Behaviour Research and Therapy xxx (2016) 1–15

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Multiple imputation as a flexible tool for missing data handling in clinical research

Craig K. Enders

UCLA, Department of Psychology, 3586 Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095, United States

ARTICLE INFO

Article history:
Received 9 September 2016
Received in revised form
16 November 2016
Accepted 17 November 2016
Available online xxx

Keywords:
Missing data
Multiple imputation
Attrition
Maximum likelihood estimation

ABSTRACT

The last 20 years has seen an uptick in research on missing data problems, and most software applications now implement one or more sophisticated missing data handling routines (e.g., multiple imputation or maximum likelihood estimation). Despite their superior statistical properties (e.g., less stringent assumptions, greater accuracy and power), the adoption of these modern analytic approaches is not uniform in psychology and related disciplines. Thus, the primary goal of this manuscript is to describe and illustrate the application of multiple imputation. Although maximum likelihood estimation is perhaps the easiest method to use in practice, psychological data sets often feature complexities that are currently difficult to handle appropriately in the likelihood framework (e.g., mixtures of categorical and continuous variables), but relatively simple to treat with imputation. The paper describes a number of practical issues that clinical researchers are likely to encounter when applying multiple imputation, including mixtures of categorical and continuous variables, item-level missing data in questionnaires, significance testing, interaction effects, and multilevel missing data. Analysis examples illustrate imputation with software packages that are freely available on the internet.

The methodological literature on missing data handling spans many decades, but the modern era of this work arguably began when Rubin (1976) established a theoretical framework for missing data problems. Since then, there has been a substantial increase in missing data research, and most software applications now implement one or more sophisticated missing data handling routines. Despite the uptick in methodological research and the concurrent publication of several missing data texts (Allison, 2002; Carpenter & Kenward, 2013; Enders, 2010; Graham, 2012; Little & Rubin, 2002; van Buuren, 2012), the migration to better analytic practices has understandably been slow. Going back to the 2000s, literature reviews revealed that researchers relied primarily on deletion methods that remove cases with missing data (Jelicic, Phelps, & Lerner, 2009; Peugh & Enders, 2004; Wood, White, & Thompson, 2004), despite warnings that these "are among the worst methods available for practical applications" (Wilkinson & Taskforce on Statistical Significance, 1999, p. 598). Although reporting practices have definitely improved in recent years, the application of modern missing data handling techniques is far from uniform in psychology and related disciplines. Consequently, the primary goal of this manuscript is to promote the awareness and application of analytic methods that enjoy strong support in the methodological literature.

Broadly speaking, the recent missing data literature supports the use of maximum likelihood estimation and multiple imputation (Schafer & Graham, 2002).[1] Maximum likelihood estimation (also known as full information maximum likelihood, or FIML) employs an iterative optimization algorithm that identifies parameter estimates that maximize fit to the observed data. For example, in a regression analysis, the maximum likelihood estimates are coefficients that minimize the sum of the squared standardized distances between the observed data and the regression line. Some methodologists have characterized maximum likelihood estimation as "implicit imputation" because it does not produce a filled-in data set (Widaman, 2006). Rather, the procedure uses all of the available data to estimate a specific set of model parameters and

[1] Bayesian estimation is a third option that I do not consider here in the interest of space. A Bayesian analysis mimics maximum likelihood estimation in the sense that it generates estimates and standard errors for a specific analysis model. However, the missing data handling aspect of Bayesian estimation resembles multiple imputation because each cycle of the iterative algorithm generates a filled-in data set.

E-mail address: cenders@psych.ucla.edu.

http://dx.doi.org/10.1016/j.brat.2016.11.008
0005-7967/

their standard errors. For example, to apply maximum likelihood to an ANOVA-type analysis, a researcher need only use a capable software package to estimate a regression model from the incomplete data. Structural equation modeling software packages are particularly useful for implementing maximum likelihood because they can accommodate a range of missing data patterns (e.g., missing values on explanatory and outcome variables).

In contrast, multiple imputation creates several versions of a data set, each of which contains different estimates of the missing values. As explained later, most incarnations of multiple imputation use a regression model to fill in the data, treating incomplete variables as outcomes and complete variables as predictors. To avoid imputations based on a single set of regression parameters, an iterative algorithm uses Bayesian estimation to update the regression model parameters, and it uses new estimates to generate each set of imputations. Having generated a set of filled-in data sets, the researcher then performs one or more statistical analyses on each complete data set to obtain imputation-specific estimates and standard errors. The final step pools the estimates and standard errors into a single set of results.

With normally distributed data, a common set of input variables, and a sufficiently large sample size, there is no theoretical reason to expect differences between maximum likelihood estimation and multiple imputation (Gelman et al., 2014; Meng, 1994; Schafer, 2003), and empirical studies suggest that the two methods usually yield similar estimates and standard errors (Collins, Schafer, & Kam, 2001).[2] All things being equal, maximum likelihood estimation is probably preferable for many situations on the basis of simplicity alone – as noted previously, a researcher need only translate the desired analysis to a capable software package. However, psychological data sets often feature complexities that are currently difficult to handle appropriately in the likelihood framework. A regression analysis with mixtures of categorical and continuous variables is a very simple, yet common, scenario where maximum likelihood estimation is not optimal. For example, consider a model with a nominal covariate (e.g., race, diagnostic category, gender) and a continuous outcome. A complete-data regression analysis uses a set of dummy codes to represent the nominal covariate, and it does so without imposing distributional assumptions on predictors. In contrast, maximum likelihood missing data handling requires distributional assumptions for the incomplete variables, and software packages would typically force the user to treat a set of incomplete dummy codes as though they were multivariate normal (and some software programs will simply exclude cases with missing predictor scores). An analysis that features scale scores computed from a set of questionnaire items is another common situation where maximum likelihood missing data handling is surprisingly difficult. Because it does not fill in the data, maximum likelihood effectively encourages the user to treat the scale as missing when one or more of its component items is missing. Specifying an analysis that leverages the typically-strong correlations among the items can be difficult, and ignoring this source of information can decimate power (Gottschall, West, & Enders, 2012; Mazza, Enders, & Ruehlman, 2015).

