Structural Equation Modeling: A Multidisciplinary Journal, 22: 294–307, 2015

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.936082

Routledge
Taylor & Francis Group

TEACHER'S CORNER

Diagnosing Global Case Influence on MAR Versus MNAR Model Comparisons

Sonya K. Sterba[1] and Nisha C. Gottfredson[2]
[1]*Vanderbilt University*
[2]*University of North Carolina at Chapel Hill*

When missingness is suspected to be not at random (MNAR) in longitudinal studies, researchers sometimes compare the fit of a target model that assumes missingness at random (here termed a MAR model) and a model that accommodates a hypothesized MNAR missingness mechanism (here termed a MNAR model). It is well known that such comparisons are only interpretable conditional on the validity of the chosen MNAR model's assumptions about the missingness mechanism. For that reason, researchers often perform a sensitivity analysis comparing the MAR model to not one, but several, plausible alternative MNAR models. In the social sciences, it is not widely known that such model comparisons can be particularly sensitive to case influence, such that conclusions drawn could depend on a single case. This article describes two convenient diagnostics suited for detecting case influence on MAR–MNAR model comparisons. Both diagnostics require much less computational burden than global influence diagnostics that have been used in other disciplines for MNAR sensitivity analyses. We illustrate the interpretation and implementation of these diagnostics with simulated and empirical latent growth modeling examples. It is hoped that this article increases awareness of the potential for case influence on MAR–MNAR model comparisons and how it could be detected in longitudinal social science applications.

Keywords: case influence, diagnostics, growth models, missing not at random, nonignorable missingness

Missing data are a reality of longitudinal social science research. Due to recent software advances and dissemination efforts (e.g., Enders, 2010; Graham, 2012; Little & Rubin, 2002; Schafer & Graham, 2002), missing data handling methods that require missing at random (MAR) assumptions—such as full information maximum likelihood (FIML) estimation—have widely replaced ad-hoc approaches—such as listwise deletion. Under MAR, the probability of missingness can depend on observed variables in the model. For instance, the probability that a repeated measure, *y,* is missing at time *t* can depend on the observed *y* score at time *t*–1, or on the value of an observed covariate such as treatment—but cannot depend on unobservables, such as a person's underlying rate of change, or what his or her *y* score at time *t* would have been. When MAR holds, we need only model the process that generated the *y*s (here termed the *outcome-generating mechanism*), and can ignore the process that generated the missingness (here termed the *missingness mechanism*).

However, the MAR assumption may not always be realistic. For instance, particularly in clinical trials and treatment-outcome settings, there is often a concern that dropout at time *t* might indeed be due to an individual's own rate of change (latent slope) or unobserved *y* score at time *t,* even after conditioning on observables (e.g., Carpenter, Pocock, &

Correspondence should be addressed to Sonya K. Sterba, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203. E-mail: Sonya.Sterba@Vanderbilt.edu

Color versions of one or more figures in the article can be found online at www.tandfonline.com/hsem.

Lamm, 2002; Little et al., 2012; Lu, Zhang, & Lubke, 2011; Michiels, Molenberghs, Bijnens, Vaneneugden, & Thijs, 2002; Muthén, Asparouhov, Hunter, & Leuchter, 2011). This constitutes missingness that is not at random (MNAR). Some participants may no longer feel the need to stay in the study and adhere to its treatment regime because they are rapidly getting better; other participants may drop out because they are getting worse. If missingness is MNAR, to avoid biases in parameter estimates and standard errors, we are obliged to jointly model the outcome-generating mechanism together with the missingness mechanism, as they are interdependent (Rubin, 1976). A variety of "MNAR models" have been proposed to jointly model both processes, for instance, selection models or pattern mixture models. These MNAR models rest on different assumptions (for reviews see Enders, 2011; Hedeker & Gibbons, 2006; Little, 1995; Verbeke & Molenberghs, 2000).

Here, we begin by explaining why longitudinal researchers are interested in comparing the fit of a target "MAR model"—of theoretical interest to the investigators, such as a conventional latent growth curve model—against the fit of MNAR model(s). Second, we describe why many methodologists recommend conducting sensitivity analyses (investigations of the impact of variation in data or modeling conditions) on the results of such fit comparisons. Third, we highlight the utility of one form of sensitivity analysis that is underused in psychology applications: the assessment of case influence on this model fit comparison. The purpose of this article is to contribute to the missing data literature by suggesting accessible diagnostics for global case influence on MAR–MNAR fit comparisons and by illustrating their use in empirical and simulated growth modeling applications. These diagnostics have not been used in this context and they are computationally simpler than current alternatives.

COMPARISONS OF MAR VERSUS MNAR MODELS

There is, unfortunately, no general test for the plausibility of MAR versus MNAR (e.g., Little & Rubin, 2002; Molenberghs, Beunckens, Sotto & Kenward, 2008). Rather, it is only possible to compare the fit (and parameter estimates) of a target MAR model and a particular substantively chosen MNAR model—that embodies one possible MNAR mechanism out of many. Many studies perform such a model fit comparison (e.g., Beunckens, Molenberghs, Thijs, & Verbeke, 2007; Cursio, 2012; Gottfredson, Bauer, & Baldwin, 2014; Gottfredson, Bauer, Baldwin, & Okiishi, 2014; Hedeker & Gibbons, 1997; Kenward, 1998; Lin, McCulloch, & Rosenheck, 2004; Maruotti, 2011; Molenberghs & Kenward, 2007; Molenberghs, Kenward, & Lesaffre, 1997; Power et al., 2012; Roy, 2003; Van Steen, Molenberghs, Verbeke, & Thijs, 2001; Verbeke, Lessafre, & Spiessens, 2001; Verbeke & Molenberghs, 2000; Verbeke, Molenberghs, Thijs, Lesaffre, & Kenward, 2001). If the specification of the outcome-generating mechanism is correct, and the specification of the missingness mechanism is described well by one of these models, that model might provide a better fit. For instance, Gottfredson, Bauer, and Baldwin (2014) found that the Bayesian information criterion (BIC) was able to distinguish between a MAR mechanism and a particular MNAR mechanism under a variety of conditions. Specifically, what is being compared in these applications is the fit of both models to the observed data. So, to consider the fit ranking as support for MAR versus the specified MNAR, we must be willing to assume a priori the plausibility of that MNAR model's assumptions about the missingness mechanism (for instance, the stipulated distribution of the missing data given the observed data), in addition to our usual assumptions about the outcome-generating mechanism.[1]

Because MAR–MNAR model fit comparisons require these assumptions, there are alternative perspectives regarding the interpretation of such fit comparisons. As summarized by Ibrahim, Chen, Lipsitz, and Herring (2005):

> there are two different points of view on this issue: (a) the appropriate nonignorable [here called MNAR] model can be determined empirically from the observed data using such approaches as . . . the Akaike information criterion (AIC), and (b) the data cannot decide on an appropriate nonignorable model, and hence sensitivity analyses are needed. (p. 341)

In this article we take the following middle-ground perspective. On the one hand, we consider the often-performed fit comparison between a MAR and MNAR model—conditional on the appropriateness of the MNAR model's assumptions—a useful additional piece of information for researchers to consider, alongside inspecting how particular parameter estimates may change between fitting MAR and MNAR models. On the other hand, given the strong assumptions required by the MNAR model, we agree it is beneficial to consider MAR–MNAR model comparisons as part of a sensitivity analysis (e.g., Rubin, 1977). Sensitivity analyses used in this context have been of two kinds: either "one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools (such as diagnostic measures)" (Molenberghs & Verbeke, 2001, p. 255). The first kind of sensitivity analysis—conducting multiple model

[1]For instance, Verbeke, Lessafre, et al. (2001) interpreted a fit comparison that yielded a better fit for the MNAR model than the MAR model as follows: "conditional on the validity of [the] model, there is a lot of evidence for nonrandom dropout" (p. 426).

comparisons of a MAR model with each of several MNAR models embodying different missingness mechanisms—has been recently demonstrated in a number of methodological papers (e.g., Enders, 2011; Feldman & Rabe-Hesketh, 2012; Muthén et al., 2011; Xu & Blozis, 2011). However, the second kind of sensitivity analysis (applying diagnostics for case influence on this model comparison) has received little attention in the social sciences, and is the focus of this article.

