Developing a Systematic Approach to Identifying Missing Data Patterns and Mechanisms in Multivariate Psychological Research

© Ting Dai & Adam Davey

2020-03-26

# Author Note

Ting Dai, Department of Educational Psychology, University of Illinois at Chicago; Adam Davey, Department of Behavioral Health and Nutrition, University of Delaware. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA194178 (Davey, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Dai was supported by grants from the Department of Education (R305C080009) and National Science Foundation (0814901, 0815245, and DRL-1252436). We thank Joseph Glutting, Scott Maitland, Tony Perez, and Jyoti Savla for helpful comments on an earlier draft of this manuscript.

# Abstract

Methods for addressing missing data have become much more accessible to applied researchers. However, little guidance exists to help researchers systematically identify plausible missing data mechanisms in order to ensure that these methods are appropriately applied. Two considerations motivate the present study. First, psychological research is typically characterized by a large number of dependent variables that may be observed across multiple waves of data collection. This situation makes it more challenging to identify plausible missing data mechanisms than is the case in other fields such as biostatistics where a small number of dependent variables is typically of primary interest and the main predictor of interest is statistically independent of other predictors. Second, there is growing recognition of the importance of systematic approaches to sensitivity analyses for treatment of missing data in psychological science. We present a systematic approach to identifying a subset of missing data patterns that characterizes the larger number of observed patterns and demonstrate how these can be used to explore potential missing data mechanisms. Results are evaluated using simulated data in order to make suggestions for which approaches are likely to be most accurate as a function of the number of variables with missing data and the number of observations. Three applications of this approach to data examples where the missing data mechanism is unknown suggests that the method appears useful in practice. Finally, we offer some suggestions for ways in which this approach may be used alongside other sensitivity analyses to evaluate the potential effects of missing data on parameter estimation.

# Introduction

All missing data introduce uncertainty about the unobserved values and, in turn, the underlying population parameters of interest. Methods for the analysis of incomplete data have proliferated in recent years (Davey & Savla, 2010; Enders, 2011a, 2011b; Enders et al., 2014, 2016; Graham, 2012; Lüdtke et al., 2017; Shin et al., 2017; Yang & Maxwell, 2014). Most statistical software (e.g., Mplus, R, Stata) now includes at least some capabilities for analysis of data under ignorable and/or nonignorable conditions. A fuller understanding of the potential reasons underlying missing data is essential for appropriately selecting and applying these methods. Several general modeling strategies have been proposed with missing data. In the case of data that are believed to be missing at random (MAR), for example, inclusion of auxiliary variables (i.e., variables that are believed to be associated with the probability of nonresponse regardless of whether they are also associated with the missing values themselves, cf. Graham, 2003) is important to correctly specify the model and also to achieve more accurate statistical inferences. In the case of data that are missing not at random (MNAR), it is necessary to explicitly model the probability of nonresponse. Approaches such as pattern mixture models (Enders, 2011a, 2011b; Hedeker & Gibbons, 1997; Little & Rubin, 2002) are important to consider discrete patterns of missing data which may be associated with model parameters.

Outside of psychology, there is also an emerging emphasis on the role of sensitivity analyses in the analysis of missing data (Jo et al., 2010; Little et al., 2012; Noble et al., 2012; Xu & Blozis, 2011). For example, it is important for researchers to understand how the results obtained appear to depend on the implicit or explicit assumptions being made about the unobserved values.

The current literature, however, offers very little guidance into the systematic identification and evaluation of potential missing data mechanisms. Biostatistical work on analysis of missing data has tended to focus on data such as those from randomized clinical trials (e.g., Little et al., 2012). Studies of this kind tend to be characterized by known assignment to groups, a small number of longitudinal measurements, and typically a small number of primary outcome variables, along with a suite of model covariates. In contrast, data from psychological studies, particularly when longitudinal, tend to be characterized by a larger array of measures and occasions, with no single outcome typically taking particular precedence. Given that a data set with k discrete variables can potentially have up to 2k – 1 different meaningful patterns of missing data , a systematic approach for data reduction would seem to be one fruitful strategy as a starting point, and this is the focus of the present paper.

Researchers looking for guidance in identifying patterns and predictors of nonresponse will typically find few examples. One exception to this is Davey, Shanahan, and Schafer (2001), which first identified discrete patterns of missing data and then examined variables associated with the probability of nonresponse in order to adjust models for dropout. However, the method in that paper was not specified in sufficient detail to provide a broadly useful and replicable framework for analysis of incomplete data, nor was the performance of alternative approaches considered and evaluated in the context of simulated data where the underlying true structure was known.

## Data Reduction

Several methods for data reduction are familiar to psychologists, including principal components analysis and factor analysis. Principal components analysis (PCA) seeks to reduce the dimensionality of a data set containing a large number of correlated variables in a way that retains much of the variation present in the full data set. This is accomplished by transforming to a new set of uncorrelated variables or principal components. Principal components are extracted such that the first few retain most of the variation present in the original variables. In other words, weights are selected to form linear combinations of the original variables in order to construct new orthogonal variables which capture a maximal variance in the original variables for the reduced number of composites. Thus, in PCA the emphasis is on capturing variance rather than covariance or explaining relations between constructs.