In my experience, multiple imputation is often a better tool for behavioral science data because it gives researchers the flexibility to tailor the missing data handling procedure to match a particular set of analysis goals. For example, mixtures of categorical and continuous variables (e.g., a regression analysis with an incomplete nominal covariate) pose no problem for multiple imputation, and composite scores with incomplete item responses are similarly benign. Because a number of accessible descriptions of maximum likelihood estimation appear in the literature (Enders, 2010, 2013; Graham, 2012; Schafer & Graham, 2002), I limit the scope of this manuscript to multiple imputation, focusing on practical issues that clinical researchers are likely to encounter in their work. Throughout the paper, I use a series of data analysis examples to illustrate the application of multiple imputation to problems that are not necessarily easy to handle with maximum likelihood estimation. Although multiple imputation is widely available in most general-use software packages, I use the Blimp application (Enders, Keller, & Levy, 2016; Keller & Enders, 2014) because it is flexible enough to accommodate a variety of scale types (nominal, ordinal, and continuous) with single-level and multilevel data, and it can be used in conjunction with any analysis program. Blimp is available for the Mac and Windows operating systems and is available for free download at www.appliedmissingdata.com/multilevel-imputation.html.

1. Motivating example

The analysis example comes from a study of an online chronic pain management program (Ruehlman, Karoly, & Enders, 2012), where individuals were randomly assigned to an intervention condition ($n = 167$) or a wait-listed control group ($n = 133$). The primary focus of this example is a 6-item depression measure, which researchers administered at pretest, 7-week follow-up, and 14-week follow-up. The data set also includes a number of background variables (e.g., gender, age, education) and baseline measures of pain severity and pain interference with daily life activities. Table 1 gives the percentage of observed values for a subset of variables that I use throughout the paper. So that interested readers can work through the data analysis examples, I used the means and correlations from real data to create an artificial data set that mimics the original. The data set and analysis scripts are available

Table 1
Percentage of observed data for analysis variables.

Variable	Name	% Complete	Range
Intervention code	TXGRP	100.0	0–1
Gender	Male	100.0	0–1
Age	Age	100.0	18–78
Education	Educ	95.0	1–7
Exercise frequency	Exercise	93.3	1–8
Pain interference	Interf	100.0	6–42
Pain severity rating	Severity	93.7	1–7
Wave 1 depression item 1	T1DEP1	100.0	1–5
Wave 1 depression item 2	T1DEP2	94.7	1–5
Wave 1 depression item 3	T1DEP3	94.7	1–5
Wave 1 depression item 4	T1DEP4	100.0	1–5
Wave 1 depression item 5	T1DEP5	100.0	1–5
Wave 1 depression item 6	T1DEP6	100.0	1–5
Wave 2 depression item 1	T2DEP1	84.7	1–5
Wave 2 depression item 2	T2DEP2	86.7	1–5
Wave 2 depression item 3	T2DEP3	91.3	1–5
Wave 2 depression item 4	T2DEP4	91.3	1–5
Wave 2 depression item 5	T2DEP5	91.3	1–5
Wave 2 depression item 6	T2DEP6	86.7	1–5
Wave 3 depression item 1	T3DEP1	76.3	1–5
Wave 3 depression item 2	T3DEP2	76.3	1–5
Wave 3 depression item 3	T3DEP3	76.3	1–5
Wave 3 depression item 4	T3DEP4	72.7	1–5
Wave 3 depression item 5	T3DEP5	72.7	1–5
Wave 3 depression item 6	T3DEP6	74.0	1–5
Wave 1 depression scale	DEP1	89.7	6–30
Wave 2 depression scale	DEP2	77.7	6–30
Wave 3 depression scale	DEP3	68.7	6–30

[2] It is difficult to identify a rule of thumb for a "sufficiently large" sample, but my experience suggests that multiple imputation and maximum likelihood can yield equivalent estimates in samples that are typical in psychological research (e.g., $N = 200$).

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

for download at appliedmissingdata.com/multilevel-imputation.html, and the analysis scripts additionally appear in the appendices of this document.

Projects like the pain management study often involve a team of researchers using different parts of a large data set consisting of hundreds or thousands of variables. The missing data literature sometimes recommends a large-scale missing data handling procedure that accounts for dozens of variables (Rubin, 1996). However, the complexity of psychological data sets usually precludes this strategy, and the sample sizes that are typical of such studies are also a limiting factor (e.g., the number of variables used to impute the data cannot exceed the number of cases, and usually needs to be much lower). In my experience, it is often better to focus on a specific analysis or family of analyses because it is easier to implement a missing data handling procedure that honors important features of the data. To illustrate this focused strategy, I consider a regression analysis that models the influence of the intervention on Wave 3 depression scores, controlling for age and baseline measures of depression and pain severity

$$DEP_3 = \beta_0 + \beta_1(DEP_1) + \beta_2(AGE) + \beta_3(SEVERITY) + \beta_4(TXGRP) + \varepsilon \quad (1)$$

where *TXGRP* is a binary dummy code (0 = control, 1 = intervention), *SEVERITY* is 7-point ordinal rating, and DEP_1 and DEP_3 are scale scores computed by summing the Wave 1 and Wave 3 questionnaire items, respectively. I chose this model because it introduces complexities that are common in behavioral science research (e.g., mixtures of categorical and continuous variables, composite scores) — complexities that, in my view, often favor multiple imputation as a solution.

