Institute for Policy Research
Northwestern University
Working Paper Series

WP-09-10

Comment on "Tests of Certain Types of Ignorable Nonresponse in Surveys Subject to Item Nonresponse or Attrition"

Christopher H. Rhoads
Postdoctoral Fellow, Institute for Policy Research
Northwestern University

DRAFT
Please do not quote or distribute without permission.

2040 Sheridan Rd. ◆ Evanston, IL 60208-4100 ◆ Tel: 847-491-3395 Fax: 847-491-9916
www.northwestern.edu/ipr, ◆ ipr@northwestern.edu

Abstract

The current paper points out some problems with the paper by Sherman (2000) referenced in the title. Misunderstandings about the terms Missing at Random (MAR) and Missing Completely at Random (MCAR) are clarified. A necessary and sufficient condition to justify a complete case analysis of bivariate, binary data when interest is in the conditional distribution of one variable given the other is presented. The non-existence of a test for MAR is noted. The impossibility of testing a condition that is sufficient to ensure unbiased estimates from an analysis of complete cases is also noted. Hence Sherman's proposed tests of ignorable nonresponse are falsified.

1. Introduction

The existence of missing data is a common problem for researchers working with data from surveys. Individuals in the original sampling frame may not respond to the survey at all, they may leave certain items blank, or, in the case of panel surveys, they may respond to earlier waves of the survey but not respond at later waves. The manner in which missing data should be handled by the data analyst will depend on two things: the nature of process that causes missing data and the type of inferences that will be made from the data.

The vast majority of procedures for handling missing data assume that the missingness process is *ignorable* in a particular sense. For instance, a well known book by Little and Rubin (2002) devotes only one of its fifteen chapters to methods that can be used when the missing data process is not ignorable.

Rubin (1976) discussed conditions that are necessary to justify ignoring the missing data process. Two key conditions are crucial to determining the ignorability of a missing data process with respect to a particular analytical procedure. They are the ideas of *missing at random* (MAR) and *observed at random* (OAR). Missing data generated by a process that is both MAR and OAR is called *missing completely at random* (MCAR).

There appears to be some confusion in the literature about (a) the meaning of the terms MAR and MCAR, (b) conditions under which it is justifiable to base an analysis on only the subset of respondents for whom there are no missing values for any variable of interest (the

so-called *complete cases*), and (c) the extent to which the assumption that a missing data process is MAR or MCAR can be tested. Confusion about all of the above issues is evident in the paper by Sherman (2000) referenced in the title of this work.

Sherman (2000) discusses the case of missing values arising in the context of bivariate categorical data. He seeks to determine when a complete case analysis is justified. In service of this goal, Sherman defines missingness processes that he describes as special cases of MCAR or MAR missingness. Sherman then claims to characterize these missingness processes in terms of odds ratios calculable from the observed data. On the basis of these characterizations, Sherman presents statistics that he claims could be used to test whether a missing data process is MCAR or MAR in the sense that Sherman intends.

The current paper uses a critique of the Sherman paper as a means of demonstrating the confusion about the points listed above, and, hopefully, as a means of clarifying some of this confusion. I assume throughout (as does Sherman) that the primary interest is in the conditional distribution of a response variable given an explanatory variable. The paper is organized as follows.

Section 2 notes that Sherman's understanding of the term "missing at random" differs from its original definition in Rubin (1976) and from its common usage by statisticians. While MAR as usually defined is neither necessary nor sufficient to ensure unbiased estimates from a complete case analysis, Sherman's MAR (which will be denoted MAR_S in the following) is a sufficient but not necessary condition to

ensure unbiased estimates from a complete case analysis. This section also provides a necessary and sufficient condition for obtaining unbiased estimates from a complete case analysis when interest is in the conditional distribution of a response variable given an explanatory variable. From here forward I shall term a missing data process that is ignorable with respect to a complete case analysis a "CCI" process.

Section 3 begins by noting that it is impossible to test the MAR hypothesis using the observed data, and it is also impossible to establish a necessary and sufficient condition for CCI using only the observed data. In view of these facts, it is clear that Sherman's proposed tests of ignorable nonresponse must fail. This section describes why Sherman's proposed tests fail to test a necessary and sufficient condition for any of the following: (a) MAR, (b) MAR_{S} , (c) MCAR, (d) CCI. The final section summarizes.