In contrast, factor analysis (FA) hypothesizes an underlying set of “common factors” that can explain, or account for, the associations among observed variables. In other words, factor analysis attempts to achieve a reduction in dimensions by invoking a model relating observed variables to a smaller number of hypothetical or latent variables. In PCA, the goal is to find optimal ways of creating subsets of variables without imposing a model or structure on the data. In contrast, the goal of FA is to identify the underlying structure with the goal of capturing covariance rather than variance of the variables.

In this paper, we emphasize a PCA approach, because the approach is deliberately empirical rather than theoretical and delivers a mathematically unique solution. However, the methods we apply can also be used with factor analysis with suitable modifications when researchers believe that is the most appropriate method (i.e., when there is interest in identifying and understanding underlying sources of missing data rather than simply ensuring that they are appropriately modeled in an analysis). While it is possible that a researcher may know (or suspect) in advance a small number of common causes for missing data in their study, more typically this is something to be determined from the data themselves. For example, what is the smallest number of missing data patterns that can account for a given proportion of variance in missing data patterns?

## Determining the Number of Components

There is a vast literature on criteria for selecting an appropriate number of principal components (cf., Jolliffe, 2002). One of the most widely applied is Kaiser’s criterion (Kaiser, 1960), which retains all components with eigenvalues greater than one (i.e., components explaining at least as much variance as a typical standardized item). Kaiser’s method is the default approach in most statistical packages (e.g., SPSS), although some studies have suggested this approach may lead to over-extraction of components in many applications (cf., Ruscio & Roche, 2012). Undeterred by this, Jolliffe (2002) suggested that it may be useful to use an even more liberal criterion, such as components with eigenvalues > 0.7). Another common method is to extract the number of components needed to account for some specified proportion of the variance (e.g., 90%, 95%) in a set of variables. This approach is particularly useful when large amounts of data can be compactly represented in a small number of components. In psychological research, Velicer (1976) has suggested using the Minimum Average Partial (MAP) correlation. MAP is designed to identify the number of components at which the average partial correlation of the variables, after partialling out m principal components, is a minimum.

Kaiser’s method ignores the fact that eigenvalues are sorted from largest to smallest, thus capitalizing on chance differences in sampling variance. To address this issue, parallel analysis (Horn, 1965) uses data simulated under an independence assumption to subtract out this sampling error variance, and solutions can be evaluated using several different criteria (e.g., mean, median, 95%ile). A wide variety of simulation studies (Ruscio & Roche, 2012) has suggested that parallel analysis provides an unbiased estimate of the number of underlying components and works well practice.

Quite recently, parallel analysis has been extended to consider data generated under data structures with varying numbers of components in the Comparison Data (CD) method (Ruscio & Roche, 2012), with additional components being added to the simulated data for as long as they produce better agreement with the structure of the original data. We do not consider this method here because it can be quite computationally intensive and generally agrees quite well with results from parallel analysis; however, it represents an important addition to the methods available for determining the number of components to be retained.

Finally, in empirical contexts, many researchers rely on scree plots in order to determine the number of components to retain in an analysis. A plot of eigenvalues against the eigenvalue number is used to identify an “elbow” or “large gap” in the data at the point where the useful “signal” degenerates into noise, or “scree.” However, this method provides no definite quantitative cutoffs, and hence is difficult to use for empirical evaluation. The method of profile likelihood recently presented by Zhu and Ghodsi (2006) attempts to address this shortcoming by quantifying the number of retained components that maximizes the observed data likelihood, thus providing an empirical (and automatic) method for determining the point at which a “gap” or “elbow” occurs within a screeplot.

## The Present Studies

This paper proceeds in the following steps. First, we use simulated data to provide guidance regarding the conditions under which each method correctly recovers the underlying number of components in data with known structure (Study 1). Next, we apply these methods to three case study data sets in order to evaluate the performance of our method compared with what is known about missing data for each of these case studies (Studies 2a, 2b, and 2c). Finally, we conclude with a synthesis of these results in order to provide guidance toward a systematic framework for evaluating missing data in psychological research.

# Study 1: A Simulation Study

In order to consider a range of conditions typical of many real-world missing data applications [e.g., Henson & Roberts (2006); Peugh & Enders (2004)), we used a factorial design to consider: (1) the number of components (one or three), (2) number of items per component (three or ten), (3) type of correlation matrix analyzed (Pearson’s r or tetrachoric rho), (4) sample size (100, 250, or 1000), and (5) probability of missing values (10%, 25%, or 50%).

## Simulation Model and Conditions

We constructed 1000 replications in each condition. Population correlations were set at .7 within variables on the same components and .3 between variables on different components, and all data were drawn from multivariate normal distributions. Variables were dichotomized based on whether observed values exceeded the threshold associated with the probability of missing data for that condition.

We evaluated five criteria for selecting the number of components to retain. These included Kaiser’s criterion (eigenvalues > 1), Jolliffe’s criterion (eigenvalues > 0.7), Velicer’s minimum average partial correlation (MAP), Horn’s parallel analysis, and Zhu and Ghodsi’s profile likelihood approach.