2. Background and terminology

Rubin and colleagues (Little & Rubin, 2002; Rubin, 1976) developed a theoretical framework for missing data handling problems that is key to understanding the strengths and limitations of different analytic approaches. Rubin's theory describes three missing data mechanisms that, roughly speaking, posit different "reasons" for the missing values: the missing completely at random (MCAR) mechanism describes situations where missingness is haphazard and unrelated to analysis variables, whereas the missing at random (MAR) and not missing at random (NMAR) mechanisms describe two types of systematic nonresponse. From a practical perspective, these mechanisms function as assumptions that dictate the accuracy of a missing data handling procedure; a method such as multiple imputation (or maximum likelihood) that assumes MAR should yield accurate estimates when an MAR mechanism is tenable for a particular analysis, whereas a procedure that assumes MCAR (e.g., deleting cases) would produce biased estimates in the same scenario.

Before describing the mechanisms in more detail, it is important to emphasize that missing data mechanisms and missing data patterns are different concepts. The regression analysis from Equation (1) involves seven missing data patterns; 182 cases have complete data on all variables, and the remaining observations are spread across subgroups with one or more missing values. The broader data set is further characterized by patterns of intermittent missingness (e.g., a participant skipped some questionnaire items but otherwise participated) and attrition (e.g., a participant permanently quit the study and is missing one or more waves of data). Note that these patterns simply tally the configuration and frequency of the holes in the data, but they say nothing about systematic tendencies that predict nonresponse. Because modern analysis methods such as multiple imputation are generally equipped to handle a wide range of patterns, the plausibility of different missing data mechanisms is usually our primary concern.

Rubin (1976)'s missing data theory defines a hypothetical data set with no missing values, and it partitions the realized data into observed and missing components. It is useful to view the missing parts as latent variable scores, the values of which reside only in the hypothetically-complete data matrix. Each incomplete variable is yoked to a corresponding dummy code that indicates whether scores are observed or missing (e.g., $M = 0$ if complete, and $M = 1$ if missing). The crux of Rubin's theory is that the nonresponse indicators may be completely unrelated to the data, or they may be systematically related to either the observed or latent scores (or both). Returning to the variables in Equation (1), we can create three missing data indicators, one each for pain severity and the two depression measures (age and treatment group membership are complete). The missing data mechanisms describe three possible relations between the nonresponse indicators and the data.

An MCAR (missing completely at random) mechanism occurs when missingness is unrelated to the data, missing or latent. The regression analysis would be MCAR if the nonresponse indicators for the incomplete variables are unrelated to other variables (e.g., cases with observed and missing depression scores at Wave 3 are identical with respect to age, gender, treatment group membership, baseline depression, and so on). MCAR values can result from haphazard events (e.g., a respondent's internet connection unexpectedly cuts out during data collection), causes of missingness that are uncorrelated with the analysis variables, or as part of a planned missingness design that saves resources and reduces respondent burden (Graham, Taylor, Olchowski, & Cumsille, 2008; Little & Rhemtulla, 2013; Mistler & Enders, 2011). It is worth noting that MCAR is the only mechanism with testable propositions. MCAR can be refuted, for example, if groups formed by the missing data indicators exhibit mean differences on other variables. Little's (1988) procedure is a multivariate test of such differences, and researchers often perform univariate *t* tests (or correlations) to examine relations between the missing data indicators and observed variables. Although the absence of mean differences does not confirm an MCAR mechanism (Raykov, 2011), the presence of mean differences does rule out a purely random nonresponse mechanism.

The MAR (missing at random) and NMAR (not missing at random) mechanisms define different types of systematic missingness. An MAR mechanism occurs when the nonresponse indicators are related to only the observed data, and an NMAR mechanism further allows the latent scores to influence missingness. For example, we could imagine a situation where individuals with high levels of pain interference are missing Wave 3 depression scores because pain impedes their ability to sit in front of a computer for extended periods of time. This situation qualifies as MAR provided that the latent (hypothetical) depression scores do not further predict attrition (i.e., two participants with the same pain interference score are equally likely to drop out, regardless of their Wave 3 depression levels). Finally, an NMAR mechanism would apply to the regression analysis, for example, if Wave 3 depression scores are missing for individuals with high levels of depression at that follow-up (e.g., because they perceive no benefit from the intervention and drop out).

In practice, it is difficult to determine which mechanism best applies to a particular analysis because Rubin's conditions involve propositions about the unobserved latent scores; MCAR and MAR stipulate that missingness is unrelated to these hypothetical values, whereas NMAR allows for a linkage. Of course, without access to the latent scores, it is impossible to know whether they predict missingness, and so we are ultimately forced to adopt an untestable assumption about the process that caused missing data (Raykov,

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

2011). As noted previously, modern missing data handling methods such as multiple imputation and maximum likelihood estimation typically require an MAR mechanism, whereas excluding cases with missing data assumes an often-unrealistic MCAR mechanism with no sources of systematic missingness. In practical terms, adopting an MAR-based approach such as multiple imputation ensures that we can obtain accurate estimates in a broader range of circumstances than we could be simply excluding incomplete cases. Importantly, this advantage is largely unrelated to the amount of missing data; if the imputation procedure satisfies MAR, the resulting estimates can tolerate rather extreme levels of missingness (e.g., 50% or more; Enders, 2010).