2. MCAR, MAR, MAR_{S}, and a necessary and sufficient condition for ignorability

Crucial to understanding when a missing data process may be ignorable are the concepts of missing at random (MAR), observed at random (OAR) and missing completely at random (MCAR). Before defining these terms it is useful to have in mind the distinction between the *latent data* and the *observed data*. Assume that the data can be arranged into a "units" by "variables" matrix in the usual fashion. The latent data, denoted $\mathbf{Z^L}$, is the data that would have been observed had there been no missing data. Let the ij^{th} element of the matrix $\mathbf{M}$ be defined by $\mathbf{M}_{ij} = 1$ if data is missing for the j^{th} variable

measured on the i^{th} unit and $\mathbf{M}_{ij} = 0$ otherwise. $\mathbf{M}$ is referred to as the missing data matrix. Then the observed data, $\mathbf{Z^O}$, is defined by $\mathbf{Z^O}_{ij} = \mathbf{Z^L}_{ij}$ if $\mathbf{M}_{ij} = 0$ and $\mathbf{Z^O}_{ij} = *$ if $\mathbf{M}_{ij} = 1$. The symbol "*" is used as a placeholder to indicate missing data.

Let $\mathbf{Z^L}_{obs}$ = the entries in $\mathbf{Z^L}$ such that $\mathbf{Z^L} = \mathbf{Z^O}$ and $\mathbf{Z^L}_{mis}$ = the entries in $\mathbf{Z^L}$ for which $\mathbf{Z^O} = *$. We characterize the random matrix $\mathbf{M}$ by its conditional distribution given the latent data and index this distribution by a parameter vector ϕ. Then the data are missing at random if

$$\text{(MAR)} \qquad P(\mathbf{M}|\mathbf{Z^L}, \phi) = P(\mathbf{M}|\mathbf{Z^L}_{obs}, \phi) \text{ for all } \mathbf{Z^L}_{mis},\ \phi.$$

The data are observed at random if

$$\text{(OAR)} \qquad P(\mathbf{M}|\mathbf{Z^L}, \phi) = P(\mathbf{M}|\mathbf{Z^L}_{mis}, \phi) \text{ for all } \mathbf{Z^L}_{obs},\ \phi.$$

The data are missing completely at random if both MAR and OAR hold, which implies that

$$\text{(MCAR)} \qquad P(\mathbf{M}|\mathbf{Z^L}, \phi) = P(\mathbf{M}|\phi) \text{ for all } \mathbf{Z^L},\ \phi.$$

In the above equations, $P(.)$ denotes a generic probability distribution over its argument(s).

The definitions given in the preceding paragraph are derived from Rubin (1976) and conform with the usual understanding of these terms in the statistics literature (see, e.g., Allison, 2002 or Little and Rubin, 2002). Sherman provides somewhat different definitions of MCAR and

MAR. Sherman describes MCAR as occurring when "the joint distribution of the complete units is equal to the joint distribution of the sample (p.362)." He 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)." Note that the definition of MAR given in equation (MAR) makes no reference to "explanatory" or "response" variables. Hence, Sherman's understanding of MAR differs from the usual understanding of that term, and I use the notation MAR_{S} to refer to Sherman's version of MAR from here forward. On the other hand, Sherman's description of MCAR is equivalent to the definition given in the equation labeled (MCAR).

Sherman defines particular types of MCAR and MAR_{S} missingness within the context of a hypothetical survey of U.S. citizens where the researcher is interested in the relationship between race (black or non black) and turnout (voter or nonvoter). It is assumed that values of race, turnout or both may be missing. Adapting the notation introduced above to this specific example, the missing data vector for subject i may be written $M_i = (M_{i1}, M_{i2})$, with M_{i1} indexing missingness on race and M_{i2} indexing missingness on turnout. The observed data vector for subject i is given by $Z_i^O = (R_i^O, T_i^O)$ with $R_i^O \in (B, \overline{B}, *_R)$ and $T_i^O \in (V, \overline{V}, *_T)$. The latent data vector for subject i is given by $Z_i^L = (R_i^L, T_i^L)$ with $R_i^L \in (B, \overline{B})$ and $T_i^L \in (V, \overline{V})$. Matrices representing the observed and latent data may be constructed by stacking the relevant row vectors.