## Results and Discussion

Table 1 presents the proportion of replications in which each method recovered the correct number of components by simulation condition. Overall, parallel analysis most often recovered the correct underlying structure (96%), followed by Kaiser (85%), profile likelihood (71%), MAP (61%), and Jolliffe (58%). However, these aggregate results masked considerable variation across study conditions. Additionally, there were striking differences in the importance of each factor in determining the situations where each criterion was more and less likely to yield correct results. In each section, we consider a method successful in a specific condition if it returns the correct number of component in more than 90% of replications.

*Kaiser.* In general, Kaiser’s criterion performed better in recovering the correct number of components with larger sample sizes (e.g., *N* = 1000) and fewer items per component (e.g., 3). It recovered the correct number of components under all conditions when *N* = 1000 (Figures 7 – 9). It also recovered the correct number of components under all conditions when probability of missing data was 50%, and except when *N* = 100 with 10% missing and there are 3 components and 3 items per component (Figure 1a). Kaiser’s criterion performed less with 10 items per component than with 3 items per component. Furthermore, results differed slightly between Pearson’s r and tetrachoric rho matrices as input, with minor differences tending to favor the use of tetrachoric matrices with 10 items per component.

*Jolliffe.* Jolliffe’s criterion more poorly than Kaiser’s in recovering the correct number of components on the conditions considered in this study. Jolliffe’s method appears to be sensitive to sample size and number of items per component, favoring larger sample sizes (e.g., *N* = 1000) and fewer items per component (e.g., 3). Given that *N* = 1000 and 3 items per component, Jolliffe’s criterion recovered the correct number of components under all conditions (95 – 100%; Figures 3), except with 10% missing data and 3 components (60%) and Pearson’s correlations as input. Results were also affected by the number of components and percentage of missing data, favoring fewer components and larger percentage of missingness. Provided the same sample size and items per component, it performed better with 1 component than 3, and 50% missing data, followed by 25% and then 10%. For instance, with *N* = 250 and 3 items per component, Jolliffe’s method recovered the correct number of components for a majority of conditions with 1 component; with 3 components it correctly recovered the number of components only with 25% or 50% missing data (Figures 2). Lastly, the results also differed between Pearson and tetrachoric matrices, with tetrachoric consistently having the same or higher correct recovery rates.

*Velicer’s MAP.* MAP recovered the correct number of components by an average of 82%. In addition to favoring larger sample sizes (e.g., *N* = 250 or 1000; Figures 2 – 3), MAP seems to also respond positively to specific combinations of input matrix and number of items per component: when there are 3 items per component MAP consistently performed better with a tetrachoric input matrix; when there are 10 items per component, MAP performed better with a Pearson input matrix.

*Parallel analysis.* Parallel analysis recovered the correct number of components under all conditions when *N* = 250 (92 – 100%; Figures 2) and *N* = 1000 (100%; Figures 3). In the cases where *N* = 100, parallel analysis still outperformed all other methods in recovering the correct number of components: With 25% and 50% missing data, it recovered the current number of components under all conditions, except for the cases where there are 3 components and 3 items per component (67 – 73%; Figures 1b – c ). In addition, Parallel analysis did not perform as ideally in conditions with *N* = 100, 10% missing data and 10 items per component (Figure 1a). Lastly, there were no differences between results with Pearson or tetrachoric input matrices.

*Profile likelihood.* Profile likelihood recovered the correct number of components (98 – 100%) under all conditions with 1 principal component. Given 1 component, profile likelihood performed as well as parallel analysis when *N* = 100, with a 98-100% of correct recovery rate (Figures 1). However, inspection of profile likelihood values tended to indicate a global maximum at one component, with an additional local maximum at 3 components for models with an underlying structure of 3 components. Profile likelihood is the only method that performed better lower probability of missing data, whereas other methods performed best with higher probabilities of missing data. The combination of 3 components and at least 25% missing data predicts poor performance of profile likelihood in recovering the correct number of components. There were no differences between results with Pearson or tetrachoric correlation matrices.