CASE INFLUENCE ON MAR VERSUS MNAR MODEL FIT COMPARISONS

Unlike in the social sciences, the biostatistics literature has placed emphasis on diagnosing whether the results of a MAR–MNAR model comparison, at the sample level, are being disproportionately *influenced* by single case(s) (e.g., Beunckens et al. 2007; Henderson, 1994; Jansen et al., 2006; Molenberghs & Verbeke, 2001; Thijs, Molenberghs, & Verbeke, 2000; Van Steen et al., 2001; Verbeke, Lessafre, et al., 2001; Verbeke, Molenberghs, & Beunckens, 2008; Zhu & Lee, 2001). For instance, it is possible that when a particular case is retained, one model fits better (e.g., a MNAR model), whereas without that single case, the alternative MAR model fits better (e.g., Kenward, 1998; Molenberghs et al., 2001; Verbeke, Molenberghs, et al., 2001). If a sample-level conclusion about model ranking can be influenced by a single case, an applied researcher would certainly want to be aware of this when interpreting results. Importantly, MAR–MNAR comparisons might be especially vulnerable to such influence (e.g., Jansen et al., 2006; Thijs et al., 2000).

In the literature on case influence in the context of MNAR modeling, both *global* influence diagnostics (typically employing iterative case deletion) and *local* influence diagnostics (introducing minor, subject-specific perturbations to parameter(s) related to dropout), have been used to assess influence on results such as model fit or parameter estimates. These two kinds of diagnostics tap somewhat different types of influence, and characteristics of each have been reviewed elsewhere (e.g., Chatterjee & Hadi, 1988). One often-noted drawback to a global influence approach is the computationally intensive nature of iterative case deletion (e.g., Beunckens et al., 2007; Molenberghs & Verbeke, 2001; Thijs et al., 2000). To address this drawback, this article suggests using computationally nonintensive global influence diagnostics (Sterba & Pek, 2012) for the MAR–MNAR model fit comparison; these approximate true iterative case deletion diagnostics. Because in social science applications, sensitivity analyses for case influence on MAR–MNAR model comparisons are currently virtually nonexistent, we hope that the user-friendly implementation of this diagnostic—obtained from widely available software for any kind of MAR and MNAR model specifications—provides an accessible entry point. (In contrast, local influence diagnostics for MAR–MNAR comparisons are not available in commercial statistics packages to our knowledge.)[2]

The remainder of this article proceeds as follows. We first describe Sterba and Pek's (2012) global influence diagnostics for model fit ranking, and provide practical information regarding their interpretation. Next, we describe some conditions that can increase the risk of case influence on MAR–MNAR comparisons. Subsequently, we illustrate the use of these diagnostics in two examples (one empirical and one simulated) that each involve comparing the fit of a target MAR latent growth curve model to a selection-type MNAR model. Syntax for computing the diagnostics for both example model comparisons is available in our online Appendix at http://www.vanderbilt.edu/peabody/sterba/appxs.htm.

Selection-type MNAR models are used here in examples for several reasons. First, they are popular in empirical applications—largely because they conveniently afford inference about the substantively interesting marginal distribution of the repeated measures,[3] unlike some types of MNAR models, such as pattern mixture models. Second, to date, methodologists have mainly studied case influence when selection-type MNAR models are used (e.g., Kenward, 1998; Molenberghs et al., 2001; Verbeke, Lesaffre, et al., 2001; Verbeke, Molenberghs, et al., 2001). Maintaining this context allows us to relate aspects of prior findings to current results using different diagnostics. Third, the strong parametric assumptions of selection-type MNAR models have been thought to render them particularly susceptible to case influence (Beunckens et al., 2007; Thijs et al., 2000; Verbeke, Lesaffre, et al., 2001; Verbeke et al., 2008). Fourth, global case influence diagnostics are designed to detect the exact kind of influence—single case influence—that is most troubling for nomothetic selection-type MNAR models, which must represent the missingness generating process for the entire sample. In contrast, global case influence diagnostics in general are not designed to detect the contribution of an entire "clump" of cases that are jointly (but perhaps not individually) influential (Atkinson & Riani, 2008). If interest specifically lies in accommodating unique features of a larger clump in a MAR–MNAR model comparison, MNAR models of the pattern mixture or latent pattern mixture variety could be employed. These models allow change parameters to differ across (possibly latent) classes with different missingness patterns. However, these models are still not immune to influential cases, and are subject to other limitations (Beunckens et al., 2007) not discussed here.

[2]Whereas source code (e.g., in GAUSS) for local influence diagnostics is available directly from authors for certain model specifications (see Verbeke, Molenberghs, et al., 2001), in general this approach might require substantial programming for other specifications.

[3]Other types of MNAR models, such as pattern mixture models, employ different factorizations of the joint distribution of the repeated measures and missingness indicators that do not give as ready access to inferences involving parameters of the marginal distribution of repeated measures.

GLOBAL CASE INFLUENCE DIAGNOSTICS FOR COMPARING MAR VERSUS MNAR MODELS

As mentioned earlier, popular selection-type MNAR models (e.g., the Diggle–Kenward [1994] model, or the Wu–Carroll [1988] model) are often compared to a counterpart MAR model that is obtained by imposing one or more constraints on the MNAR model.[4] The fit of these (more restricted) MAR and (less restricted) MNAR model pairs have been compared with likelihood-ratio difference tests (LRTs) as well as information criteria—such as the Bayesian information criterion (BIC) and Akaike's information criterion (AIC). However, regularity conditions for LRTs are not met for such MAR–MNAR model comparisons (see Hens, Aerts, Molenberghs, & Thijs, 2003; Jansen et al., 2006; Molenberghs & Kenward, 2007). Use of information criteria has been recommended instead (e.g., Dmitrienko, Molenberghs, Chanung-Stein, & Offen, 2005; Maruotti, 2011; Molenberghs & Kenward, 2007; Muthén, Asparouhov, & Hunter, 2009; Power et al., 2012; Wang & Daniels, 2011). Thus, here we employ only BIC and AIC for fit comparisons. A researcher could consider one or both of these indices. BIC is intended to select the model closest to the true generating process, whereas AIC is intended to select the most generalizable model (see Kuha, 2004, or Vrieze, 2012, for reviews). As such, these indices need not agree. Consider Model A the MAR model and Model B the MNAR model. For Models A and B, define L^A and L^B as their sample likelihoods and k^A and k^B as their number of free parameters, respectively. For MAR–MNAR comparisons considered here involving conventional selection-type MNAR models, $k^A < k^B$. Following Kuha (2004), in the sample, define the between-model difference in BIC and in AIC, respectively, as:

$$\Delta\text{BIC} = -2(\ln L^A - \ln L^B) + \ln N(k^A - k^B) \quad (1)$$

$$\Delta\text{AIC} = -2(\ln L^A - \ln L^B) + 2(k^A - k^B) \quad (2)$$

Negative ΔBIC or ΔAIC indicate support for Model A at the sample level (vice versa for Model B). Assuming FIML estimation was used to fit both models, denote the individual contributions to the likelihood for each fitted model as L_i^A and L_i^B where i indicates case and $i = 1 \ldots N$. In the longitudinal models considered later, a case (i.e., the highest level unit in the analysis) is a person. These individual likelihood contributions can be obtained as a by-product of using FIML estimation to fit each model, and can be outputted using widely available software, including M*plus,* Mx, and OpenMx (e.g., Boker et al., 2011; Muthén & Muthén, 1998–2013; Neale, Boker, Xie, & Maes, 2003).

Sterba and Pek (2012) defined approximate influence diagnostics for model ranking using ΔBIC and ΔAIC as:

$$\Delta\text{ind}_{\text{BIC}_i} = -2(\ln L_i^A - \ln L_i^B) + (k_A - k_B)\ln\,(N/(N-1)) \quad (3)$$

$$\Delta\text{ind}_{\text{AIC}_i} = -2(\ln L_i^A - \ln L_i^B) \quad (4)$$

where to compute $\Delta\text{ind}_{\text{BIC}_i}$ and $\Delta\text{ind}_{\text{AIC}_i}$ requires only a single model fitting of each Model A and B. These $\Delta\text{ind}_{\text{BIC}_i}$ and $\Delta\text{ind}_{\text{AIC}_i}$ diagnostics were shown to approximate their iterative, exact case deletion counterparts: $\Delta\text{BIC}_i = \Delta\text{BIC} - \Delta\text{BIC}_{(-i)}$ and $\Delta\text{AIC}_i = \Delta\text{AIC} - \Delta\text{AIC}_{(-i)}$. Here, ΔBIC_i and ΔAIC_i represent the exact change in the sample-level index associated with deleting a case, where the subscript $(-i)$ denotes that case i was excluded from the analysis.