3. Selecting variables for imputation

The MAR assumption that is crucial to multiple imputation requires that, after conditioning on the observed data, the hypothetical latent scores carry no information about nonresponse. In practical terms, MAR implies that the missing data indicators carry no additional information about the missing scores above and beyond that contained in the observed data. MAR provides an important simplification that allows us to generate imputations (or perform an analysis, in the case of maximum likelihood estimation) without including the nonresponse indicators in the missing data handling procedure. On the other hand, obtaining accurate estimates from an NMAR mechanism requires complicated modeling approaches that introduce the indicators in one form or another (Enders, 2011; Muthen, Asparouhov, Hunter, & Leuchter, 2011). Although MAR may seem straightforward, satisfying this assumption often requires that we look beyond the variables in our analysis and condition on (control for) so-called auxiliary variables that differentiate the complete and incomplete cases. This section describes the logic of the so-called inclusive analysis strategy (Collins et al., 2001) that attempts to identify auxiliary variables that are useful for imputation, either because they reduce nonresponse bias or improve power.

To illustrate the role that auxiliary variables play in satisfying the MAR assumption, Panel A of Fig. 1 depicts the true associations among six variables in a data set: X and Y are variables in the analysis model, M_Y is the missing data indicator for Y, and A_1, A_2, and A_3 are potential auxiliary variables that could be included in imputation. In Panel A, the absence of an arrow connecting Y and M_Y indicates that MAR is satisfied if the imputation procedure includes A_1 because this variable fully explains the relation between Y and M_Y (i.e., there is no residual relation between Y and M_Y after controlling for A_1). Panel B depicts the relations that result from excluding A_1 from the imputation regression model. Ignoring this variable induces a correlation between Y and M_Y, such that the missing data indicator carries information about the values of Y. The resulting NMAR mechanism would likely introduce bias because the imputation model could not generate accurate predictions without conditioning on M_Y (this is usually very difficult and requires special analytic procedures). Panels C and D depict the consequences of ignoring A_2 and A_3, respectively, during imputation. Omitting A_2 may decrease power because the imputation routine cannot leverage the information contained in its correlation with Y, but this variable is not a source of nonresponse bias because it does not predict missingness. Finally, although A_3 predicts nonresponse, ignoring this variable does not affect the mechanism because it is uncorrelated with the analysis variables, and thus there are no indirect pathways that can absorb the relation with M_Y.

Fig. 1 highlights that some analyses will satisfy MAR only if the missing data handling procedure incorporates auxiliary variables that are not part of the analysis plan. The figure also suggests that we should be most concerned with auxiliary variables that predict

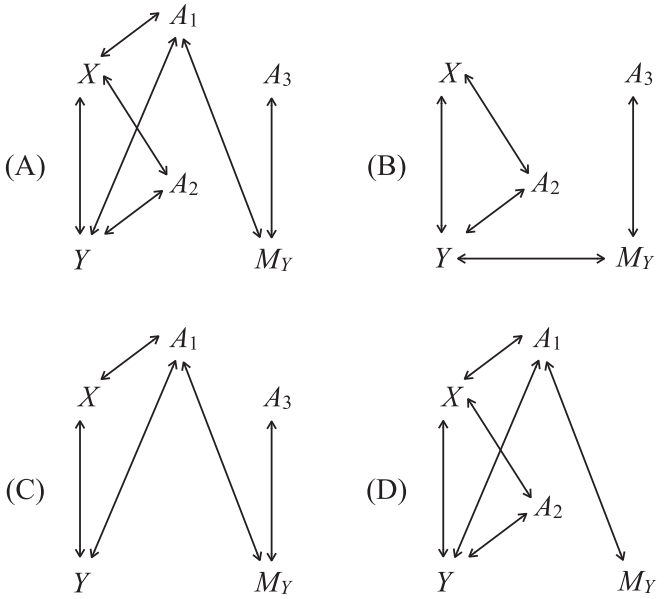

Fig. 1. Panel A depicts the true associations among six variables in a data set: X and Y are variables in the analysis, M_Y is the missing data indicator for Y, and A_1, A_2, and A_3 are potential auxiliary variables. In Panel A, the absence of an arrow connecting Y and M_Y indicates that MAR is satisfied after controlling for A_1 because the missing data indicator is unrelated to the incomplete variable. Panel B depicts an NMAR mechanism that results from ignoring (failing to condition on) A_1. Panels C and D show MAR mechanisms that result from ignoring A_2 and A_3, respectively.

nonresponse and are correlated with the analysis variables (e.g., A_1) because ignoring these variables fails to eliminate systematic differences between the complete and incomplete cases, thereby introducing bias. Notice that satisfying MAR parallels the logic of ANCOVA in a quasi-experimental design, where the goal is to introduce covariates that remove pre-existing differences between groups. In the context of missing data handling, the goal is to control for (condition on) auxiliary variables that differentiate the complete and incomplete cases. Simple bivariate correlations can help identify potential auxiliary variables, as can path analysis models (Raykov & West, 2015).

To illustrate the search for auxiliary variables, Table 2 gives correlations between the three missing data indicators from the regression analysis and other variables in the data set (graphically, these correlations align with the double-headed arrows in Fig. 1 that connect M_Y to X, A_1, A_2, and A_3). Positive correlations indicate that higher scores on a variable in the left-most column are associated with higher rates of missingness on one of the three analysis variables (e.g., older participants are more likely to have missing depression scores), and negative correlations indicate that lower scores predict nonresponse (e.g., lower levels of education are

Table 2
Missing data indicator.