# Study 2: Applications

The value of any method lies in its performance under real-world circumstances. In a second study, we apply the information gained from the initial simulation with three different data sets in which the missing data mechanisms were unknown to us, but partially knowable. A first application (Study 2a), uses data from Mplus example 11.4, which demonstrates pattern mixture modeling with MNAR data. The characteristics of this data set (relatively large sample size, *N* = 1000, small number of variables, and complex missing data structure) make it an interesting starting point for data that were not generated as part of the current study because several methods can be expected to perform well under these circumstances, thus providing opportunities to achieve consensus. Additionally, because they are also simulated data, it is likely that the underlying structure can be inferred with considerable accuracy even though it is unknown in advance. A second application (Study 2b) uses data from a well-known data set on psychiatric clinical trials examined in detail by Hedeker and Gibbons (1997). This data set has the advantage of having sample size and probability of missing data that make it quite typical for longitudinal psychological research (Henson & Roberts, 2006). This study design is also typical of the kinds of randomized controlled trials studied in health sciences. Additionally, because it was studied in such careful detail by Hedeker and Gibbons (1997), it provided a good opportunity to evaluate the solution suggested by our proposed method and the results obtained in previous research. The final example (Study 2c) considered data on depressive symptoms in the Health and Retirement Study (HRS; Juster & Suzman, 1995). This data set provided the opportunity to evaluate our methods using a large sample size (*N* = 31,169) with data on a large number of variables derived from a large number of waves (10). A further advantage of considering HRS data is that is contains known design features that are responsible for some of the missing values. For example: cohorts were added into the study in various waves; vital status is known for this sample of older adults; physical or cognitive impairment precluded completion of various self-report measures for some individuals in some waves. Thus, we could link the observed pattern of missing data for depressive symptoms measures with external information from a study-wide tracking database that included information about whether individuals should be expected to provide data in each wave, along with whether those data were actually collected. However, missing data patterns were extracted independently of this information, which was used only after the fact in order to validate and interpret the results. We applied the five methods of missing data pattern reduction (i.e., Kaiser, Jolliffe, MAP, parallel analysis, and profile likelihood) with each application. For each application, we describe the example data set, summarize the procedure for missing pattern reduction with each of the five different methods, present the results, and interpret the findings in relations to what is known or can be discerned with regard to missing data in the original study.

## Study 2a (Application 1): Mplus Data Example 11.4 (Muthén & Muthén, 2017)

*Data*. We used the data set, Example 11.4, from the software package, Mplus Version 8 (Muthén & Muthén, 2017), to test our method of missing data pattern reduction. Example 11.4 in the User’s Guide of Mplus is a demonstration of using Mplus to conduct pattern mixture modeling with MNAR data. The variables, y0 to y5, of the data set are 6 waves of observed data, in which y0 has no missing data. The missing data are nonignorable according to Muthén and Muthén (2017).

*Procedure*. We sought to reduce the missing data patterns from the possible 31 (i.e., 25 – 1) to a reasonable number with our method. Specifically, we recoded y1 to y5 into five corresponding missing data indicators, m1 to m5 (Missing = 1, Observed = 0), and ran the five methods on the missing data indicators to see how many principal components to extract as each method indicated, and decided the most appropriate number of components to extract from these data. We then conducted the PCA according to the previous step, and dichotomized each principal component to make a binary variable that indicates data missing or not on the principal missing data pattern. To examine how well the reduced pattern(s) capture the overall missingness, we conducted logistic regression of each binary missing pattern variable on the m1 to m5 in the data set. The percentage of correct classifications indicates that extent to which the reduced pattern(s) capture the overall missing data patterns.

*Results and discussion*. We first used the five criteria to determine the number of principal components to retain with PCA. The Kaiser’s criterion (i.e., Eigenvalue > 1.0), Velicer’s MAP, parallel analysis, and profile likelihood all suggested retaining a single principal component (see Table 2). Given the conditions of this data set (i.e., a total of 5 variables, *N* = 1000, and 26.4–38.4% of missing observations), Kaiser, parallel analysis, and profile likelihood are all indicated (see Table 1) to determine how many principal components (i.e., missing data patterns) should be extracted, which means that, based on the simulation study results, one component should be extracted from the missing data indicators, m1 to m5. Accordingly, we conducted a PCA extracting 1 component from the 5 missing data indicators. All indicators loaded strongly on the extracted component; standardized loadings are all above .40 (See Table 3). We saved the estimated component scores, and recoded the scores (below or equal to 0 = 0, Not Missing; above 0 = 1, Missing) into a new binary variable, Component 1, to indicate whether or not a variable has this pattern of missing data. We found 42.9% of the cases that have missing data on this pattern. As such, our method reduced the possible missing data patterns to a single dichotomous indicator for the Mplus Example 11.4 data set. To examine how well the missing data pattern describes the actual missingness of the data set, we conducted logistic regressions of the dichotomous variable, Component 1 (Not Missing = 0, Missing = 1), on m1 to m5 in the original data set. We estimated 5 logistic regression models of Component 1, on m1, m2, m3, m4 and m5, individually, and then a 6th regression on m1 to m5 together. Results of the logistic regressions are presented in Table 4. Although all shown as significant predictors of the dichotomized component, none of the observed variables was able to completely predict Component 1 as a sole predictor (i.e., correct classification rate below 85%). However, in Model 6, m1 to m5 together predicted the dichotomized Component 1 successfully with a correct classification rate of 100%, indicating that the reduced missing pattern completely describes the overall missingness of the Mplus Example 11.4 data set, and offering support for the idea that a single pattern underlies missing data on all five indicator variables.