Computing ΔBIC_i and ΔAIC_i for all cases requires N jackknife iterative refittings of *each* the MAR and MNAR models. Even at modern computing speeds, N refittings of selection-type MNAR models can be computationally demanding, despite normally distributed repeated outcomes (see Hogan & Laird, 1997; Vonesh, Greene, & Schluchter, 2006). For instance, in the selection-type MNAR models considered later, computational burden increases as a function of the number of random effects on which dropout depends (in the Wu–Carroll [1988] MNAR model) or as a function of the number of repeated outcomes with dropout (in the Diggle–Kenward [1994] MNAR model). This is because obtaining the joint likelihood of outcomes and dropout indicators for these MNAR models requires integrating over missing response variables in the Diggle–Kenward MNAR model (two dimensions of integration for Example 1 and four dimensions of integration when later applied to Example 2) and requires integrating over the random effects distribution in the Wu–Carroll MNAR model (two dimensions of integration for Example 2). Monte Carlo numerical integration (with sharp increases in total integration points required to maintain accuracy as the dimensions of integration increase) has been recommended for estimating these models (Muthén et al., 2011), together with multiple sets of random starting values to decrease the chance of local optima (Enders, 2011; Muthén et al., 2011). For a given number of dimensions of integration, larger N, more integration points, and additional sets of starting values all increase computation time. Other kinds of selection-type MNAR models not illustrated here can be even more computationally demanding (see examples in Muthén et al., 2011).

[4]Depending on the MNAR model, such constraints could include fixing to 0 the effects of certain predictors on the probability of missingness. These predictors could be outcome scores at the time of dropout or latent growth coefficients from the outcome-generating model (as described later, in Examples 1 and 2). Another possible constraint involves fixing to 0 the covariance between random effects from the outcome-generating and the missingness mechanisms.

Fortunately, the noniterative $\Delta\text{ind}_{\text{BIC}_i}$ and exact ΔBIC_i provide highly consistent *rank orders* for cases' influence, as do the noniterative $\Delta\text{ind}_{\text{AIC}_i}$ and exact ΔAIC_i (e.g., Kendall's Tau-b = .94–.99; Sterba & Pek, 2012; see also Sadray, Jonsson, & Karlsson, 1999).[5] Closeness of rank order is important because it means, for instance, that when the sample level ΔBIC is positive, the case with the most positive $\Delta\text{ind}_{\text{BIC}_i}$ likely has the greatest potential for influence according to the exact ΔBIC_i. Also, when the sample level ΔBIC is negative, the case with the most negative $\Delta\text{ind}_{\text{BIC}_i}$ likely has the greatest potential for influence according to the exact ΔBIC_i. The same holds for $\Delta\text{ind}_{\text{AIC}_i}$ and ΔAIC_i. Hence, as a practical strategy, in the context of MAR–MNAR model comparisons, we follow Sterba and Pek's recommendation to (a) first compute $\Delta\text{ind}_{\text{BIC}_i}$ and/or $\Delta\text{ind}_{\text{AIC}_i}$ for *all* cases, flagging any potential influential case; and then (b) confirm influence by calculating the exact ΔBIC_i and/or ΔAIC_i *only* for a flagged case. In particular, a case would be flagged as potentially influential on sample-level model fit ranking:

$$\text{for } \Delta\text{BIC if } \begin{cases} \Delta\text{ind}_{\text{BIC}_i} > \Delta\text{BIC when } \Delta\text{BIC} > 0 \\ \Delta\text{ind}_{\text{BIC}_i} < \Delta\text{BIC when } \Delta\text{BIC} < 0 \end{cases} \quad (5)$$

$$\text{for } \Delta\text{AIC if } \begin{cases} \Delta\text{ind}_{\text{AIC}_i} > \Delta\text{AIC when } \Delta\text{AIC} > 0 \\ \Delta\text{ind}_{\text{AIC}_i} < \Delta\text{AIC when } \Delta\text{AIC} < 0 \end{cases} \quad (6)$$

Under these circumstances, the presence or absence of the case might alter the sign of the ΔBIC or ΔAIC index at the sample level. Note that it is possible for a given case to be influential on the sample-level ΔBIC fit ranking but not ΔAIC fit ranking (or vice versa), due to the fact that these sample-level indices (and their corresponding case-level diagnostics) evaluate model quality differently.

Case influence could also or instead occur on the *degree of evidence* for a given model. For instance, using ΔBIC we could define degree of evidence according to Raftery's (1995) guidelines that rely on the relationship between ΔBIC and Bayes factors, where |ΔBIC | of 0–2 is weak, 2–6 is positive, 6–10 is strong, and > 10 is very strong. Then if ΔBIC was –7 but $\Delta\text{ind}_{\text{BIC}_i}$ was –5, case *i* would be potentially influential on the degree of evidence for Model A, from strong to weak. See Burnham and Anderson (2002) for suggestions on effect sizes for ΔAIC, which could similarly be used to flag case(s) that might be influential on the degree of evidence in terms of ΔAIC.

[5] $\Delta\text{ind}_{\text{BIC}_i}$ exactly equals ΔBIC_i and $\Delta\text{ind}_{\text{AIC}_i}$ exactly equals ΔAIC_i when parameters are fixed to their estimates from the full *N* analyses for Models A and B when calculating $\Delta\text{BIC}_{(-i)}$ and $\Delta\text{AIC}_{(-i)}$. One degenerate special case in which the rank orders of these indices will diverge is discussed in Sterba and Pek (2012), but it is very unlikely to be seen in practice.

CONDITIONS INCREASING THE RISK OF CASE INFLUENCE ON MAR VERSUS MNAR MODEL COMPARISONS

Conceptually, we can distinguish between general and case-specific conditions increasing the risk of case influence on MAR versus MNAR model fit comparisons. General conditions can pertain to the models and sample size under consideration. Specific conditions can pertain to characteristics of a given case in relation to the specified models.

Regarding general conditions that can increase the risk of influence, all else being equal, the risk of case influence is increased when ΔBIC or ΔAIC (whichever is being used to rank models) is closer to 0 at the sample level. Then it will be easier for a given case to show influence according to the definitions in Equations 5 and 6. For $\Delta\text{AIC} = 0$, for instance, $-2(\ln L^A - \ln L^B)$ needs to equal $-2(k^A - k^B)$, using the definitions of L^A, L^B, k^A, and k^B provided earlier. In empirical settings, $-2(\ln L^A - \ln L^B)$ will generally increase along with the number of cases, *N,* and the effect size difference between the models. For instance, in MAR–MNAR comparisons in which the models differ by a single parameter (which should be 0 when MAR is upheld), $k^A - k^B = -1$, and the effect size difference between models can be quantified simply by the size of this parameter value. When this parameter value is modest (neither large nor very small) and *N* is low, there would be a generally higher risk of case influence on ΔAIC, all else being equal. Still under low *N,* consider what can happen if, instead of being modest, this parameter value were large or very small. If this parameter value were large (large departure from MAR), implying $-2(\ln L^A - \ln L^B) > -2(-1)$, the risk of influence can generally decrease. If this parameter value were very small (tiny departure from MAR), implying $-2(\ln L^A - \ln L^B) < -2(-1)$, the risk of influence can again generally decrease.

Regarding case-specific conditions that could increase the risk of influence, all else being equal (i.e., given the chosen MAR and MNAR model specifications and the total number of cases), *heterogeneity* in the outcome-generating process and/or the missingness mechanism are contributing factors. Note that heterogeneity could be defined in a discrete or continuous distribution sense, as follows. In a discrete distribution sense, population heterogeneity could be defined as when a case's outcomes and/or missingness probabilities are literally generated from a population model distinct from the other cases (e.g., Muthén, 1989). However, a case literally generated from a different population might or might not have extreme outcome scores or missingness probabilities, as a function both of sampling variability and of how different its generating model was, as compared to that for the rest of the cases. Conversely, in a continuous distribution sense, a case whose outcomes and missingness probabilities were literally generated from the same models

as the other cases could still have an extreme pattern simply due to continuous sampling variability. Here, in line with prior MNAR literature on case influence, we focus on heterogeneity manifesting in relatively extreme patterns, from any cause.

Prior research has found that population heterogeneity exclusively in the outcome-generating mechanism is a risk factor for case influence on a MAR–MNAR model comparison. Perhaps unintuitively, even cases with complete data but an anomalous longitudinal trend can have influence on the conclusions of a MAR–MNAR model comparison. For instance, Kenward (1998; followed by Crouchley & Ganjali, 2002; Molenberghs et al., 2001; Thijs et al., 2000; Verbeke, Molenberghs, et al., 2001) were interested in predicting Time 2 milk yield from Time 1 milk yield in a sample of cows. Two ill cows had qualitatively different yield patterns from the other cows. Although these two ill cows had complete data, they were influential on the model fit comparison between a MAR model (that allowed dropout to depend only on Time 1 milk yield) versus a MNAR model (that also allowed dropout to depend on Time 2 milk yield). The absence of these two ill cows flipped the model ranking from supporting MNAR to supporting MAR.[6] Although cases with extreme or outlying observed scores are not necessarily influential on fit results, and vice versa (e.g., Pek & MacCallum, 2011), in several previous studies a case with an extreme longitudinal trend in the outcome (e.g., "an unusually high profile, or a somewhat atypical serial correlation behavior"; Jansen et al., 2006, p. 844; see also Beunckens et al., 2007; Enders, 2011; Henderson, 1994) was also found influential on a MAR–MNAR model comparison. Later, the Example 2 illustration involves a case with complete data but an unusual outcome trend, and this case influences conclusions of a MAR–MNAR comparison.