Variable	Severity	DEP1	DEP3
TXGRP	−0.10	−0.03	−0.03
Male	0.10	0.02	0.07
Age	0.06	**0.24**	**0.27**
Educ	0.05	**−0.31**	−0.18
Exercise	−0.02	−0.07	−0.15
Interf	0.01	0.02	**0.35**
Severity	NA	0.02	**0.20**
DEP1	−0.01	NA	0.02
DEP2	−0.03	0.00	0.15
DEP3	0.01	0.06	NA

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

Table 3
Correlations among study variables.

Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. TXGRP	1.00									
2. Male	−0.04	1.00								
3. Age	0.09	0.23	1.00							
4. Educ	−0.09	0.01	0.04	1.00						
5. Exercise	0.00	−0.13	0.04	0.25	1.00					
6. Interf	0.00	0.04	0.03	−0.11	−0.34	1.00				
7. Severity	0.00	0.10	0.03	−0.14	−0.28	0.59	1.00			
8. DEP1	−0.04	0.13	−0.13	0.04	−0.23	0.36	0.29	1.00		
9. DEP2	−0.10	0.17	−0.05	0.11	−0.14	0.39	0.34	0.72	1.00	
10. DEP3	−0.21	0.16	0.01	0.13	−0.09	0.32	0.27	0.73	0.75	1.00

Note: Correlations based on maximum likelihood estimation with full sample.

associated with missingness). Although somewhat arbitrary, the table highlights correlations greater than ± 0.20, which is roughly in the middle of the small effect size range (Cohen, 1988). The bolded correlations with education and pain interference are particularly important because these two variables are not part of the regression analysis, and thus we may need to control for them if they are also correlated with the analysis variables.

Table 3 gives bivariate correlations among the study variables. Because education level has relatively weak correlations with the analysis variables (the strongest correlation is $r = -0.13$), it essentially functions like A_3 from Fig. 1, and thus we can safely ignore this variable during missing data handling (±0.40 is a rough rule of thumb for salient auxiliary variable correlations; Collins et al., 2001). However, like A_1 in the figure, pain interference with daily life is a useful auxiliary variable because of its higher correlations with the analysis variables (*r*'s in the range of 0.32–0.59). In addition to variables that predict nonresponse, we may also identify auxiliary variables that correlate with only the incomplete analysis variables, as these variables can improve power by increasing the precision of the imputations. The Wave 2 depression scale is an excellent candidate as an auxiliary variable (*r*'s with Wave 1 and Wave 3 depression exceed 0.70), but its utility may be diminished somewhat by missing data (e.g., only 16 cases with missing Wave 1 scores have Wave 2 scores to borrow from, and 57 cases with missing Wave 3 scores have data at Wave 2; Enders, 2008).

Collectively, the correlations in Tables 2 and 3 suggest that pain interference and Wave 2 depression scores are important auxiliary variables; controlling for the former mitigates nonresponse bias by improving the chances of satisfying the MAR assumption, and conditioning on the latter can improve power by borrowing information from a strong correlation. As we will see in the next section, multiple imputation readily handles auxiliary variables with no additional effort – these additional variables are simply included in the imputation routine along with the variables from the analysis model. This is a distinct advantage over maximum likelihood estimation, which requires structural equation modeling software and a rather cumbersome model setup (Graham, 2003).

4. Multiple imputation

Multiple imputation consists of three phases: an imputation phase, analysis phase, and pooling phase. The imputation phase creates multiple copies of the data (e.g., 20 or more is a current rule of thumb; Graham, Olchowski, & Gilreath, 2007), each with different imputed values. The basic idea behind imputation is to use a regression model to define a distribution of plausible replacement values for each case, then use computer simulation to "draw" a value at random from this distribution. For example, if we assume a normal distribution for an incomplete variable, each imputation is drawn a normal curve with a predicted value and residual variance defining the mean and spread, respectively. More formally, the distribution of replacement values for each case can be written as

$$Y_{(mis)} \sim N\left(\widehat{Y}, \sigma_{\varepsilon}^2\right) \quad (2)$$

where $Y_{(\text{mis})}$ is an estimate of the missing value, ~*N* denotes the normal distribution, $\widehat{Y}$ is the predicted value from a regression equation, and σ_{ε}^2 is the residual variance from the regression. Conceptually, the imputations from Equation (2) can be viewed as the sum of a predicted score and a random noise term (deviation score), the variance of which equals σ_{ε}^2 (i.e., $Y_{(\text{mis})} = \widehat{Y} + \varepsilon$).

The imputation phase typically uses a two-step iterative algorithm that creates imputations and then applies Bayesian estimation methods to the filled-in data to generate a new set of regression parameters for the next round of imputation. The updating process for regression parameters mimics the imputation step in the sense that new estimates are sampled from a distribution of plausible values. For example, the procedure computes ordinary least squares estimates and standard errors from the filled-in data, then it uses these quantities to define the mean and standard deviation, respectively, of a normal distribution, from which it draws new coefficients. These updated parameter values carry forward to the next round of imputation, where they are used to create a different set of imputations. Repeating these two steps many times accounts for missing data uncertainty by producing imputations from a range of plausible regression parameters. As such, no two sets of imputations will be exactly alike – in fact, they may be quite different.

