for the claim that, "The assumption that the data were missing at random (MAR) was adequate to obtain unbiased estimates." But there is no real support for this claim. The fact that the estimates from the various methods were alike doesn't mean they were unbiased. The estimates may have all been biased in the same direction.

A further problem that afflicts some sensitivity analyses is that the parameters varied in the sensitivity analysis frequently have no useful substantive interpretation. For instance, Raghunathan (2000) describes a sensitivity analysis of outcomes from the National Comorbidity Study (NCS) conducted within a Bayesian framework. He shows how the 95% posterior interval varies with different assumed values of the parameters important to the sensitivity analysis. However, no justification is given for using particular values for the unknown parameters, nor is any framework proposed that would allow one to determine what values of these parameters might be plausible.

Nonetheless, when done carefully sensitivity analysis can be a useful tool for the analysis of incomplete data. It is most useful when the parameter(s) varied for the purposes of the sensitivity analysis have clear interpretations in terms of the policy context. It is also helpful if the parameters are well understood by policy experts. If this is the case then the space of plausible non-ignorable models can be substantially reduced by consulting expert opinion and/or by gathering auxiliary data. One helpful approach is to change the value of the parameter of interest until the primary estimand of interest changes sign (or perhaps changes from statistically significant to no longer significant). Conducting the sensitivity analysis in this fashion obviates the need for experts to sketch out an entire range of plausible parameter values. Instead only the plausibility of the parameter value that causes the change in sign (or significance) needs to be evaluated.

5. CONCLUSION

Assuming that the missing data mechanism is MAR or MCAR is not a way to get around the untestable assumptions needed to identify MNAR models, since MAR and MCAR are themselves untestable assumptions. One should not be fooled by the tests that have appeared in the literature into thinking that these tests can help us determine which procedure is best for handling missing data in the context of a particular policy study. All methods (with the exception of the bounds approach) rely on untestable assumptions about the missing data mechanism. The bounds approach is frequently non-informative, implying the need to make some sort of assumption about the distribution of the missing data. The relatively weak assumptions made in sensitivity analysis may be the best we can do. However, the decision about which method to use cannot be informed by the data and instead must be made on the basis of a deep understanding of the particular policy context.

Appendix: Details of the Park and Davis test

Park and Davis (1993) develop their test within the context of a model for categorical repeated measures data originally introduced by Grizzle, Starmer and Koch (GSK) (1969), and described in Woolson and Clarke (1984). The model assumes that interest is in modeling a response variable which is discrete with $G-1$ possible response categories. This response will be measured at d time points, or more generally, under d different conditions. Thus, there are a total of $s = (G-1)^d$ distinct response profiles. Missing data on the categorical response is simply regarded as another response category, so that when there is missing data, there are a total of $s = G^d - 1$ distinct response profiles (since the response profile with all data missing is not modeled). Let there be H different strata in the data. For the h^{th} stratum, let $\mathbf{p}_h$ be the s-dimensional vector of sample cell proportions having covariance matrix $\mathbf{V}_h$. Let $\mathbf{F}(\mathbf{p}_h)$ be the vector of response functions of interest. Then, regression models of interest are defined within each stratum as

$$\mathbf{E}[\mathbf{F}(\mathbf{p}_h)] = \mathbf{X}_h\beta_h, \quad \text{for } h = 1, \ldots, H. \tag{15}$$

Define $\mathbf{p}' = (\mathbf{p}'_1, \ldots, \mathbf{p}'_H)$ and $\mathbf{V} = var(\mathbf{p}) = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_H)$. Thus, the model given in equation (15) can be written as

$$\mathbf{E}[\mathbf{F}(\mathbf{p})] = \mathbf{X}\beta,$$

where $\mathbf{X} = \text{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_H)$ and $\beta' = (\beta'_1, \ldots, \beta'_H)$. It is assumed that values in $\mathbf{X}$ are never missing. To facilitate later discussion I also introduce the notation $\mathbf{p}_L$ to stand for the vector of cell proportions that would have been observed had there been no missing data and $\mathbf{X}_L$ to stand for the $\mathbf{X}$ matrix that would have been used if there had been no missing data.