Sherman defines the MCAR(item) condition to hold if and only if:

(1) The $M_{ij}$are independent for $i = 1, \ldots, n;\ j = 1, 2$.

(2) For $j = 1, 2$ The M_{ij} are identically distributed for $i = 1, \ldots, n$.

Note that MCAR(item) is a special case of MCAR as defined in equation (MCAR) since

$$\begin{aligned}
& P(M_{i1}, M_{i2}|R_i^L, T_i^L) \\
&= P(M_{i1}|R_i^L, T_i^L)P(M_{i2}|R_i^L, T_i^L) && \text{by (1)} \\
&= P(M_{i1})P(M_{i2}) && \text{by (2)} \\
&= P(M_{i1}, M_{i2}). && \text{by (1)}
\end{aligned}$$

Hence, $P(\mathbf{M}|R^L, T^L) = P(\mathbf{M})$ as required by MCAR. The first equality given above is justified because the true values of race and turnout are constant for each value of i. Thus, the independence of M_{i1} and M_{i2} for each i implies conditional independence of M_{i1} and M_{i2} given R^L and T^L.

Sherman defines the condition MCAR(unit) to hold if and only if:

(1) $M_{i1} = M_{i2}, i = 1, \ldots, n$.

(2) $M_{i1}, i = 1, \ldots, n$, are independent and identically distributed

Note that MCAR(unit) is a special case of MCAR since

$$\begin{aligned}
P(M_{i1}, M_{i2}|R_i^L, T_i^L) &= P(M_{i1}|R_i^L, T_i^L) && \text{by (1)} \\
&= P(M_{i1}) && \text{by (2)} \\
&= P(M_{i1}, M_{i2}). && \text{by (1)}
\end{aligned}$$

Sherman defines the MAR_S(item) condition to hold if and only if:

(1) $M_{ij}, i = 1, \ldots, n, j = 1, 2$ are independent.

(2) $M_{i2}, i = 1, \ldots, n$ are identically distributed.

(3) $M_{i1}, i = 1, \ldots, n$ are identically distributed within race category.

Sherman defines the condition MAR_S(unit) to hold if and only if:

(1) $M_{i1} = M_{i2}, i = 1, \ldots, n$.

(2) $M_{i1}, i = 1, \ldots, n$, are independent, and, within race categories, identically distributed.

Unlike Sherman's MCAR conditions, which are, in fact, implied by MCAR, Sherman's MAR_S conditions are not implied by MAR, nor do they imply MAR. For instance, a missingness process where $P(M_i = (1,0)|R_i^L, T_i^L = V) \neq P(M_i = (1,0)|R_i^L, T_i^L = \overline{V})$ is consistent with MAR (since turnout is observed when $M_i = (1,0)$) . However, such a missingness process is not consistent with either MAR_S(item) or MAR_S(unit) (unless the correlation between T_i^L and R_i^L is one). On the other hand, both MAR_S(item) and MAR_S(unit) allow the probability of a missing value for the race variable to depend on the potentially unobserved value of race, which is a violation of MAR.

Sherman's "item" and "unit" types of MCAR and MAR_S can be regarded as endpoints of a continuum between the case where $Corr(M_{i1}, M_{i2}) = 0$ (the "item" type) and the case where $Corr(M_{i1}, M_{i2}) = 1$ (the "unit" type). In fact, it seems unlikely that either of these extremes would ever hold in practice. One would expect that individuals for whom the value of one variable is missing would also be more likely to be

missing the value of another variable (hence $Corr(M_{i1}, M_{i2}) > 0$). On the other hand, it will likely be the case that some individuals will have data values recorded for some variables but not for others (hence $Corr(M_{i1}, M_{i2}) < 1$). Of course, the case where $Corr(M_{i1}, M_{i2}) < 0$ is also possible, but this possibility seems unlikely. The next subsection will show that the correlation between the individual missingness indicators is immaterial for the purposes of determining if CCI holds.