After generating the desired number of imputations, the researcher then performs one or more statistical analyses on each complete data set to obtain imputation-specific estimates and standard errors. For example, I later illustrate an analysis that fits the regression model from Equation (1) to 20 imputed data sets. With careful planning, a single collection of imputed data sets can support a variety of different analyses, but this step should avoid variables or effects that were not part of the imputation procedure (e.g., if the imputation model included only zero-order relations, then the analysis phase should not consider interactive effects). The analysis phase produces parameter estimates and standard errors for each data set, and the researcher subsequently aggregates these quantities into a single set of results in the pooling phase. A single collection of point estimates is obtained by taking the arithmetic average of the imputation-specific estimates, and standard errors are combined in a similar fashion. The ultimate product of the pooling phase is a single set of estimates and standard errors that have the same interpretation and meaning as those from a complete-data analysis. Although the analysis and pooling phases sound tedious, most popular software packages have facilities that automate the procedures.

Thus far I have been somewhat vague about the composition of the regression model that generates the distribution of missing values in Equation (2). The classic incarnations of multiple imputation (Rubin, 1987; Schafer, 1997) use a multivariate regression model where incomplete variables are outcomes and complete variables are predictors. A distinguishing feature of this so-called joint model approach is that all incomplete variables are imputed in a single computational step. A second approach to imputation – termed chained equations imputation or fully conditional specification – uses a series or univariate regression models to generate imputations (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001; van Buuren, 2012; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). In this framework, variables are imputed in a round robin fashion, where each incomplete

variable is an outcome at one step and a predictor at all other steps. Fully conditional specification is ideal for mixtures of categorical and continuous variables because each imputation step can be tailored to the incomplete variable's metric (van Buuren, 2007). The joint model framework can also accommodate certain combinations of variable metrics (Asparouhov and Muthen, 2010; Carpenter & Kenward, 2013; Schafer, 1997; Schafer & Olsen, 1998), but I focus on fully conditional specification for the remainder of the paper because it is relatively easy to understand, and it is implemented in the Blimp application.

5. Analysis example 1

In this section I use the regression model from Equation (1) to illustrate a basic multiple imputation analysis. So that readers may recreate the analysis examples with freely available software, I use the Blimp application for imputation and R for the analysis and pooling phases. The imputation software, raw data file, and analysis scripts are available for download at www.appliedmissingdata.com/multilevel-imputation.html, and the analysis scripts are also found in Appendices A and B. Note that the Blimp application also features a graphical interface for Mac and Windows that allows researchers to specify all features of imputation without any coding. Technical details about algorithmic approach implemented in Blimp are available in Enders et al. (2016).

At a minimum, the imputation procedure should include all variables in the analysis model, but it can (and perhaps should) include auxiliary variables that predict missingness and/or correlate with the analysis variables. We previously determined that pain interference with daily life activities and Wave 2 depression scores are good auxiliary variables, and so I implemented FCS imputation with seven variables: three waves of depression scales, age, pain severity, pain interference, and the treatment indicator. For now, I treat the incomplete variables as though they are normally distributed, which is not necessarily optimal for the 7-point rating scale (the resulting imputations will have decimals). I return to this issue later where I illustrate imputation for mixtures of categorical and continuous variables. As an aside, estimating the regression model with maximum likelihood would treat all variables as multivariate normal, including the treatment indicator. Although I temporarily violate distributional assumptions for the incomplete severity ratings, fully conditional specification does not require distributional assumptions for complete predictors.

The imputation algorithm for this problem cycles through the incomplete variables one at a time, using a series of regression models to define distributions of plausible replacement values. A single iteration of the algorithm draws replacement values from the following sequence of univariate normal distributions.

In words, each equation says that the missing values (denoted with a "mis" subscript) are drawn at random from a normal distribution, the mean and variance of which are determined by the predicted value in square brackets and residual variance from a regression model, respectively. To avoid complicating the notation, I use the same symbols for all equations, but it is important to note that the numeric values differ (e.g., each equation has a unique value of β_0 and other model parameters). The vertical dots between the equations (i.e., ⋮) represent Bayesian estimation steps that generate the new regression parameters for the next round of imputation (Enders, 2010; Gelman et al., 2014; Sinharay, Stern, & Russell, 2001; van Buuren, 2012).

Notice that the right side of each equation includes complete and previously imputed variables (denoted by the "imp" subscript), such that the target of imputation at one step functions as a complete predictor at all other steps. For example, an individual's predicted DEP_1 score at the first step depends on the complete variables (age, pain interference, and treatment group membership) and on previously imputed variables (Wave 2 and 3 depression scores, pain severity). After this variable is imputed and its regression parameters are updated, it moves to the right side of all other equations where it functions as a complete predictor. Imputing variables in a round robin fashion ensures that the imputed values preserve all possible zero-order relations among the variables. Because the intervention effect is the primary interest in the subsequent analysis phase, it is important to emphasize that imputation preserves any group differences that may exist on the incomplete variables (it does not create differences that do not already exist, however). For example, each β_0 term in Equation (3) can be viewed as the adjusted mean for the control group, and the β_3 coefficients allow the treatment group means to differ. The imputation models do assume that the treatment and control groups share a common variance-covariance matrix, however (i.e., imputation does not model interactive effects).