Population heterogeneity in the missingness mechanism, on the other hand, is an intuitive reason for case influence on a MAR–MNAR model comparison (e.g., Thijs et al., 2000). However, prior research has actually found population heterogeneity in the missingness mechanism to be a less likely risk factor for influence (Jansen et al., 2006), except in combination with other predisposing factors. Consider the situation in which a single case dropped out nonrandomly (e.g., its probability of missingness was generated by an MNAR process), and all other cases dropped out randomly (e.g., their probabilities of missingness were generated by, say, a MAR process). Prior research found that, in this situation, the single case is unlikely to be influential on a MAR–MNAR comparison if, for instance, it dropped out at the beginning of the study (which implies little information contributed toward discriminating between the competing models) and if its data deviated little from the sample mean trajectory (Jansen et al., 2006). However, Example 1 later illustrates that a single case that dropped out nonrandomly, but with *neither* of these characteristics, can influence a MAR–MNAR comparison.

In sum, this section reviewed some general model and data conditions as well as some case-specific conditions that might increase the risk of case influence on a MAR–MNAR model fit comparison. Although these conditions tend to increase the risk of case influence, they do not guarantee the presence of influence. In an empirical setting, we will be able to diagnose *whether* a case is influential, but we may not be able to definitively determine *why* it is influential. This determination might require obtaining information beyond the diagnostics by further substantive study, rechecking data integrity, and considering alternative model specification(s).

EXAMPLES

In the two illustrations that follow (one simulated, one empirical), we focus on comparisons of a target MAR model and a common selection-type MNAR model. In each illustration, there is an influential case displaying some of the case-specific characteristics that the previous section described as potentially increasing the risk of case influence. Although in practice it may be useful to compare each MAR model to more than one MNAR model (see Discussion), for now we limit ourselves to one illustrative MNAR model per example. A different selection-type MNAR model is applied in each example, for theoretical reasons (following Graham, 2012; Hedeker & Gibbons, 2006). M*plus* 7.0 (Muthén & Muthén, 1998–2013) was used for example analyses.

Simulated Example

Often in published individual trajectory plots from developmental studies and longitudinal treatment-outcome studies with attrition there is a single case that (a) drops out in the middle of the study, and (b) also deviates considerably from the sample mean trajectory prior to dropout (i.e., some heterogeneity in the outcome pattern). For instance, this pattern is observed in data from Neri et al. (2013), Barry et al. (2005), Pan, Rowe, Singer, and Snow (2005), Yancy et al. (2010), and Grober et al. (2008). Example 1 is a simulated illustration of the possibility for influence on a MAR–MNAR model comparison when a single case of this description has a heterogeneous missingness mechanism. Specifically, one case is generated with an MNAR missingness mechanism

[6] Previous model comparisons involving this milk yield example used an LRT. We replicated their pattern of results using the diagnostics described in this article. For instance, in the full sample $\Delta BIC = 1.508$, favoring the MNAR model. But Cows 4 and 5 were each flagged as potentially influential on model ranking using Equation 5: for Cow 4 $\Delta ind_{BIC_i} = 3.30$ and for Cow 5 $\Delta ind_{BIC_i} = 3.28$. To confirm their influence, first Cow 4 was deleted, which indeed reversed support at the sample level to the MAR model, $\Delta BIC_{(-i)} = -1.17$; additionally, deleting Cow 5 increased sample-level support for the MAR model from weak to strong ($\Delta BIC_{(-i)} = -4.63$).

whereas the rest of the cases adhere to a MAR missingness mechanism. We apply the diagnostics to assess case influence in this simulated example. This simulated illustration involves the common context of a treatment-outcome study with dropout, although the diagnostics would apply also to nontreatment studies and to MNAR models for intermittent missingness (see Discussion).

In Example 1, the outcome-generating mechanism—a Gaussian conditional latent growth model in Equation 7—was used to generate five repeated measures for 150 cases. For 149 cases, the missingness mechanism was the MAR logistic dropout submodel in Equation 8. For one case, the missingness mechanism was the MNAR Diggle–Kenward (1994) logistic dropout submodel in Equation 9.[7] In total, 18% of persons dropped out and dropout occurred at the final two time points out of five repeated measures. Figure 1 depicts a subset of 30 trajectories from this sample. Before cases dropped out, their trajectory line is solid; if cases dropped out, the dashed line depicts what their y scores would have been. Case 117, with a relatively extreme outcome pattern due to sampling variability, dropped out nonrandomly (MNAR). Substantive motivation for considering a Diggle–Kenward (1994) MNAR model is based on the expectation that the probability of dropout at a given time point might depend directly on what that person's unobserved y score would have been, after controlling for observables (e.g., treatment status; y score at a previous time point).

The *MAR model* is given in Equations 7 and 8 and the *MNAR model* is Equations 7 and 9:

$$\begin{aligned} y_{it} &= \eta_{0i} + \eta_{1i}\lambda_t + \varepsilon_{it} \\ \eta_{0i} &= \gamma_{00} + \gamma_{01} treat_i + \zeta_{0i} \\ \eta_{1i} &= \gamma_{10} + \gamma_{11} treat_i + \zeta_{1i} \end{aligned} \tag{7}$$

where $\varepsilon_{it} \sim N(0, \sigma_t^2)$ and

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{00} & \\ \psi_{10} & \psi_{11} \end{bmatrix} \right)$$

$$\text{logit}(\Pr(d_i = t | d_i \geq t; \mathbf{y}_i, treat_i)) = \alpha_{0t} + \alpha_{10} treat_i + \alpha_{20} y_{it-1} \tag{8}$$

$$\text{logit}(\Pr(d_i = t | d_i \geq t; \mathbf{y}_i, treat_i)) = \alpha_{0t} + \alpha_{10} treat_i + \alpha_{20} y_{it-1} + \alpha_{30} y_{it} \tag{9}$$

[7]Parameters for the outcome-generating mechanism were: $\gamma_{00} = 2; \gamma_{10} = -.5; \gamma_{01} = 0; \gamma_{11} = -1;$ $\sigma_1^2 - \sigma_T^2 =$ (1.4,.9,.73,.65,.6); $\tau_{00} = .8;$ $\tau_{10} = .1; \tau_{11} = .4$. Parameters for the missingness mechanism included: $\alpha_{04} = \alpha_{05} = -2.5; \alpha_{10} = .1; \alpha_{20} = .2; \alpha_{30} = 1$. Parameters were chosen to give rise to a pattern of observed and missing data with features related to Neri et al. (2013), Barry et al. (2005), Pan et al. (2005), Yancy et al. (2010), or Grober et al. (2008) in the manner described in the text.

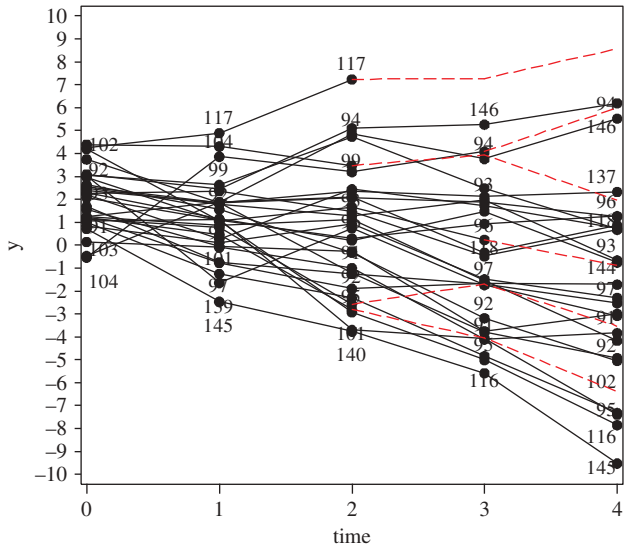

FIGURE 1 Example 1 simulated data trajectories before dropout (solid lines) and after dropout (dashed lines) for subset of 30 cases. *Notes.* Solid lines connect observed data points. Dashed lines denote unobserved data.

In the conditional latent growth model in Equation 7, y_{it} is the outcome for person i at time t, where $t = 1 \ldots 5$. η_{0i} and η_{1i} are intercept and linear growth coefficients, respectively, which are themselves predicted by treatment, $treat_i$. Note that λ_t represents fixed linear time scores of 0, 1, 2, 3, 4 for times $t = 1$–5. ε_{it} is a normally distributed time-specific residual and ζ_{0i} and ζ_{1i} are bivariate normally distributed person-specific random effects. d_i is a dropout indicator for person i. In Equations 8 and 9 a person's probability of dropout at time t, where $t = 4$ or 5, is predicted by $treat_i$ and that person's outcome score at time t–1 (i.e., y_{it-1}). In Equation 9, dropout at time t is also predicted by that person's (unobserved) outcome score at time t, consistent with MNAR.