2.1. **Complete Case Ignorability.** The differences between MAR_S and MAR are worthy of note, but the larger question of how these assumptions about the missingness process relate to complete case ignorability remains. It appears to be widely acknowledged that in order to ensure unbiased estimates in an analysis based on only complete cases, the MCAR condition must hold (Peugh and Enders, 2004; Schafer and Graham, 2002) . While this point is in general true, if interest is only in the conditional distribution of a response variable given an explanatory variable, the following, weaker, condition is sufficient.

Theorem 1. *A necessary and sufficient condition for complete case ignorability with respect to the conditional distribution of* T^L *given* R^L *is given by*

$$P(M_i = (0,0)|R_i^L, T_i^L) = P(M_i = (0,0)|R_i^L). \tag{1}$$

In words, equation (1) states that the probability of a complete case may depend on the unobserved value of the independent variable, but it may not depend on the value of the dependent variable, regardless of whether the value of the dependent variable is observed or not.

A proof of the above theorem may be found in Allison (2002, p. 87). In fact, Allison shows that the result given in theorem 1 holds more generally for any regression of a dependent variable Y on a set of independent variables $X_1, \ldots, X_q$. That is, complete case ignorability holds provided that the probability of a complete case conditional on the independent variables doesn't depend on the latent value of the dependent variable (but this probability may depend on the latent values of the independent variables).

One immediate consequence of theorem 1 is the following.

Corollary 1. *MAR is neither a necessary nor a sufficient condition for CCI.*

Proof. Suppose that $P(M_i = (0,0)|R_i^L, T_i^L = V) \neq P(M_i = (0,0)|R_i^L, T_i^L = \overline{V})$. A missingness process of this sort is consistent with MAR, but not with CCI. Thus MAR is not sufficient for CCI. Now suppose that $P(M_i = (1,0)|R_i^L = \overline{B}, T_i^L) \neq P(M_i = (1,0)|R_i^L = B, T_i^L)$. A missingness process of this sort is consistent with CCI but not with MAR. Thus MAR is not necessary for CCI. □

It is interesting to compare Sherman's MAR_S conditions to the necessary and sufficient condition for ignorability given in theorem 1. I assume that the independence of M_i for $i = 1, \ldots, n$ is guaranteed by the sampling process and so is trivially true. Conditional on this assumption, the first condition in the definitions of MAR_S(unit) and MAR_S(item) has to do with the correlation between M_{i1} and M_{i2} and

the remaining conditions relate to the marginal distributions of M_{i1} and M_{i2}.

The condition on $Corr(M_{i1}, M_{i2})$ distinguishes between "unit" and "item" types of missingness, but is irrelevant from the perspective of CCI. In the case of $\mathrm{MAR_S}$(unit), given the condition on $Corr(M_{i1}, M_{i2})$, the additional condition given by Sherman is necessary and sufficient for CCI. On the other hand, given $Corr(M_{i1}, M_{i2}) = 0$, the other two conditions that Sherman gives to define $\mathrm{MAR_S}$(item) are sufficient for CCI but not necessary. Condition (2) can be weakened to allow missingness on turnout to depend on race.

3. A test of the ignorability condition?

Nicoletti (2006) has pointed out that it is impossible to test the MAR condition using only the observed data. An application of Bayes rule to equation (MAR) shows that an equivalent characterization of MAR is given by

$$P(\mathbf{Z^L}_{mis}|\mathbf{Z^L}_{obs}, \mathbf{M}) = P(\mathbf{Z^L}_{mis}|\mathbf{Z^L}_{obs}) \Leftrightarrow \text{(MAR)}. \tag{2}$$

This characterization makes immediately clear why MAR cannot be tested: MAR is an assumption purely about the distribution of missing data. Since the missing data is not observed, this assumption cannot be tested. On the other hand, OAR may be characterized by the equation

$$P(\mathbf{Z^L}_{obs}|\mathbf{Z^L}_{mis}, \mathbf{M}) = P(\mathbf{Z^L}_{obs}|\mathbf{Z^L}_{mis}) \Leftrightarrow \text{(OAR)}. \tag{3}$$

Since OAR is an assumption about the observed data, it can be tested against the data obtained. Since MCAR is the conjunction of OAR and MAR it is possible to test a necessary condition for MCAR, but not possible to test a sufficient condition for MCAR. To summarize, MCAR may be falsified by the observed data but not verified, and MAR cannot be either falsified or verified from the observed data. Since the complete case ignorability condition given in theorem 1 makes assumptions about both the observed and the missing data, it can be falsified, but not verified, from the observed data.