Park and Davis suggest the following as a test of the ignorability of the missing data mechanism: Define strata based on patterns of missing data, and then

test for the homogeneity of the model parameters across strata. Specifically, define

$$H_0 : \quad \beta_1 = \ldots = \beta_H.$$

Note that H_0 can be expressed as $\mathbf{C}\beta$ for appropriately chosen $\mathbf{C}$, and H_0 can be tested via the Wald statistic:

$$Q_c = (\mathbf{Cb})' \left[\mathbf{C} \left(\mathbf{X'V^{-1}X} \right)^{-1} \mathbf{C}' \right] \mathbf{Cb}. \tag{16}$$

If we define ν to be the dimension of the β_h vectors (where ν is assumed to be constant across strata), then Q_c has a large sample chi-square distribution under H_0 with $\nu(H-1)$ degrees of freedom.

Park and Davis claim that, "if the model results are not significantly different across strata, then it is reasonable to conclude that the missing data mechanism is ignorable" (p. 632). Their justification in claiming to test ignorability vs. nonignorability appears to be the fact that their test will not reject H_0 when the distribution of $\mathbf{X}_h$ varies across missing data patterns, whereas tests of the MCAR assumption would reject H_0 in this case.

It is true that the missing data process can be consistent with Park and Davis' null hypothesis without being MCAR. The full force of the MCAR assumption is not needed for ignorability in this case because interest is only in the conditional distribution of $\mathbf{p}_L$ given $\mathbf{X}_L$. MCAR makes no distinction between response variables and explanatory variables, and so it requires $f(\mathbf{X}_L, \mathbf{p}_L|\mathbf{R}) = f(\mathbf{X}_L, \mathbf{p}_L)$. Thus, tests of MCAR will reject if the marginal distribution of $\mathbf{X}_L$ varies with $\mathbf{R}$. On the other hand, if we distinguish $\mathbf{F}(\mathbf{p}_L)$ as the vector of response variables and $\mathbf{X}_L$ as a matrix of explanatory vectors, then, in the spirit of Little (1995), we might define *covariate dependent missingness* (CDM) as follows:

$$f(\mathbf{F}(\mathbf{p}_L)|\mathbf{X}_L, \mathbf{R}) = f(\mathbf{F}(\mathbf{p}_L)|\mathbf{X}_L). \qquad \text{(CDM)} \tag{17}$$

Dividing $\mathbf{F}(\mathbf{p}_L)$ into sub-vectors so that $\mathbf{F}(\mathbf{p}_L) = (\mathbf{F}(\mathbf{p}_{obs}), \mathbf{F}(\mathbf{p}_{mis}))$ it is clear that the null hypothesis of the Park and Davis test is equivalent to

$$f(\mathbf{F}(\mathbf{p}_{obs})|\mathbf{X}_L, \mathbf{R}) = f(\mathbf{F}(\mathbf{p}_{obs})|\mathbf{X}_L). \qquad \text{(COAR)} \tag{18}$$

This is the exact same "observed at random" relation described in equation (10), except we now condition on the fully observed $\mathbf{X}_L$. However, this COAR condition is only part of what is needed to justify CDM, and hence, ignorability. The condition

$$f(\mathbf{F}(\mathbf{p}_{mis})|\mathbf{X}_L, \mathbf{R}) = f(\mathbf{F}(\mathbf{p}_{mis})|\mathbf{X}_L, \mathbf{F}(\mathbf{p}_{obs})), \qquad \text{(MAR)} \tag{19}$$

is also needed. Equation (19) is simply the alternative characterization of MAR given in equation (9). It now becomes clear that the alternative hypothesis space for Park and Davis' test includes MAR. Park and Davis test a necessary but not sufficient condition for ignorability, a condition that we might call conditional OAR.