Designating the MAR model (Equations 7 & 8) as Model A[8] and the MNAR model (Equations 7 & 9) as Model B, the comparison of Models A and B has $\Delta df = 1$. Fitting both models yields $\Delta\text{BIC} = 10.45$ and $\Delta\text{AIC} = 13.46$. Both indices indicate support for the MNAR model (of a very strong degree according to Raftery [1995], for ΔBIC). However, one case, Case 117, is flagged as potentially influential on model ranking according to $\Delta\text{ind}_{\text{BIC}_i}$ or $\Delta\text{ind}_{\text{AIC}_i}$ using the definitions in Equations 5 and 6. The $\Delta\text{ind}_{\text{BIC}_i} = 13.84$ (as depicted in the index plot in Figure 2)[9] and

[8]Technically no missingness model is needed under MAR, and as such the estimates for the outcome-generating process parameters in the MAR model should stay the same regardless of whether this logistic dropout model is included or not. Nonetheless, it is conventional to include the logistic dropout model conditional on observables in the MAR specification when it is to be compared to a more elaborated logistic dropout model, in the MNAR specification, so the likelihoods are on the same metric (see Muthén et al., 2011).

[9]$\Delta\text{ind}_{\text{AIC}_i}$ and $\Delta\text{ind}_{\text{BIC}_i}$ index plots will only differ by a constant so we only present one of them here.

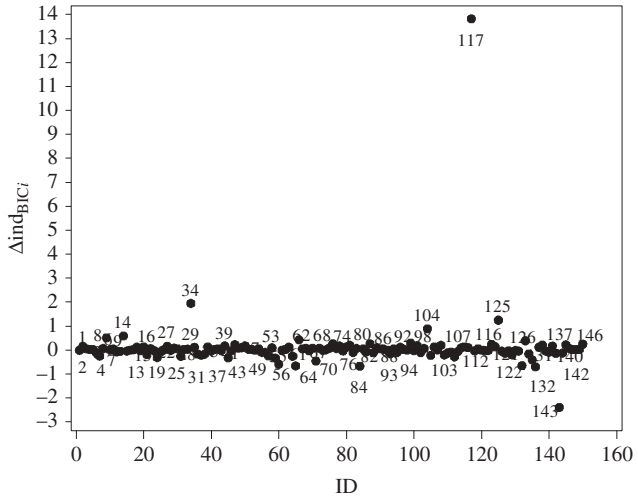

FIGURE 2 Index plot for Δind_{BIC_i} in Example 1.

$\Delta ind_{AIC_i} = 13.83$. Deleting Case 117 confirms that this case is influential on the sample-level ranking of ΔBIC, which now supports the MAR model ($\Delta BIC_{(-i)} = -1.38$), and that Case 117 is influential on degree, but not direction, of evidence for ΔAIC ($\Delta AIC_{(-i)} = 1.63$).

Example 1 illustrates the possibility that a case generated from a heterogeneous missingness mechanism (here, an MNAR mechanism), could have influence on sample-level results of a MAR–MNAR model comparison. Other characteristics that increased this case's potential for influence included when it dropped out and how the case's outcome data deviated from the sample mean trajectory (Case 117's y_{i4} and y_{i3} scores were relatively large at and before time of dropout, as shown in Figure 1). In an empirical study, we would not be able to conclude from these results that the influential Case 117 was MNAR, but we could consider heterogeneity in the missingness mechanism a potential contributing factor.

Empirical Example

Example 2 is an empirical illustration that cases with anomalous trends can dominate the MAR–MNAR model ranking result *even* if they have complete data (see also Enders, 2011; Jansen et al., 2006). This example involves $N = 170$ cases from a depression treatment trial (see Dmitrienko, Chuang-Stein, & D'Agostino, 2007). Treated and control participants were assessed five times on the outcome of interest: the Hamilton Depression Rating Scale (Hamilton, 1980), where higher scores denote more depressive symptoms. Whereas 10% of participants dropped out by Time 2, 36% dropped out by Time 5. Here we are interested in comparing the fit of a MAR latent growth curve model with an MNAR model allowing for random-coefficient dependent missingness (in particular, a Wu–Carroll [1988] random coefficient selection model). Substantive motivation for considering a Wu–Carroll (1988) MNAR model is based on the expectation that the probability of dropout at a given time point may depend on one or more latent growth coefficients (here, η_{0i} and η_{1i}) characterizing an individual's entire change process (e.g., Feldman & Rabe-Hesketh, 2012)—rather than depending directly on unobserved outcome scores at particular time point(s), as in Example 1's Diggle–Kenward (1994) MNAR model.

The rate of change was theoretically expected to be nonlinear—with depression scores on average decreasing more rapidly at earlier time points (as in other similar studies; Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009). The outcome-generating process in both our MAR and MNAR models represented the nonlinear mean trend semiparametrically using a shape-factor model (also called a freed-loading model; see Bollen & Curran, 2006). Specifically, the outcome-generating process was specified as in Equation 7, except that λ_2, λ_3, and λ_4 were freely estimated to allow the model-implied mean trend to take on a flexible functional form. The missingness mechanism under MAR, in Equation 10, allows the probability of dropout at time t, (with dropout occurring at $t = 2$–5), to depend only on treatment condition. Under MNAR the missingness mechanism is given in Equation 11, where dropout is also a function of η_{0i} and η_{1i}.

$$\text{logit}(\Pr(d_i = t | d_i \geq t; treat_i)) = \alpha_{0t} + \alpha_{10} treat_i \tag{10}$$

$$\begin{aligned}\text{logit}(\Pr(d_i = t | d_i \geq t; \boldsymbol{\eta}_i, treat_i)) = \alpha_{0t} + \alpha_{10} treat_i \\ + \alpha_{20}\eta_{0i} + \alpha_{30}\eta_{1i}\end{aligned} \tag{11}$$

Designate the *MAR model* (Equation 10 and Equation 7—with λ_2–λ_4 estimated) as Model A. Designate the *MNAR model* (Equation 11 and Equation 7—with λ_2–λ_4 estimated) as Model B. This model comparison has $\Delta df = 2$. At the sample level, information criteria support the MAR model; ΔAIC = –3.96, and ΔBIC = –10.23 (with the latter indicating very strong support for the MAR model at the sample level). Inspection of the influence diagnostics, however, indicates that two cases contribute disproportionately to this sample-level model ranking, and that these two cases to some extent counterbalance each other's contributions. These two cases are here labeled Case 11 and Case 145 in the Δind_{AIC_i} index plot in Figure 3. Case 11 has $\Delta ind_{AIC_i} = -7.53$ and $\Delta ind_{BIC_i} = -7.54$; it strongly favors the MAR model, has complete data, and is in the control group. Case 145 has $\Delta ind_{AIC_i} = 5.22$ and $\Delta ind_{BIC_i} = 5.21$; it heavily favors the MNAR model, dropped out (after the third time point), and is in the treatment group. Whereas both cases are potentially influential on degree of evidence, most seriously Case 11 is flagged as potentially influential on ΔAIC model ranking according to the definition in Equation 6. To conserve space, we focus further pedagogical investigation on Case 11.

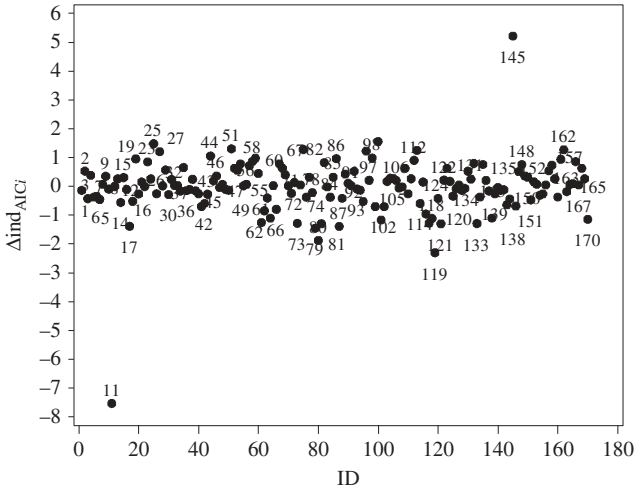

FIGURE 3 Index plot for $\Delta\text{ind}_{\text{AIC}_i}$ in Example 2.