The MAR assumption is essential to justify the use of virtually all of the most common modern methods for handling missing data (Little and Rubin, 2002; Schafer and Graham, 2002), and complete case analysis is perhaps the most widely used method for handling missing data. So it is unsurprising that authors would attempt to devise tests to justify maintaining the MAR assumption, or to justify the use of a complete case analysis. Sherman's paper is not the only example. Other examples include Park and Davis (1993) and Donaldson and Moinpour (2005). Unfortunately, these tests, like Sherman's, fail to achieve their goal.

Sherman claims that necessary and sufficient conditions for MCAR(item) and MAR_{S}(item) may be derived using only the joint distribution of the fully observed random variables R_i^O and T_i^O. The following subsection shows that the conditions on the joint distribution of R_i^O and T_i^O that Sherman claims are necessary and sufficient for MAR_{S}(item) (respectively MCAR(item)) are in fact necessary but not sufficient.

	Turnout		
Race	M_T	$\overline{V}$	V
M_R	p_{00}	p_{01}	p_{02}
$\overline{B}$	p_{10}	p_{11}	p_{12}
B	p_{20}	p_{21}	p_{22}

TABLE 1. Notation for joint probabilities defining the distribution of observed race and turnout outcomes

3.1. **Sherman's proposed characterizations of MCAR(item) and $\mathbf{MAR_S}$(item).** Consider the joint distribution of R_i^O and T_i^O defined by the probabilities given in Table 1. Sherman defines

$$\Theta = \frac{p_{00}(p_{11}+p_{12}+p_{21}+p_{22})}{(p_{01}+p_{02})(p_{10}+p_{20})},\ R_{12} = \frac{p_{10}(p_{21}+p_{22})}{(p_{11}+p_{12})p_{20}},\ C_{12} = \frac{p_{10}(p_{21}+p_{22})}{p_{02}(p_{11}+p_{21})}$$

and makes the following claims:

(A) MCAR(item) holds if and only if $\Theta = 1, R_{12} = 1,$ and $C_{12} = 1$.

(B) MAR(item) holds if and only if $\Theta = 1$ and $R_{12} = 1$.

Sherman's proofs that

(1) MCAR(item) $\Rightarrow \Theta = 1, R_{12} = 1,$ and $C_{12} = 1$; and that

(2) MAR(item) $\Rightarrow \Theta = 1$ and $R_{12} = 1$

are correct. However his proofs that

(i) $\Theta = 1, R_{12} = 1,$ and $C_{12} = 1 \Rightarrow$ MCAR(item), and that

(ii) $\Theta = 1$ and $R_{12} = 1 \Rightarrow$ MAR(item)

are flawed. The claims in (i) and (ii) above can be falsified through a counterexample.

For simplicity of exposition suppose for the argument below that sample proportions estimate population probabilities without error. Now consider the 3 x 3 table of data given in table 2. It is easy to verify

	Turnout		
Race	M_T	$\overline{V}$	V
M_R	20	20	20
$\overline{B}$	15	15	15
B	15	15	15

TABLE 2. Hypothetical observed data

	Turnout	
Race	$\overline{V}$	V
$\overline{B}$	55	20
B	20	55

TABLE 3. Hypothetical latent data

that this table satisfies the conditions $\Theta = 1, R_{12} = 1$ and $C_{12} = 1$. Suppose that if we had been able to observe race and turnout for the entire sample we would have observed the 2x2 table for the complete data given in table 3. Note that the numbers in table 3 are perfectly consistent with those in table 2. However, it is clear that a complete case analysis of the table of incomplete data would give erroneous results. Such an analysis would conclude that there is no association between race and voting turnout. Yet an examination of the 2 x 2 table of the hypothetical complete data makes clear that there is in fact a strong association between race and voting turnout, with blacks much more likely to turnout. The 3 x 3 table summarizing the observed data simply doesn't contain enough information to show whether or not the MCAR(item) or MAR_S(item) conditions hold.