## Study 2b (Application 2): National Institute of Mental Health Schizophrenia Collaborative Study (Hedeker & Gibbons, 1997)

*Data*. Hedeker and Gibbons (1997) analyzed data on changes in severity of illness in the National Institute of Mental Health Schizophrenia Collaborative Study to illustrate the influence of missing data with a pattern-mixture approach. The target variable was Item 79, Severity of Illness, on the Inpatient Multidimensional Psychiatric Scale (IMPS79). Responses were on a scale of 1 (normal, not at all ill) to 7 (among the most extremely ill), which were treated in the analysis as continuous variables. Participants were grouped by placebo and drug; the main measurements occurred at Weeks 0, 1, 3, and 6. Defining completion as being observed at Week 6 (i.e., the last measurement), Hedeker and Gibbons (1997) found that the completion rate of the Placebo group (81%) was significantly higher than that of the Drug group (65%). For the completers, across all four measurements both patients on drug and on placebo experienced decreases in severity of illness, and the decrease was greater for those on drug than for those on placebo. For the dropouts, across the first three measurements, the patients on drug experienced significant decreases, whereas those on placebo did not experience much change in severity of illness. Hedeker and Gibbons’ (1997) approach reduced the missing patterns to one—missing at Week 6 or not—which was found to be non-ignorable due to the significant interaction between dropout, on drug, and the change in IMPS79 scores over time (*b* = -.635, SE = .196, *p* = .002; p. 71). Patients on drug also demonstrated significantly different change in IMPS79 scores compared to those on placebo (*b* = -.539, SE = .086, *p* = .001).

*Procedure*. We sought to reduce the missing data patterns from the possible 15 (i.e., 24 – 1) to a reasonable number with our method. Specifically, we recoded IMPS79\_0, IMPS79\_1, IMPS79\_3, and IMPS79\_6 into four corresponding missing data indicators, IMPS79.0m, IMPS79.1m, IMPS79.3m, and IMPS79.6m (Missing = 1, Observed = 0), and ran the five methods on the missing data indicators to see how many principal components to extract as each method indicated, and decided the most appropriate number of components to extract from these data. We then conducted the PCA according to the previous step, and dichotomized each principal component to make a binary variable that indicates data missing or not based on the principal missing data pattern. We then analyzed changes in severity of illness with the reduced missing data pattern(s), to compare our findings with those by Hedeker and Gibbons (1997).

*Results and discussion*. We first conducted the five methods on the 4 missing data indicators to determine the number of principal components to be extracted with a PCA. The Kaiser criterion (i.e., Eigenvalue > 1.0), parallel analysis, and profile likelihood all suggested retaining a single principal component (see Table 2). Given the characteristics of this data set (i.e., a total of 4 variables, *N* = 437, and 1.5–54.6% of missing observations), Kaiser, parallel analysis, and profile likelihood are all suitable (see Table 1) to determine how many principal components (i.e., missing data patterns) should be extracted, which means that, based on the simulation study results, 1 component should be extracted from the missing data indicators. Accordingly, we conducted a PCA extracting 1 component from the four missing data indicators. The indicators loaded on the extracted component with standardized loadings ranging from .130 to .863 (See Table 3). We saved the estimated component scores, and recoded the scores (below or equal to 0 = 0, Not Missing; above 0 = 1, Missing) into a new variable, Component 1, to indicate whether or not a variable has this pattern of missing data. We found 27.9% of the cases that have missing data on this pattern. As such, our method identified one missing data pattern for the Hedeker and Gibbons (1997) data set. The high loadings of IMPS79.3m and IMPS79.6m on the principal component (Table 3) indicate that the missing pattern is to a large extent driven by missing at Weeks 3 and 6. This result is to a large extent consistent with the pattern found by Hedeker and Gibbons (1997) in describing missing at Week 6. To compare our missing data pattern with the one found by Hedeker and Gibbons (1997) in terms of its influence on changes in IMPS79 scores, we modeled the change in IMPS79 scores from Week 0 to Week 6 using mixed models with three covariates: 1) Drug (on drug vs. on placebo), 2) Component 1 (missing vs. observed based on Component 1), and 3) the Drug×Component1 interaction. Results indicated that the main effect of Drug and the Drug×Component1 interaction were both significant on the change in IMPS79 scores (*b* = -.521, SE = .127, *p* < .001; *b* = -.514, SE = .225, *p* = .022, respectively), and that the main effect of Component 1 was not significant on the change in IMPS79 scores (*b* = -.113, SE = .188, *p* = .547). We found one missing data pattern—Component 1—with our missing data reduction method. Consistent with the findings by Hedeker and Gibbons (1997), this pattern was associated with changes in IMPS79 scores, in the form of an interaction with Drug. As shown in the prototypical trajectories (Figure 6), 1) on average, patients on drug experienced greater decreases in severity of illness (i.e., IMPS79 scores) than those on placebo; 2) within the group on placebo, the missing data group started with initially higher levels of severity but also decreased to a greater degree compared to those who completed the treatment, such that the resulting levels of severity of illness were comparable for both the missing and the complete data groups on placebo; 3) most interestingly, among those receiving treatment, the missing data group had initially similar levels of severity but experienced significantly greater decreases compared to those without missing data. Our findings highlight the crucial role of missing data at Weeks 3 and 6 in the estimated decrease of illness severity within the group on placebo and the one on drug.