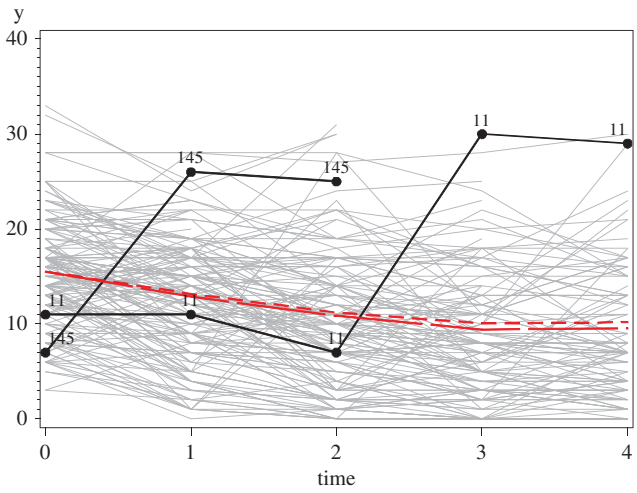

FIGURE 4 Observed data trajectories for Example 2 with Cases 11 and 145 shown in bold and the model-implied means for the missing not at random (MNAR) model (short dash) and missing at random (MAR) model (long dash) superimposed. *Note*. Model-implied means were obtained from the full *N* analysis.

Potential influence for the flagged Case 11 is confirmed with exact case deletion. When Case 11 is deleted, the model ranking indeed reverses for ΔAIC; that is, $\Delta\text{AIC}_{(-i)} = 3.08$ (now preferring MNAR at sample level), whereas for ΔBIC the degree of evidence for the MAR model is reduced; $\Delta\text{BIC}_{(-i)} = -3.19$. In other words, Case 11 has influence on the sign of ΔAIC and the magnitude of ΔBIC, favoring MNAR.

Although Cases 11 and 145 differ on a variety of features (dropout vs. completer; treatment vs. control), they both have relatively anomalous functional forms, which may relate to their potential for influence on the model comparison result. Figure 4 depicts their *observed* trajectories (bold solid lines), alongside other cases' *observed* trajectories (nonbold solid lines), and the overall *model-implied mean* trajectories for both models (dashed lines). Additionally, *model-implied, individual-specific* trajectories for Case 11 are, for the MAR model, $\hat{y}_i|\boldsymbol{\eta}_i, treat_i = [12.04, 17.59, 21.80, 24.81, 24.59]$; and, for the MNAR model, $\hat{y}_i|\boldsymbol{\eta}_i, treat_i = [18.05, 19.71, 21.11, 21.93, 21.83]$. Comparison of these model-implied, individual-specific trajectories against Case 11's (bold, solid) observed trajectory in Figure 4 might suggest why Case 11 greatly prefers the MAR model according to $\Delta\text{ind}_{\text{AIC}_i}$ and $\Delta\text{ind}_{\text{BIC}_i}$. That is, even though both models accommodate individual differences in level and shape via two random effects, Case 11's observed trajectory in Figure 4 is closer to $\hat{y}_i|\boldsymbol{\eta}_i, treat_i$ from the MAR model.

In sum, Example 2 illustrates how one or a few cases can disproportionately affect sample-level conclusions about the plausibility of the specified MAR versus MNAR missingness processes. Cases, such as Case 11, with unusual trends might be able to dominate the MAR–MNAR model ranking result despite having complete data (also see Enders, 2011; Jansen et al., 2006).

In practice, a researcher would not need to permanently omit an influential case (e.g., Case 11) from the analysis, particularly without additional substantive rationale from follow-up investigation (e.g., indication of an illness, adverse event, or data entry or coding error). Instead, after deciding which fit index or indices to interpret (here, AIC), here it could simply be reported that a MAR–MNAR model comparison was performed, the result of which was sensitive to the influence of a particular case. Parameter estimates from the pair of fitted models—with and without the influential case—could be inspected, as shown in Table 1. The researcher might start by interpreting the best fitting model in the full *N* analysis (here, the MAR model; Table 1, column 2), while noting the sensitivity of the results. Then it could be noted whether different substantive conclusions would be drawn about any particular parameter estimates based on the best fitting $(N - 1)$ model, omitting the influential case (here, the MNAR model in Table 1, column 3). For instance, comparing Table 1 columns 2 versus 3 (both in bold), there is the same pattern of significant and nonsignificant parameter estimates, and furthermore significant estimates are of the same sign. However, particular parameter estimates do differ in magnitude (e.g., the mean slope, γ_{10}, is 21% smaller in absolute value in column 3 than in column 2). In conclusion, although inferences about the missingness mechanism are sensitive to the influence of particular case(s) in this application, the overall substantive conclusions for the outcome-generating process are quite robust. That is, in both columns, there is a similar nonlinearly decreasing mean trend, a similar magnitude of individual variation in level and change, and a nonsignificant effect of treatment on intercepts and slopes.

TABLE 1
Parameter Estimates for the Full *N* and the *N* – 1 Model Comparisons in Example 2

	Full N Analysis				*N – 1 Analysis (Omitting Case 11)*			
	MNAR Model		***MAR Model***		***MNAR Model***		*MAR Model*	
Parameter	*Estimate*	*SE*	***Estimate***	***SE***	***Estimate***	***SE***	*Estimate*	*SE*
γ_{00}	15.364*	.609	**15.286***	**.629**	**15.128***	**.610**	15.372*	.631
γ_{10}	−1.157*	.204	**−1.295***	**.201**	**−1.018***	**.204**	−1.355*	.193
γ_{01}	.328	.853	**.456**	**.878**	**.383**	**.846**	.377	.879
γ_{11}	−.352	.257	**−.399**	**.267**	**−.338**	**.243**	−.306	.250
λ_2	1.765*	.241	**1.768***	**.216**	**1.807***	**.258**	1.798*	.221
λ_3	3.240*	.294	**3.112***	**.262**	**3.447***	**.323**	3.243*	.269
λ_4	4.107*	.290	**4.071***	**.246**	**4.387***	**.330**	4.141*	.256
ψ_{00}	19.898*	3.198	**26.823***	**4.300**	**20.811***	**3.201**	25.722*	4.194
ψ_{11}	1.248*	.322	**1.856***	**.407**	**1.280***	**.297**	1.529*	.364
ψ_{10}	.569	.692	**−1.618**	**1.077**	**.429**	**.724**	−1.032	.991
σ_1^2	14.275*	2.549	**6.573***	**3.172**	**12.204***	**2.533**	7.836*	3.100
σ_2^2	15.419*	2.239	**15.686***	**2.235**	**15.059***	**2.165**	15.384*	2.177
σ_3^2	19.203*	2.860	**19.866***	**2.907**	**17.406***	**2.530**	18.158*	2.685
σ_4^2	6.648*	2.206	**4.772***	**2.014**	**4.140***	**1.775**	3.550*	1.835
σ_5^2	17.147*	2.902	**16.949***	**2.903**	**18.278***	**3.003**	17.703*	2.938
α_{02}	2.417*	.813	**2.198***	**.290**	**2.686***	**.771**	2.183*	.290
α_{03}	2.623*	.844	**2.465***	**.331**	**2.880***	**.802**	2.449*	.331
α_{04}	2.285*	.855	**2.206***	**.313**	**2.535***	**.808**	2.190*	.313
α_{05}	1.734*	.877	**1.802***	**.290**	**1.962***	**.831**	1.784*	.290
α_{10}	.314	.332	**.001**	**.271**	**.297**	**.331**	−0.014	.271
α_{20}	.047	.042	—	—	**.059**	**.042**	—	—
α_{30}	.706*	.285	—	—	**.724***	**.278**	—	—

Note. MNAR = missing not at random; MAR = missing at random. Parameter estimates shown in bold correspond to the best fitting model for the full *N* and the best fitting model for the *N* – 1 sample, according to Akaike's Information Criterion.

[a]In Example 2, the MAR model is Equation 10 and Equation 7—with $\lambda_2 - \lambda_4$ estimated, and the (Wu–Carroll) MNAR model is Equation 11 and Equation 7—with $\lambda_2 - \lambda_4$ estimated.

*$p < .05$.