	Observed Data							
$t=1$	N	N	S	S	N	S	M	M
$t=2$	N	S	N	S	M	M	N	S
Treat	50	20	30	20	35	25	4	2
Cont	50	20	30	20	35	25	4	2

Table 1(a)

	Latent Data			
$t=1$	N	N	S	S
$t=2$	N	S	N	S
Treat	82	26	52	26
Cont	56	52	52	26

Table 1(b)

Example In order to illustrate that a nonignorable missing data processes can produce observed data that is perfectly consistent with Park and Davis' null hypothesis I present the following hypothetical example. Imagine a policy study meant to evaluate the effectiveness at preventing smoking in teens of a tobacco education program. Participants are randomly assigned to the tobacco education class or to a different class not focused on the ill effects of smoking. Prior to the class, participants are asked if they have smoked a cigarette in the last 7 days, and they are asked the same question again after completion of the class. Interest in is the proportion of subjects smoking at times $t = 1, 2$ for both the treatment and control groups. Values may be missing at either the first or second time point. The letter "N" is used to denote "not smoking", "S" to denote "smoking", and "M" to denote "missing." Then the data might look like the "Observed data" table in Table 1(a).

To perform Park and Davis' test, we divide the data into two strata, one for complete cases and one for incomplete cases. $\mathbf{F}(\mathbf{p}_1)$ is a 4×1 dimensional response vector for complete cases. Its first two elements are the proportions smoking at times $t = 1, 2$ for the treatment group and its second two elements are the proportions smoking at times $t = 1, 2$ for the control group. $\mathbf{F}(\mathbf{p}_2)$ is defined analogously for incomplete cases except for proportions smoking at a given time t are expressed relative to the total number of subjects giving data at time t. $\mathbf{X}_h$ is simply the 4×4 identity matrix for $h = 1, 2$. Then, using the GSK methodology, the estimates of β_h are given in table 2(a).

Note that the parameter estimates are exactly the same for the complete and incomplete cases. The Q_c statistic is thus 0, and when compared to the χ^2 distribution with 4 degrees of freedom we find a p value of 1. Thus, as far as the Park and Davis test is concerned, there can be no doubt that the missing data mechanism is ignorable.

Observed Data proportion S			
		Complete	Incomplete
Treat	$t = 1$	.417	.417
	$t = 2$	.333	.333
Cont	$t = 1$	.417	.417
	$t = 2$	.333	.333

Table 2(a)

Latent Data proportion S		
Treat	$t = 1$	.419
	$t = 2$	.279
Cont	$t = 1$	.419
	$t = 2$	.419

Table 2(b)

Now suppose that if we could have observed the missing values we would have obtained the data in Table 1(b). The data in Table 1(b) is perfectly consistent with the data in 1(a). Tables 1(a) and 1(b) would result from a missingness process where missingness at time $t = 1$ is random for both treatment and control groups, but in the control group the probability of missingness at time $t = 2$ is higher for those who are smoking whereas in the treatment group the probability of missingness at time $t = 2$ is higher for those who are not smoking. If we had been able to see the latent data in Table 1(b) we would have obtained the parameter estimates in Table 2(b). The interpretation of the results of the experiment based on the latent data would be very different than an interpretation based on the observed data. An analysis focusing only on the observed data would conclude that randomization was successful (since the proportions smoking at $t = 1$ are equal for both the treatment and control group), that there was no effect of treatment, and that both treatment and control participants had reduced smoking over time. In contrast, the analysis of the latent data would have found a strong reduction in smoking across time in the treatment group and no change in the control group. Clearly the missing data mechanism is not ignorable in this case.

REFERENCES

[1] Allison, P.D. (2002) *Missing Data*. Thousand Oaks, CA: Sage.

[2] Buckler, K. and Unnever, J.D. (2008). Racial and ethnic perceptions of injustice: testing the core hypotheses of comparative conflict theory. *Journal of Criminal Justice*. 36, 270-278.

[3] Chen, H.Y. and Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*. 86(1), 1-13.

[4] Cochran, W. (1977). *Sampling Techniques*, Third edition. New York: Wiley.

[5] Cochran, W.F., Mosteller, F., and Tukey, J. (1954). *Statistical problems of the Kinsey report on sexual behavior in the human male*. Washington, D.C.: American Statistical Association.

[6] Donaldson, G. W., and Moinpour, C. M. (2005). Learning to live with missing quality-of-life data in advanced-stage disease trials. *Journal of Clinical Oncology*. 23 (30), 7380-7384.