To avoid ending up with imputations that are too similar, the computational steps from Equation (3) are often repeated for hundreds or thousands of cycles, with imputed data sets saved at specified intervals in the process. For example, the program in Appendix A generates 20 sets of imputations by saving a data set after every 200th computational cycle (the BURN command specifies the number of initial computational cycles, and the THIN command specifies the number of iterations between data sets). It is difficult to offer good rules of thumb for this aspect of imputation because the appropriate interval depends on data-specific features (e.g., the amount of missing data, magnitude of the correlations, number of variables), but software packages generally provide graphical or numeric diagnostics that facilitate this decision (Gelman & Rubin, 1992; Schafer & Olsen, 1998; Schafer, 1997). I chose an interval of 200 based on values of the potential scale reduction factor (PSR; Gelman & Rubin, 1992), a numeric diagnostic

$$
\begin{aligned}
DEP_{1(\text{mis})} &\sim N\Big(\Big[\beta_0 + \beta_1 AGE + \beta_2 INTERF + \beta_3 TXGRP + \beta_4 DEP_{2(\text{imp})} + \beta_5 DEP_{3(\text{imp})} + \beta_6 SEVERITY_{(\text{imp})}\Big], \sigma_\varepsilon^2\Big)\\
&\vdots\\
DEP_{2(\text{mis})} &\sim N\Big(\Big[\beta_0 + \beta_1 AGE + \beta_2 INTERF + \beta_3 TXGRP + \beta_4 DEP_{3(\text{imp})} + \beta_5 SEVERITY_{(\text{imp})} + \beta_6 DEP_{1(\text{imp})}\Big], \sigma_\varepsilon^2\Big)\\
&\vdots\\
DEP_{3(\text{mis})} &\sim N\Big(\Big[\beta_0 + \beta_1 AGE + \beta_2 INTERF + \beta_3 TXGRP + \beta_4 SEVERITY_{(\text{imp})} + \beta_5 DEP_{1(\text{imp})} + \beta_6 DEP_{2(\text{imp})}\Big], \sigma_\varepsilon^2\Big)\\
&\vdots\\
SEVERITY_{(\text{mis})} &\sim N\Big(\Big[\beta_0 + \beta_1 AGE + \beta_2 INTERF + \beta_3 TXGRP + \beta_4 DEP_{1(\text{imp})} + \beta_5 DEP_{2(\text{imp})} + \beta_6 DEP_{3(\text{imp})}\Big], \sigma_\varepsilon^2\Big)\\
&\vdots
\end{aligned}
\tag{3}
$$

based on stability of the regression parameters across iterations. The Blimp application produces PSR tables when the "psr" keyword is listed on the OPTIONS line of the syntax, as it is in the appendix (or activated via a radio button in the graphical interface).

After generating imputations, one or more statistical analyses are performed on each filled-in data set. It is important to reiterate that the analyses should be restricted to variables or effects from the imputation model, as any variables not included in imputation are uncorrelated with the filled-in values (e.g., although gender is complete, adding it as a covariate in the analysis model is inappropriate because this variable did not contribute to the imputations). To illustrate the analysis and pooling phases, I used the MITML package (Grund, Robitzsch, & Lü;dke, 2016) in R to fit the regression model to each data set and combine the resulting estimates and standard errors. Table 4 gives the estimates and standard errors from the first three data sets. Notice that the estimates vary across data sets (e.g., the intervention slope ranges between −1.50 and −1.98), which is a natural consequence of analyzing different imputations. This so-called between-imputation variation is an important component of the analysis that captures uncertainty due to missing data.

Table 5 gives the pooled estimates and standard errors from the regression analysis. As noted previously, the pooled point estimates are simply arithmetic averages of the imputation-specific values (Little & Rubin, 2002; Rubin, 1987). However, notice that the pooled standard errors from Table 5 are markedly larger than those of the individual data sets in Table 4. Because the imputation-specific standard errors in Table 4 derive from complete data sets, their arithmetic averages are too small because they fail to account for missing data uncertainty. Rubin's pooling formula incorporates the average standard error from the imputed data sets, but it also incorporates a correction factor that inflates the standard error to compensate for imputation noise. This correction term is based on the variability of the estimates across the imputed data sets (e.g., the variation in the intervention slope estimates that I noted previously), and its magnitude depends on the amount of missing data and the correlations among the variables. Ultimately, the estimates and standard errors from Table 5 are interpreted in the same fashion as those from a complete-data analysis. For example, the treatment group slope, $\beta_4 = -1.87$, $SE = 0.46$, $p < 0.001$, indicates that the intervention group mean was 1.87 points lower than that of the control group, controlling for the covariates. Enders (2010, Ch. 11) gives recommendations for reporting the results from a multiple imputation analysis.

Table 4
Parameter estimates from three imputed data sets.

Variable	Imputation 1		Imputation 2		Imputation 3	
	Est.	*SE*	Est.	*SE*	Est.	*SE*
Intercept	2.244	1.196	2.368	1.239	0.928	1.205
DEP1	0.661	0.037	0.705	0.038	0.658	0.037
Age	0.059	0.017	0.059	0.017	0.075	0.016
Severity	0.294	0.146	0.023	0.149	0.407	0.147
TXGRP	−1.982	0.380	−1.498	0.384	−1.820	0.375

Table 5
Pooled estimates and standard errors from analysis example 1.

Variable	Est.	*SE*	*t*	*df*	*p*
Intercept	2.119	1.632	1.299	56.711	0.100
DEP1	0.660	0.047	14.136	74.721	<0.001
Age	0.062	0.024	2.591	45.596	0.006
Severity	0.287	0.191	1.505	66.434	0.069
TXGRP	−1.871	0.459	−4.079	92.360	<0.001

6. Imputing mixtures of categorical and continuous variables

The previous analysis treated pain severity ratings as normally distributed, which had the effect of generating fractional rather than discrete imputations (e.g., an imputed value of 4.32). Rounding the fractional imputes to the nearest integer is one way to handle this issue, but methodology studies suggest that post-imputation rounding can introduce serious biases (Allison, 2005; Horton, Lipsitz, & Parzen, 2003). Although rounding can provide acceptable results in some situations (Bernaards, Belin, & Schafer, 2007; Carpenter & Kenward, 2013), advances in categorical imputation have largely negated the need to apply such procedures. Categorical imputation routines are now widely available in software packages, although the capabilities of these programs vary (e.g., some programs are limited to ordinal or nominal variables, fewer packages can handle both). The Blimp application can accommodate mixtures of nominal, ordinal, and continuous (normal) variables, both in single-level and multilevel analyses. As noted previously, tailoring the missing data handling routine to each variable's scale is difficult and usually impossible in the context of maximum likelihood estimation, so this flexibility is a major advantage to multiple imputation.