DISCUSSION

Social scientists might at times have substantive reasons to expect that their missing data mechanism does not adhere to MAR assumptions. In this situation, researchers often continue to use missing data handling methods that assume MAR, hoping that after conditioning on diverse observed variables potentially related to missingness and to outcome(s), bias is mostly ameliorated (Schafer & Graham, 2002). In this situation, as another option, an MNAR model can be applied to jointly model interdependent outcome-generating and missingness mechanisms. Such MNAR models have the potential to prevent bias in parameters of the outcome-generating process (of key interest to investigators) that could otherwise arise due to MNAR missingness. As awareness of MNAR models has increased, more empirical applications conduct fit comparisons between a target MAR model and substantively reasonable MNAR model. Because the results of such fit comparisons are valid under the assumptions of the stipulated MNAR mechanism (e.g., Verbeke, Lesaffre, et al., 2001), methodologists have encouraged two kinds of sensitivity analyses to examine the robustness of these results. One kind of sensitivity analysis—recently demonstrated in several social science applications (e.g., Enders, 2011; Feldman & Rabe-Hesketh, 2012; Muthén et al., 2011; Xu & Blozis, 2011)—involves examining parameter estimates and fit across not one but several competing MNAR models, each imposing alternative assumptions. A second kind of sensitivity analysis—presently underused in social science applications—involves examining the robustness of a given MAR–MNAR fit comparison to influential cases(s). As explained by Thijs et al. (2000), "influential subjects can have large impact on the substantive conclusions, especially in the context of selection models for incomplete data, due to the well-known sensitivity to model assumptions, and therefore formal tools for their detection are to be welcomed" (p. 644).

In response, this article suggested two convenient diagnostic tools for detection of global case influence on MAR–MNAR model comparisons; neither has been used before in this context. Global case influence diagnostics previously used in this context were not convenient (i.e., required *N* iterative refittings per model; e.g., Kenward, 1998; Molenberghs et al., 2001; Thijs et al., 2000). This is particularly burdensome for selection-type MNAR models, as they can require high-dimensional numerical integration. Here we

suggested that two noniterative approximate global influence diagnostics, $\Delta\text{ind}_{\text{AIC}_i}$ and $\Delta\text{ind}_{\text{BIC}_i}$, be used to *flag* case(s) that are potentially influential on MAR–MNAR model fit rankings according to ΔAIC or ΔBIC. Following the rationale in Sterba and Pek (2012), we suggested—for flagged cases only—confirming the presence or absence of influence with exact case deletion. This strategy drastically reduces potential refittings per model to perhaps one or two needed to confirm whether conclusions are susceptible to case influence. Moreover, the $\Delta\text{ind}_{\text{AIC}_i}$ and $\Delta\text{ind}_{\text{BIC}_i}$ diagnostics we suggested are generally applicable and user-friendly; their calculation requires by-products of FIML estimation that several commercial software programs already output, as demonstrated for one software program (M*plus*) in the online Appendix.

Examples 1 and 2 were chosen for pedagogical purposes to illustrate interpretation of the diagnostics in the context of influential case(s). In practice, however, it is not uncommon to find no influential cases. If no cases are flagged as influential, researchers can report that the sample-level conclusions about model ranking are robust to the contribution of an individual case, leading to additional confidence in these results. On the other hand, if at least one case is flagged (and confirmed) as influential on a MAR–MNAR model comparison, such sensitivity needs to be reported. Particularly, it would be useful to report to what extent the final model ranking and the parameter estimates of substantive interest would change if that case were excluded or included. An example of how such sensitivity analyses can be reported was shown in the empirical illustration.

The fundamental importance of a case influence analysis for a MAR–MNAR model comparison is to ensure that a researcher is informed if his or her model comparison result can be unduly influenced by a particular case. Earlier we discussed several data and model conditions and case-specific characteristics that could increase the risk of case influence, all else being equal. Two of these case-specific characteristics were highlighted in the examples. In practice, several contributing risk factors could apply in a particular application (e.g., Verbeke et al., 2008).

Multiple Model Comparisons

In illustrative examples presented here, one target MAR model was compared to a particular substantively motivated MNAR model, and case influence was investigated in a sensitivity analysis. As mentioned earlier, another kind of sensitivity analysis involves comparing a target MAR model to more than one MNAR model, or to more than one specification of a given MNAR model (e.g., Hedeker & Gibbons, 2006; Michiels et al., 2002; Verbeke, Lesaffre, et al., 2001). Combining these two kinds of sensitivity analyses, case influence could be investigated for some or all pairs of models under comparison. For instance, we might be interested in an additional MAR–MNAR comparison for empirical Example 2—say, a comparison of a Diggle–Kenward MNAR model to a MAR model[10] where dropout is conditioned on y_{it-1} (Equations 8 and 7 with λ_2–λ_4 estimated). In this comparison, ΔAIC and ΔBIC support the MNAR Diggle–Kenward model (ΔAIC = 5.79, ΔBIC = 2.65) at the sample level. Case 11 again has the most extreme $\Delta\text{ind}_{\text{BIC}_i}$ and $\Delta\text{ind}_{\text{AIC}_i}$, and can be flagged (and confirmed) as influential on ranking for ΔBIC, but not ΔAIC. (Hence, without Case 11, the same sample-level ranking pattern is obtained as in Example 2: AIC supports the MNAR model and BIC weakly supports the MAR model.) Several factors described earlier are operating here again to increase the risk of influence: a case-specific condition (Case 11's anomalous observed trend) as well as general data and model conditions (the overall modest N and relative closeness of ΔBIC and ΔAIC to 0 at the sample level).

When conducting multiple model comparisons, a researcher needs to decide whether to retain or omit an influential case identified in one model comparison when conducting other comparisons. This decision can be informed by the researcher's outside investigation into potential explanations for that case's influence. If the case was determined to be a data coding error, for example, removing it from other model comparisons seems reasonable. Otherwise, if a known cause for the case's influence was not determined, a researcher may consider retaining the case in other comparisons for the purposes of examining and reporting when it, or a different case, might be influential. If a different case is influential in a subsequent comparison, this could highlight which aspects of a model specification are needed to accommodate a particular case's data characteristics, or suggest ideas for new specifications.

Extensions and Limitations

Several extensions and limitations deserve mention. First, global case influence diagnostics (including those discussed here) traditionally are not designed to detect "clumps" of cases that are jointly influential; a clump might mask detection of influence for cases in the clump (Lawrance, 1995). Some global influence diagnostics have been modified to detect joint influence among k cases, for a prespecified number $k > 1$ (Bruce & Martin, 1989; Thijs et al., 2000). Alternatively, it would be possible to repeatedly apply the diagnostics that we suggested here for several cases in a suspected clump. We could then recheck for influential cases with Equations 3 or 4 after deleting one flagged case at a time (see Sterba & Pek, 2012). See Footnote 6 for an illustration.

[10]Here we are not interested in comparing the MAR model from Example 2 to a Diggle-Kenward MNAR model because these models differ in parameters pertaining to not only MNAR (the effect of y_{it} on dropout) but also MAR (the effect of y_{it-1} on dropout). If this comparison were of interest, note that ΔAIC and ΔBIC support the MNAR model (ΔAIC = 22.29 and ΔBIC = 16.02, where $\Delta df = 2$) and no influential cases are flagged.

When at least one case is influential, a lack of robustness of the sample-level result to single-case contributions has been demonstrated. In theory, (latent) pattern mixture MNAR models could instead be employed to accommodate larger clumps possibly thought to arise from unobserved population heterogeneity in the missingness mechanism (e.g., Beunckens, Molenberghs, Verbeke, & Mallinckrodt, 2014; Dantan, Proust-Lima, Letenneur, & Jacqmin-Gadda, 2008). A practical complication is that classes extracted by latent pattern mixture models might serve to approximate data features other than heterogeneity in the missingness mechanism (Gottfredson, Bauer, & Baldwin, 2014).

Second, selection-type MNAR models that were illustrated here in examples only considered one particular kind of missingness—dropout. Certain MNAR models, including shared-parameter models, readily accommodate intermittent missingness (e.g., Follmann & Wu, 1995; Lin et al., 2004; Little, 1995). Case influence diagnostics described here would be equally applicable in that setting.

Third, although in this article we have focused on MAR–MNAR model comparisons, a researcher might also have interest in comparing fit and parameter estimates across different MNAR models. Researchers must keep in mind that comparing fit with ΔAIC or ΔBIC (and diagnosing influence using $\Delta\text{ind}_{\text{BIC}_i}$ and $\Delta\text{ind}_{\text{AIC}_i}$) requires that the same set of dependent variables be used across the models under comparison (e.g., the repeated measure outcomes and the dropout indicators; Burnham & Anderson, 2002). For instance, the Wu–Carroll MNAR dropout model and Diggle–Kenward MNAR dropout model meet the requirement of maintaining the same set of dependent variables with each other, but neither may meet the requirement with other MNAR models allowing for intermittent missingness.

Conclusions

Commonly used estimation routines (e.g., FIML) accommodate MAR missingness, but not MNAR missingness. Incorrect assumptions about the missingness mechanism have the potential to induce bias in parameter estimates of inferential interest. Comparisons of MAR and MNAR model(s) have become more common, particularly in treatment-outcome studies where MNAR missingness might be suspected on substantive grounds. The fact that the results of these fit comparisons can be sensitive to case influence is underappreciated in social science research. This article illustrated the calculation, interpretation, and implementation of two convenient diagnostics suited for detecting case influence on MAR–MNAR model comparisons. We hope that this article increases awareness of the potential for case influence on MAR–MNAR model comparison results and increases understanding of how to detect it.