## Study 2c (Application 3): Depressive Symptoms in the Health and Retirement Study (Juster & Suzman, 1995)

*Data*. We used the interview data on depressive symptoms from 10 waves (1992 through 2008) of the Health and Retirement Study (HRS). The HRS is a large, nationally representative multi-wave survey containing detailed information on a wide range of topics. Baseline data for the HRS were first collected in 1992 through face-to-face interviews conducted with respondents aged 51-61 (birth cohort 1931-1941) and their spouses. The original sample consisted of 7,600 households and more than 12,600 persons and was based on a multi-stage area probability design oversampling Hispanics, African Americans and Florida residents. The response rate was over 80% (for further details see Juster & Suzman, 1995). Follow-ups by phone are performed biannually, with proxy exit interviews for deceased individuals. Beginning in 1998, additional cohorts were added to the sample, drawn from Aging and Health Dynamics in the Oldest Old (AHEAD, born 1923 or before), Children of the Depression Age (CODA, 1924-1930) and War Babies (WB, 1942-1947). The overall sample size is 31,169. Our primary variables were depressive symptoms as measured by the Center for Epidemiological Studies Depression (CES-D) scale (Radloff, 1977). The HRS uses a subset of 8-items from the original 20-item version of the Center for Epidemiological Studies Depression (CES-D) scale. In terms of the structure of the CES-D items assessed in the HRS, there is representation from the most important three of four CES-D subscales. Specifically, there are three items (depressed, lonely, sad) from the depressed affect (DA) subscale, two items (happy, enjoyed life) from the positive affect (PA) subscale, and three items (effort, sleep, get going) from the somatic complaint (SO) subscale. Not represented is the domain of interpersonal problems. The resulting variables were subscale scores on depressed affect (DA), positive affect (PA), and somatic complaint (SO) for each wave, which makes a total of 30 variables over the 10 waves that were examined in the present study.

*Procedure*. We sought to reduce the missing data patterns from the possible 1,073,741,823 (i.e., 230 – 1) to a reasonable number with our method. Specifically, we recoded the 30 CES-D variables into 30 corresponding missing data indicators, DA1m to SO10m (Missing = 1, Observed = 0), and ran the five methods on these missing data indicators to see how many principal components to extract as each method indicated, and decided the most appropriate number of components to extract from these data. We then conducted the PCA according to the previous step, and recoded each principal component into a binary variable that indicates data missing or not on the missing data pattern. To examine how well the reduced missing data pattern(s) capture the overall missingness, we conducted logistic regressions of each binary missing pattern variable on 10 variables that inform the study design and participant status. A high percentage of correct classifications indicates that the reduced pattern(s) to a great extent capture the overall missingness of the data.

*Results and discussion*. We first conducted the five methods on the 30 missing data indicators to determine the number of principal components to be extracted with a PCA. The Kaiser criterion (i.e., Eigenvalue > 1.0) and parallel analysis suggested 4 principal components, Jolliffe suggested 5 components, Velicer’s MAP indicated 3, and Profile Likelihood indicated 2 principal components (see Table 2). Given the conditions of this data set (i.e., a total of 30 variables, *N* = 31,169, and 38.1 – 65.8% of missing observations), Kaiser and parallel analysis are both suitable (see Table 1) to determine how many principal components (i.e., missing data patterns) should be extracted, which means that, based on the simulation study results, 4 components should be extracted from the missing data indicators, DA1m to SO10m. Accordingly, we conducted a PCA extracting 4 components from the 30 missing data indicators. With a cutoff value of .40 for the item loadings, we were able to identify four components (i.e., four patterns of missing data). The first component was loaded with the CES-D variables from Waves 8 to 10, the second component was loaded with variables from Waves 4 to 6, the third component was loaded with variables from Waves 6 to 8, and the fourth component was loaded with variables from Waves 1 to 3 (See Table 3 for standardized item loadings). We saved the estimated component scores, and recoded the scores (below or equal to 0 = 0, Not Missing; above 0 = 1, Missing) into 4 new variables, Components 1 to 4, each indicating whether or not a variable has missing data on the corresponding pattern. We found 53.1% of the cases that have missing data on pattern 1, 44.6% have missing data on pattern 2, 42.6% have missing data on pattern 3, and 64.1% have missing data on pattern 4. As such, our method reduced the possible missing data patterns to four for the CES-D data of HRS, each of which is structured around missing data across contiguous waves of data collection. In order to further understand what information the four reduced missing data patterns convey about the HRS data set, we integrated external information from a small number of variables on participant status and study design of the HRS data set, including the participant’s vital status (i.e., alive), whether a proxy provided interview data for the participant (i.e., proxy, which by study protocol indicated that self-reported depressive symptom data should not be collected), whether interview data was obtained in a given wave (i.e., iwwave), and the year when first interviewed (i.e., firstiw) , with which we created 10 index variables (index01 to index10) to show whether or not a participant’s CES-D data should be missing at a measurement occasion based on the aforementioned information. For instance, the variable, index03, represents whether or not a participant’s CES-D data should be missing in 1996: if a participant was interviewed before or in 1996 for the first time, and alive in 1996, his or her interview data were obtained within this wave and were not provided by a proxy, then this participant’s CES-D data should not be missing in 1996 and the corresponding index03 variable should have a score of 0. See Table 5 for frequencies of index01 to index10. We conducted four logistic regression models of Component1 to Component4, each on all ten variables (See Table 6a for model coefficients). As presented in Table 6b, all four 10-predictor logistic regressions converged normally and provided better fit to the data than the corresponding null models with only the intercept. In each model, there were multiple predictors which had significant effects on the dependent variable. Most importantly, the models all had high sensitivity and specificity, and an overall rate of correct classification close to 1.