Sherman's proofs of his claims that $\Theta = 1, R_{12} = 1,$ and $C_{12} = 1 \Rightarrow$ MCAR(item) and that $R_{12} = 1$ and $C_{12} = 1 \Rightarrow$ MCAR(item) note that the MCAR(item) and MAR_S(item) conditions place certain constraints

on the values p_{ij} for $i, j = 0, 1, 2$. For instance, he notes that both MCAR(item) and MAR_S(item) require $P(M_R, M_T) = P(M_R)P(M_T)$. Sherman shows that $\Theta = 1$ and $R_{12} = 1 \Rightarrow p_{00} = (p_{00} + p_{01} + p_{02})(p_{00} + p_{10} + p_{20})$ as required. Yet MAR_S (item) and MCAR(item) place additional constraints on the data that Sherman does not consider, and the information in the table of incomplete data is not sufficient to determine if these constraints are met or not. For instance, both MAR_S (item) and MCAR(item) require that

$$P(M_{i1} = 1|R_i^L = \text{B}, V_i^L = \text{V}) = P(M_{i1} = 1|R_i^L = \text{B}, V_i^L = \overline{\text{V}}) \tag{4}$$

Writing

$$P(M_{i1} = 1|R_i^L = \text{B}, V_i^L = \text{V}) = \frac{P(R_i^L = \text{B}, V_i^L = \text{V}|M_{i1} = 1)P(M_{i1} = 1)}{P(R_i^L = \text{B}, V_i^L = \text{V})}.$$

it should be clear that it is impossible to know from the incomplete data whether or not (4) holds since we cannot know $P(R_i^L = \text{B}, V_i^L = \text{V})$ when race is missing.

One flaw in Sherman's logic is that he appears to have confused $\mathbf{Z^O}$ and $\mathbf{Z^L}$. For instance, he describes R_{12} as the ratio of the odds that turnout is observed rather than not observed given that the race outcome is B to the odds that turnout is observed rather than not observed given that the race outcome is $\overline{B}$ (p.366). Thus, he implies that

$$R_{12} = \frac{odds(M_{i2} = 0|R_i^L = B)}{odds(M_{i2} = 0|R_i^L = \overline{B})}. \tag{5}$$

In fact,

$$R_{12} = \frac{odds(M_{i2} = 0 | R_i^O = B)}{odds(M_{i2} = 0 | R_i^O = \overline{B})}. \tag{6}$$

However, even if we could interpret R_{12} as characterized by equation (5), $R_{12} = 1$ and $\Theta = 1$ would not be a sufficient condition for MAR_S(item). For instance, there is nothing to guarantee that condition (3) in the definition of MAR_S(item) is met. Similar statements could be made about MCAR(item). Even if C_{12} could be interpreted as Sherman intends, $R_{12} = 1$, $C_{12} = 1$ and $\Theta = 1$ would not be sufficient for MCAR(item).

3.2. **Sherman's proposed characterizations of MCAR(unit) and MAR_S(unit).** Sherman claims to be able characterize of MCAR(unit) and MAR_S(unit) in the context of standard panel surveys using only observable quantities. He suggests that these characterizations can be used to test for ignorable attrition. Sherman considers the hypothetical example of married women followed over two time periods with observations at each period taken on their work status, employed (E) or unemployed (U), and their husband's income, high(H) or low (L). He assumes that all women are measured at the first period, but some drop out prior to the second period. Work status is viewed as the response variable and income as an explanatory variable. Let table 4 denote the joint distribution of income and work status for the first period for all women. Let table 5 represent the joint distribution of income and work status during the first period for only those women who did not drop out prior to the second period. Sherman claims that:

All Women		
	work status	
Income	E	U
L	π_{11}	π_{12}
H	π_{21}	π_{22}

TABLE 4. Notation for probabilities relating to all women

Women who didn't drop out		
	work status	
Income	E	U
L	p_{11}	p_{12}
H	p_{21}	p_{22}

TABLE 5. Notation for probabilities relating to women in study for both time periods

(i) MCAR(unit) holds if and only if $\pi_{ij} = p_{ij}$ for all i and j .

(ii) MAR(unit) holds if and only if $\frac{\pi_{i1}p_{i2}}{\pi_{i2}p_{i1}} = 1$ for $i = 1, 2$.