[7] Evans, W. and Farrelly, M. (1998). The compensating behavior of smokers: Taxes, Tar and Nicotine. *The RAND Journal of Economics*. 29(3), 578-595.

[8] Fairclough, D. L., Peterson, H. F., Cella, D, and Bonomi, P. (1998). Comparison of several model-based methods for analyzing incomplete quality of life data in cancer clinical trials. *Statistics in Medicine* 17, 781-796.

[9] Fitzmaurice, G. (2008). Missing data: implications for analysis. *Nutrition*. 24, 200-202.

[10] Flook, L., Repetti, R. and Ullman, J. (2005). Classroom Social Experiences as Predictors of Academic Performance. *Developmental Psychology*. 41(2), 319-327.

[11] Gadbury, G.L., Coffey, C.S. and Allison, D.B. (2003). Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. *Obesity Reviews*. 4, 175-184.

[12] Gibb, B.F., Beevers, C.G., Andover, M.S. and Holleran, K. (2006). The hopelessness theory of depression: a prospective multi-wave test of the vulnerability-stress hypothesis. *Cogn Ther Res*. 30, 763-772.

[13] Gonzalez, L. (2005) Nonparametric bounds on the return to language skills. *Journal of Applied Econometrics*. 20: 771-795.

[14] Gonzalez, G., Desai, R., Sofuoglu, M., Poling, J., Oliveto, A., Gonsai, K. and Kosten, T. (2007). Clinical efficacy of gabapentin versus tiagabine for reducing cocaine use among cocaine dependent methadone-treated patients. *Drug and Alcohol Dependence*. 87, 1-9.

[15] Graham, J.W. (2009). Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.*. 60: 549-576.

[16] Grella, C.E., Scott, C.K., Foss, M.A., and Dennis, M.L. (2008) Gender similarities and differences in the treatment, relapse and recovery cycle. *Evaluation Review*. 32, 113-137.

[17] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*. 25, 489-504.

[18] He, Y. (2010) Missing data analysis using multiple imputation: Getting to the heart of the matter. *Circulation: Cardiovascular Quality and Outcomes*. 3, 98-105.

[19] Heckman, J. (1979). Sample selection bias and specification error. *Econometrica*. 47(1). 153-61.

[20] Hedeker, D. (1999). MIXNO: A computer program for nominal mixed-effects regression. *Journal of Statistical Software*. 4, 1-92.

[21] Katz, M.H. (2006). *Multivariable Analysis: A Practical Guide for Clinicians, Second edition*. Cambridge: Cambridge University Press.

[22] Kunina, O., Wilhelm, O., Formazin, M., Jonkman, K. and Schroeders, U. (2007). Extended criteria and predictors in college admission: Exploring the structure of study success and investigating the validity of domain knowledge. *Psychology Science*. 49(2), 88-114.

[23] Lang, F., Baltes, P. and Wagner, G. (2007). Desired lifetime and end-of-life desires across adulthood from 20 to 90: A dual-source information model. *Journal of Gerontology: Psychological Sciences*. 62B(5), P268-P276.

[24] Lewis-Beck, M.S., Bryman, A. and Liao, T.F. (2004). *The Sage encyclopedia of social science research methods*. Thousand Oaks, CA: Sage.

[25] Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1994). Weighted lest squares analysis of repeated categorical measurements with outcomes subject to nonresponse. *Biometrics*. 50, 11-24.

[26] Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*. 83 (404), 1198-1202.

[27] Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data: second edition*. Hoboken, NJ: Wiley.

[28] Manski, C.F. (1989). Anatomy of the selection problem. *The Journal of Human Resources*. 24(3), 343-360.

[29] Manski, C.F. (1995). *Identification problems in the social sciences*. Cambridege, MA: Harvard University Press.

[30] Manski, C. F. (2003). *Partial identification of probability distributions*. New York: Springer.

[31] Manski, C. F. (2007). *Identification for Prediction and Decision*. Cambridege, MA: Harvard University Press.

[32] Molenberghs, G, Beunckens, C., Sotto, C. and Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J.R. Stat. Soc. B*. 70(2) 371-388.