# Discussion

Growing accessibility of methods for analysis of missing data mechanisms has not been matched by consensus on methods for systematically evaluating the assumptions on which successful application of these methods rest. In biostatistics, where a single outcome is typically of interest, the missing data mechanism is relatively straightforward to explore. In psychological and educational research, data are often multivariate and longitudinal, which requires a reasoned approach to data reduction prior to examination of missing data mechanisms, since the number of potential missing data patterns to consider grows exponentially with the number of variables in the analysis. We evaluate principal components analysis with a number of commonly applied methods for determining the number of components to retain, using simulated and real data examples representing the broad spectrum of conditions encountered in psychological and educational research. Under nearly all conditions, our results point to a number of clear and concrete recommendations. In terms of evaluating performance of various methods for determining the number of components to extract, our results align well with the general consensus in the literature. Further, because we are primarily interested in the number of components rather than the component loadings, our results suggest that these methods work just as well with Pearson’s *r* as with more computationally intensive tetrachoric correlations. Consistent with the broader literature on component extraction, our results suggest that parallel analysis should be the preferred method to use. Using simulated data, this method recovers the correct number of components in over 90% of cases under every condition we studied for sample sizes of at least 250. Because this intermediate sample size was selected on the basis of a “typical” study in psychological and educational research, this suggests that the approach we outline here is likely to be effective in a majority of circumstances encountered by psychological and educational researchers. These recommendations should be qualified when applied to small (*N* = 100) sample sizes. Under these circumstances, although parallel analysis generally performs best even under these cases, the percentage of replications that returns the correct number of components (~70%) is substantially lower than our selected threshold of 90% correct. In these cases, we note that parallel analysis has a tendency to under-extract the number of components by one (27%). This suggests that our recommended strategy of extracting one more and one fewer component captures the true number of components in 97% of cases even under the poorest performing condition (*N* = 100). Another possibility for analyses with small sample sizes is that methods such as bootstrap aggregation (“bagging,” Breiman, 1996) may provide more accurate results in this situation. Finally, we note that even when profile likelihood is incorrect with small sample sizes, the extent of difference between the true and extracted number of components is very small, with a (very slight) bias toward over/under extraction. We also find evidence that this approach is likely to work well in practice, using one simulated data set generated according to an unknown mechanism known to be MNAR, a data set with demonstrated MNAR mechanism that has received considerable careful analysis in the literature, and a large longitudinal population-based study where several design characteristics partially determine the pattern of missing data. Analysis of these applications leads us to recommend a set of specific procedures for examining potential missing data mechanisms in practice (Table 7).

Step 1 is to construct dichotomous indicators of all variables with missing data (1 = missing, 0 = observed). Step 2 is to perform a PCA on missing data indicators. Step 3 is to determine the number of components to retain using parallel analysis or similar criterion. Step 4 is to extract the desired number of principal components and calculate predicted component scores. Step 5 is to dichotomize component scores. This can be determined empirically, such as by examining the distribution of components scores or else based on a split at zero. Step 6 is to examine distributions of observed variables and missing data indicators by each component score. Step 7 is to predict dichotomized component scores from observed variable and/or missing data indicators using logistic regression or similar method. Step 8, if the pattern is difficult to interpret, examine cross-tabulations by dichotomized component scores. Step 9, if the pattern of results is still not clear, consider examining cross-tabulations by one or more stratifying variables (e.g., treatment condition, gender). We recommend this step in all situations with an experimental manipulation and/or stratification variable. This is especially important for any models that will include evaluation of one or more interaction terms. In Step 10, examine missing data patterns by any known characteristics of the study design in order to evaluate the external validity of the missing data patterns obtained. We recommend repeating this process using one more and one fewer than the number of components extracted. Missing data always introduce uncertainty about unobserved values. In the best circumstances, they do so in a similar fashion to the way that analysis of data from a sample introduces uncertainty about parameters in the population, but informative non-response can have considerable consequences for study conclusions if not more fully evaluated. Further, methods for the analysis of MNAR data are still emerging and this problem has been shown theoretically to be intractable based only on the observed data (Mohan, Pearl, & Tian, 2013). For this reason, recommendations for analyses with missing data increasingly revolve around sensitivity analyses. This shifts the emphasis toward methods that reduce, or at least better characterize, the extent of potential bias. Of particular interest is the extent to which study conclusions depend on the assumptions, implicit or explicit, being made about the nature of unobserved values. The approach we recommend here is strongly consistent with a sensitivity analysis framework in two ways. First, examination of associations with a smaller number of missing data components allows each of them to be considered in greater depth. We have outlined a minimal set of specific and systematic recommendations that can be used to guide this process for a given number of missing data components. Second, we also advocate consideration of different numbers of missing data components (e.g., one more and one fewer than extracted on the basis of parallel analysis). In many cases, results agree very closely across these conditions. In circumstances where results diverge, this is important information that can be used to guide sensitivity analyses and/or qualify results.