It is not obvious how to evaluate Sherman's claims regarding MAR_{S}(unit) and MCAR(unit). For one, the definition Sherman gives for MAR_{S}(unit) is ambiguous in the context of the example he describes. Condition (2) of MAR_{S}(unit) requires that the missing data indicator (M_{i1}) be identically distributed within categories of the independent variable. However, in the context of a panel survey, it is unclear if Sherman intends this definition to apply to categories measured at the first wave of the study or the second wave of the study.

One thing that is clear is that the condition $\pi_{ij} = p_{ij}$ for all i and j is not sufficient to show that attrition is ignorable (and since $\pi_{ij} = p_{ij}$ for all i and j implies $\frac{\pi_{i1}p_{i2}}{\pi_{i2}p_{i1}} = 1$ for $i = 1, 2$ this second condition is also not sufficient). The condition $\pi_{ij} = p_{ij}$ for all i and j ensures that the joint distribution of variables measured at the first

wave is the same for those who drop out prior to the second period as it is for those who do not drop out. However, this condition implies nothing about the distribution of variables measured at the second wave. Hence, any complete case analysis that involves inference about population quantities that relate to second wave variables (for instance, an analysis of time trends) could well be incorrect even when $\pi_{ij} = p_{ij}$ for all i and j

3.3. **Use of Sherman's test.** Tests such as those proposed in Sherman (2000) give applied researchers the false impression that the ignorability of the missing data mechanism is a testable proposition. Indeed, some researchers seem to have been misled by the tests given in Sherman. Boehmke (2003) implies that Sherman's test can determine whether observations are randomly missing. Barabas (2004) utilizes Sherman's tests to justify certain claims about the missingness process in the data he analyzes. Barabas concludes on the basis of Sherman's test for unit missingness that attrition is ignorable in a two wave panel survey. Barabas also analyzed "item nonresponse" within each wave of the survey and concluded on the basis of Sherman's test that there were "violations of the missing at random assumption" and that "item nonresponse was nonrandom for certain questions". Given the above discussion of Sherman's tests there is reason to doubt Barabas' claims about the missing data in his study. Attrition in this study may or may not have been ignorable. Item nonresponse may or may not have been generated by a missing at random process. The data provides no answers to these questions.

4. Summary

This paper has attempted to clarify (a) the meaning of the terms MAR and MCAR, (b) necessary and sufficient conditions for a complete case analysis when interest is in the conditional distribution of a response variable given an independent variable, and (c) the testability of the MAR and MCAR assumptions. The paper by Sherman (2000) demonstrates confusion about all of the topics just listed.

This paper notes that Sherman's definition of MAR differs from the usually accepted definition. A condition which is necessary and sufficient to justify a complete case analysis with respect to conditional distributions computed from bivariate categorical data is presented. Sherman's claim to have discovered a test of certain types of MCAR and MAR missingness is falsified. In fact, as noted in Nicoletti (2006), it is impossible to test a sufficient condition for MAR, for MCAR, or for complete case ignorability using only the observed data.

Researchers should recognize that the justification for maintaining a MAR or MCAR hypothesis must come from some prior knowledge unrelated to the data at hand, rather than from the data itself. If such assumptions cannot be justified, the probability distributions in question are only partially identified. Researchers should then consider an approach that explicitly recognizes the partial identification problem, such as the non-parametric bounds approach described in Manski (2003).

References

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

[2] Barabas, Jason (2004). How deliberation affects policy opinions. *American Political Science Review*, 98(4), 687-701.

[3] Boehmke, Frederick (2003). Using auxiliary data to estimate selection bias models, with an application to interest group use of the direct initiative process. *Political Analysis*, 11, 234-254.

[4] 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.

[5] Little, Roderick and Rubin, Donald (2002). *Statistical analysis with missing data: second edition*. Hoboken, NJ: Wiley.

[6] Manski, Charles (2003). *Partial identification of probability distributions*. New York: Springer.

[7] Nicoletti, Cheti (2006). Nonresponse in dynamic panel data models. *Journal of Econometrics*, 132, 461-489.

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

[9] Peugh, James and Enders, Craig (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74 (4), 525-556.

[10] Rubin, Donald (1976). Inference and missing Data. *Biometrika*, 63 (3), 581-592.

[11] Schafer, Joseph and Graham, John (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2), 147-177.

[12] Sherman, Robert (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.