[33] Myers, W.R. (2000). Handling missing data in clinical trials: an overview. *Drug Information Journal*. 34, 525-533.

[34] Neron, S., Correa, J.A., Dajczman, E., Kasymjanova, G., Kreisman, H. and Small, D. (2007). Screening for depressive symptoms in patients with unresectable lung cancer. *Supportive Care in Cancer*. 15:1207-1212.

[35] Park, T. and Davis, C. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*. 49(2), 631-638.

[36] Park,T. and Lee, S. (1997). A test of missing completely at random for longitudinal data with missing observations. *Statistics in Medicine*. 16, 1859-1871.

[37] Pepper, J. (2000). The intergenerational transmission of welfare receipt: A nonparametric bounds analysis. *Review of Economics and Statistics*. 82(3), 472-88.

[38] Peugh, J.L. and Enders, C.K. (2004). Missing data in educational research. *Review of educational research*. 74(4), 525-556.

[39] Puma, M. J., Olsen, R.B., Bell, S.H. and Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials*. (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

[40] Puustinen, M., Lyyra, A., Metsapelto, R. and Pulkkinen, L (2008). Children's help seeking: The role of parenting. *Learning and Instruction*. 18, 160-171.

[41] Raghunathan (2000). Comment on "Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*. 95(449), 85-87.

[42] Rhoads, C. (2009). Comment on "Tests of Certain Types of Ignorable Nonresponse in Surveys Subject to Item Nonresponse or Attrition", Institute of Policy Research working paper, # WP-09-10.

[43] Rubin, D. B. (1976). Inference and missing Data. *Biometrika*. 63 (3), 581-592.

[44] Salmon, P., Holcombe, C., Clark, L., Krespi, R., Fisher, J. and Hill, J. (2007). Relationships with clinical staff after a diagnosis of breast cancer are associated with patients' experience of care and abuse in childhood. *Journal of Psychosomatic Research*. 63, 255-262

[45] Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*. 7 (2), 147-177.

[46] Scharfstein, D., Manski, C. and Anthony, J. (2004). On the construction of bounds in prospective studies with missing ordinal outcomes: Application to the Good Behavior Game trial. *Biometrics*. 60(1), 154-64.

[47] Sherman, R. P. (2000). Tests of certain types of ignorable nonresponse in surveys subject to item nonresponse or attrition. *American Journal of Political Science*. 44 (2), 356-368.

[48] Simmons, D., Hairelll, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V., Rupley, W. and Byrns, G. (2010). A Comparison of Multiple-Strategy Methods: Effects on Fourth-Grade Students' General and Content-Specific Reading Comprehension and Vocabulary Development *Journal of Research on Educational Effectiveness*. 3, 121-156.

[49] SPSS, Inc. (2008). SPSS Statistics Release 17.0.1, Missing Value Analysis Help file.

[50] Steginga, S., Pinnock, C., Gardner, M., Gardiner, R.A. and Dunn, J. (2005). Evaluating peer support for prostate cancer: the Prostate Cancer Peer Support Inventory. *BJU International*. 95, 46-50.

[51] Sundin, E.C. and Ogren, M. (2006). Supervisees' and supervisors' experiences of group climate in group supervision in psychotherapy: Effects of admission procedure. *Issues in Educational Research*. 16, (online).

[52] Troxel, A. B. Fairclough, D. L. Durran, D., and Hahn, E. (1998). Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine*. 17, 653-666.

[53] Van Ness, P., Murphy, T., Araujo, K., Pisani, M. and Allore, H. (2007). The use of missingness screens in clinical epidemiologic research has implications for regression modeling. *Journal of Clinical Epidemiology*. 60, 1239-45.

[54] Winship, C. and Mare, R. (1992). Models for sample selection bias. *Annual Review of Sociology*. 18, 327-350.

[55] Woolson, R.F. and Clarke, W.R. (1984). Analysis of categorical incomplete longitudinal data. *Journal of the Royal Statistical Society, Series A*. 147, 87-99.

[56] Zhou, X., Castellussio, P, Hui, S. L., and Rodenbery, C. A. (1999). Comparing two prevalence rates in a two-phase design study. *Statistics in Medicine*. 18, 1171-1182.