REFERENCES

Atkinson, A., & Riani, M. (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society*, *38*, 3–14.

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, *77*, 203–211. doi:10.1037/a0015235

Barry, S., Zeger, S., Selnes, O., Gega, M., Borowicz, L., & McKahnn, G. (2005). Quantitative methods for tracking cognitive change 3 years after coronary artery bypass surgery. *The Annals of Thoractic Surgery*, *79*, 1104–1109.

Beunckens, C., Molenberghs, G., Thijs, H., & Verbeke, G. (2007). Incomplete hierarchical data. *Statistical Methods in Medical Research*, *16*, 457–492. doi:10.1177/0962280206075310

Beunckens, C., Molenberghs, G., Verbeke, G., & Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics*, *64*, 96–105.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., & Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*, 306–317.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. Hoboken, NJ: Wiley.

Bruce, A. G., & Martin, R. D. (1989). Leave-*k*-out diagnostics for time series. *Journal of the Royal Statistical Society, Series B*, *51*, 363–424.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.

Carpenter, J., Pocock, S., & Lamm, C. (2002). Coping with missing values in clinical trials: A model based approach applied to asthma trials. *Statistics in Medicine*, *21*, 1043–1066.

Chatterjee, S., & Hadi, A. (1988). *Sensitivity analysis in linear regression*. New York, NY: Wiley.

Crouchley, R., & Ganjali, M. (2002). The common structure of several recent statistical models for dropout in repeated continuous responses. *Statistical Modeling*, *2*, 39–62.

Cursio, J. F. (2012). *Latent trait pattern-mixture mixed-models for ecological momentary assessment data* (Unpublished doctoral dissertation) University of Illinois, Chicago, IL.

Dantan, E., Proust-Lima, C., Letenneur, L., & Jacqmin-Gadda, H. (2008). Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts. *The International Journal of Biostatistics*, *4*, 1–26.

Dmitrienko, A., Chuang-Stein, C., & D'Agostino, R. (2007). *Pharmaceutical statistics using SAS*. Cary, NC: SAS Institute.

Dmitrienko, A., Molenberghs, G., Chanung-Stein, C., & Offen, W. (2005). *Analysis of clinical trials using SAS*. Cary, NC: SAS Institute.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data-analysis. *Applied Statistics, C*, *43*, 49–93.

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

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16. doi:10.1037/a0022640

Feldman, B., & Rabe-Hesketh, S. (2012). Modeling achievement trajectories when attrition is informative. *Journal of Educational and Behavioral Statistics*, *37*, 703–736.

Follmann, D., & Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, *51*, 151–168.

Gottfredson, N. C., Bauer, D. J., & Baldwin, S. (2014). Modeling change in the presence of non-randomly missing data: Evaluating a shared parameter mixture model. *Structural Equation Modeling*, *21*, 196–209.

Gottfredson, N. C., Bauer, D. J., Baldwin, S., & Okiishi, J. (2014). Using a shared parameter mixture model to estimate change during treatment when termination is related to recovery speed. *Journal of Consulting and Clinical Psychology*, *82*, 813–827. doi:10.1037/a0034831

Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.

Grober, E., Hall, C., Lipton, R., Zonderman, A., Resnick, S., & Kawas, C. (2008). Memory impairment, executive dysfunction, and intellectual decline in preclinical Alzheimer's disease. *Journal of the International Neuropsychological Society*, *14*, 266–278.

Hamilton, M. (1980). Rating depressive patients. *Journal of Clinical Psychiatry*, *41*, 21–24.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Henderson, R. (1994). Discussion to P. J. Diggle & M. G. Kenward, "Informative dropout in longitudinal data analysis." *Applied Statistics*, *43*, 77.

Hens, N., Aerts, M., Molenberghs, G., & Thijs, H. (2003). The behavior of the likelihood ratio test for testing missingness. In G. Verbeke, G. Molenberghs, M. Aerts, & S. Fieuws (Eds.), *Proceedings of the 18th International Workshop on Statistical Modelling* (pp. 183–187). Leuven: Katholieke Universiteit Leuven.

Hogan, J., & Laird, N. (1997). Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine*, *16*, 259–272.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S., & Herring, A. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., & Kenward, M. G. (2006). The nature of sensitivity in missing not at random models. *Computational Statistics & Data Analysis*, *50*, 830–858.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, *17*, 2723–2732.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33*, 188–229.

Lawrance, A. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society*, *57*, 181–189.

Lin, H., McCulloch, C., & Rosenheck, R. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, *60*, 295–305.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112–1121.

Little, R. J. A., D'Agostino, R., Cohen, M., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., Hogan, J., Molenberghs, G., Murphy, S., Neaton, J., Rotnitzky, A., Scharfstein, D., Shih, W., Siegel, J., & Stern, H. (2012). The prevention and treatment of missing data in clinical trials: Special report. *The New England Journal of Medicine*, *367*, 1355–1360.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Lu, Z. L., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research*, *46*, 567–597.

Maruotti, A. (2011). A two-part mixed-effects pattern-mixture model to handle zero-inflation and incompleteness in a longitudinal setting. *Biometrical Journal*, *53*, 716–734.

Michiels, B., Molenberghs, G., Bijnens, L., Vaneneugden, T., & Thijs, H. (2002). Selection models and pattern mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*, *21*, 1023–1041.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*, 371–388.

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. Chichester, UK: Wiley.

Molenberghs, G., Kenward, M. G., & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, *84*, 33–44.

Molenberghs, G., & Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, *1*, 235–269.

Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., & Kenward, M. G. (2001). Influence analysis to assess sensitivity of the dropout process. *Computational Statistics & Data Analysis*, *37*, 93–113.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B., Asparouhov, T., & Hunter, A. M. (2009). *Nonignorable missing data and growth mixture modeling using M*plus. Presentation to the UK Mplus Users Group, Cambridge, England.

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, *16*, 17–33. doi:10.1037/a0022634

Muthén, B., & Muthén, L. K. (1998–2013). *M*plus (version 7) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). Mx: Statistical modeling (6th ed) [Computer software]. Richmond, VA: Psychiatry Department, Virginia Commonwealth University.

Neri, D., Somarriba, G., Schaefer, N., Chaparro, A., Scott, G., Mitnik, G., Ludwig, D., & Miller, T. (2013). Growth and body composition of uninfected children exposed to human immunodeficiency virus: Comparison with a contemporary cohort and United States national standards. *Journal of Pediatrics*, *163*, 249–254.

Pan, B., Rowe, M., Singer, J. & Snow, C. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families, *Child Development*, *76*, 763–782.

Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research*, *46*, 202–228.

Power, R. A., Muthén, B., Henigsberg, N., Mors, O., Placentino, A., Mendlewicz, J., Maier, W., McGuffin, P., Lewis, C., & Uher, R. (2012). Non-random dropout and the relative efficacy of escitalopran and nortriptyline in treating major depressive disorders. *Journal of Psychiatric Research*, *46*, 1333–1338.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, *59*, 829–836.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*, 538–543.

Sadray, S., Jonsson, E., & Karlsson, M. (1999). Likelihood-based diagnostics for influential individuals in non-linear mixed effects model selection. *Pharmaceutical Research*, *16*, 1260–1265.

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

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods*, *17*, 582–599.

Thijs, H., Molenberghs, G., & Verbeke, G. (2000). The milk protein trial: Influence analysis of the dropout process. *Biometrical Journal*, *42*, 617–646.

Van Steen, K., Molenberghs, G., Verbeke, G., & Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling*, *2*, 125–142.

Verbeke, G., Lessafre, E., & Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, *35*, 419–434.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Verbeke, G., Molenberghs, G., & Beunckens, C. (2008). Formal and informal model selection with incomplete data. *Statistical Science*, *23*, 201–218.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., & Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics*, *57*, 7–14.

Vonesh, E., Greene, T., & Schluchter, M. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, *25*, 143–163.

Vrieze, S. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criteron (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243.

Wang, C., & Daniels, M. (2011). A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in pattern mixture models with and without covariates for incomplete data. *Biometrics*, *67*, 810–818.

Wu, M., & Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188.

Xu, S., & Blozis, S. (2011). Sensitivity analyses of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, *36*, 237–256.

Yancy, W., Westman, E., McDuffie, J., Grambow, S., Jeffeys, A., Bolton, J., Chalecki, A., & Oddone, E. (2010). A randomized trial of a low carbohydrate diet vs. orlistat plus a low fat diet for weight loss. *Archives of Internal Medicine*, *170*, 136–145.

Zhu, H.-T., & Lee, S.-Y (2001), Local Influence for Incomplete-Data Models. *Journal of the Royal Statistical Society: Series B*, *63*, 111–126.