## Limitations

We have considered a broad subset of conditions that reflect the range of most psychological research. However, researchers might want to conduct similar simulations that more closely align with their own data. In real data applications, we don’t know the ground truth, and so cannot truly evaluate performance with real data. However, our results with real data are both broadly interpretable and consistent with previous analyses that are available for two of the data sets we considered here. Another important consideration is that study results, particularly adjustments for factors associated with the missing data mechanism are only going to be as good as the quality of ancillary data available. This might suggest that researchers devote greater attention to collecting additional follow-up data from participants as possible, something that has long been recommended practice in the field of survey research (e.g., Little & Rubin, 2002). Future Directions This study presents one validated approach to promote consideration of missing data mechanisms in multivariate and/or longitudinal contexts, but considerable opportunities remain with regard to better aligning missing data theory and practice. For example, our missing data mechanisms had known structure for the simulated data, but not for the real data applications. Future research should examine the performance of these procedures using real data with deliberately masked observations. Similarly, in many areas of research, such as analysis of clinical trials data, there is appropriate concern on missing data methods that can be “gamed” or manipulated by the data analyst. This can sometimes lead to widespread adoption of some less favorable but objective treatments of missing data, such as last observation carried forward (Little et al., 2012). Although this is one of our goals in presenting this approach in as systematic and replicable a manner as possible, and situating it explicitly within a sensitivity analysis framework, future research should carefully evaluate where missing data protocols need to be modified in order to further remove or reduce subjectivity in the application of missing data methods. In summary, we present a standardized procedure for exploring missing data mechanisms with multivariate and/or longitudinal data. We evaluate several commonly used criteria for selecting the number of principal components using dichotomous missing data indicators as one tool for guiding investigation of potential missing data mechanisms and sensitivity analyses when considering the assumptions about missing data for study conclusions. Results suggest that profile likelihood performed best under nearly all circumstances considered here.

# References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. <https://doi.org/10.1007/BF00058655>

Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach*. Routledge.

Davey, A., Shanahan, M. J., & Schafer, J. L. (2001). Correcting for selective nonresponse in the national longitudinal survey of youth using multiple imputation. *The Journal of Human Resources*, *36*(3), 500. <https://doi.org/10.2307/3069628>

Enders, C. K. (2011a). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*(4), 267–288. <https://doi.org/10.1037/a0025579>

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

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19*(1), 39–55. <https://doi.org/10.1037/a0035314>

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–240. <https://doi.org/10.1037/met0000063>

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer. <https://doi.org/10.1007/978-1-4614-4018-5>

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. <https://doi.org/10.1207/S15328007SEM1001_4>

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*(1), 64–78. <https://doi.org/10.1037/1082-989X.2.1.64>

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393–416. <https://doi.org/10.1177/0013164405282485>

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. <https://doi.org/10.1007/BF02289447>

Jo, B., Ginexi, E. M., & Ialongo, N. S. (2010). Handling missing data in randomized experiments with noncompliance. *Prevention Science*, *11*(4), 384–396. <https://doi.org/10.1007/s11121-010-0175-4>

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Juster, F. T., & Suzman, R. (1995). An overview of the health and retirement study. *The Journal of Human Resources*, *30*, S7. <https://doi.org/10.2307/146277>

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. <https://doi.org/10.1177/001316446002000116>

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed). Wiley.

Little, R. J., D’Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., & Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355–1360. <https://doi.org/10.1056/NEJMsr1203730>

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. <https://doi.org/10.1037/met0000096>

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user’s guide* (Eighth). Muthén & Muthén.

Noble, S. M., Hollingworth, W., & Tilling, K. (2012). Missing data in trial-based cost-effectiveness analysis: The current state of play. *Health Economics*, *21*(2), 187–200. <https://doi.org/10.1002/hec.1693>

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556. <https://doi.org/10.3102/00346543074004525>

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. <https://doi.org/10.1177/014662167700100306>

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282–292. <https://doi.org/10.1037/a0025697>

Shin, T., Davison, M. L., & Long, J. D. (2017). Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychological Methods*, *22*(3), 426–449. <https://doi.org/10.1037/met0000094>

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327. <https://doi.org/10.1007/BF02293557>

Xu, S., & Blozis, S. A. (2011). Sensitivity analysis of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, *36*(2), 237–256. <https://doi.org/10.3102/1076998610375836>

Yang, M., & Maxwell, S. E. (2014). Treatment effects in randomized longitudinal trials with different types of nonignorable dropout. *Psychological Methods*, *19*(2), 188–210. <https://doi.org/10.1037/a0033804>

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, *51*(2), 918–930. <https://doi.org/10.1016/j.csda.2005.09.010>