Logistic and probit regression are the principal frameworks for categorical imputation. Like the linear regression methods described previously, discrete imputes are a function of a deterministic component (i.e., predicted value) and random stochastic noise. To illustrate the basic idea, consider an incomplete binary variable coded as zero or one (e.g., 0 = subclinical depression, 1 = clinical depression). Logistic regression expresses the log odds (logit) of the outcome (e.g., a clinical diagnosis) as a linear function of a set of predictors. Substituting predictor variable scores into a logistic regression equation yields the predicted log odds, and a simple transformation converts the predicted logit to a predicted probability (Agresti, 2012). In the context of imputation, the predicted probability defines the distribution of plausible replacement values for each case. For example, suppose that a logistic model estimated the predicted probability of clinical depression at 0.85 for a particular individual. The imputation step would then draw a replacement value at random from a two-category distribution with group proportions of 0.15 and 0.85. Like linear regression imputation, the distribution of replacement values varies across individuals, such that the shape of the distribution (the category proportions) depends on a predicted value. Several popular software programs offer logistic regression imputation, including SAS, SPSS, Stata, and the MICE package in R, to name a few. These software packages differ in the types of data they can address; some packages implement multinomial logistic imputation for nominal outcomes (e.g., SPSS), others use a so-called proportional odds model for ordinal variables (e.g., SAS), and others offer both options (e.g., Stata and MICE).

Probit regression is a second option for categorical imputation. This approach is sometimes referred to as latent variable imputation (Carpenter & Kenward, 2013; Enders, Mistler, & Keller, 2016; Quartagno & Carpenter, 2016) because it views discrete responses as arising from an underlying normal latent variable distribution (or distributions of latent variable difference scores, in the case of nominal variables with more than two categories).[3] For example,

[3] The logistic regression model can also be represented with an underlying normal latent variable, albeit with a different variance structure. However, I refer to probit regression as a latent variable approach because this is common in the literature, and because Bayesian estimation for the probit model works directly with the underlying latent variable scores (Bayesian estimation for logistic regression does not).

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

reconsidering the binary depression diagnosis, the latent variable approach defines an underlying normal variable that represents an individual's propensity for clinical depression (for identification purposes, the underlying normal variable is usually scaled as a standard normal *z*-score). Further, a single threshold parameter divides the distribution into two discrete outcomes, such that latent scores below the threshold correspond to a value of zero (i.e., subclinical depression), and scores above the threshold correspond to a one (i.e., clinical depression). The mechanics of probit-based imputation are quite similar to those of linear regression imputation. The Bayesian estimation routine first replaces discrete responses with latent variable scores, after which it estimates the regression of the latent variable scores on a predictor set. Imputations are then generated by drawing latent variable scores from a standard normal distribution, and the latent values are subsequently converted to discrete imputes (e.g., in the case of a binary or ordinal variable, by comparing the location of the imputed *z*-scores relative to the threshold parameters). Latent variable imputation is available in MLwiN, Mplus, Blimp, and the JOMO package for R. I illustrate categorical imputation with Blimp later in the manuscript.

7. Questionnaires with item-level missing data

Recall from Table 1 that the depression data is characterized by intermittent missingness where participants skipped questionnaire items but otherwise participated and attrition where participants are missing one or two waves of data. Perhaps the most common approach to handling the item-level missingness is to compute a scale score by averaging the available items (e.g., if a participant answered three items, the scale score is defined as the mean of those items). This approach, which is commonly referred to as "proration" in the applied literature, requires the restrictive MCAR mechanism (i.e., no systematic sources of missingness), but it further assumes that the item means and inter-item correlations are identical (e.g., all items have a mean of 3.0 and all inter-item correlations are 0.40). Proration can introduce substantial biases when either of these conditions is violated (Mazza et al., 2015) — one or both usually are — and thus should be avoided. Multiple imputation is almost always a superior approach to handling item-level missing data.

The previous analysis example treated the depression scale scores as missing when one or more of the component items was missing, and it applied a scale-level imputation procedure that ignores any complete item responses. This approach is not ideal because it fails to leverage the strong correlations between the items and scale scores, and thus can result in a substantial loss of power relative to a procedure that imputes the items (Gottschall et al., 2012; Mazza et al., 2015). In fact, the power gain from imputing items rather than scales can be equivalent to increasing the sample size by 50% or more (Gottschall et al., 2012). Not only does item-level imputation provide a dramatic boost in precision, but the procedure is very easy to implement: apply a categorical imputation routine to the incomplete items, compute scale scores from the filled-in item responses, and analyze the scale scores. As an aside, maximum likelihood estimation is naturally suited for scale-level missing data handling (i.e., treating the scale as missing when one or more items are missing), and thus it suffers from the same limitations as scale-level imputation. Incorporating item-level data into a maximum likelihood analysis requires the researcher to specify a complex auxiliary variable model (Eekhout et al., 2015; Mazza et al., 2015) or recast the scale scores as latent variables with items as indicators.

Table 6
Pooled estimates and standard errors from analysis example 2.

Variable	Est.	*SE*	*t*	*df*	*p*
Intercept	1.245	1.410	0.883	138.062	0.189
DEP1	0.679	0.041	16.384	181.605	<0.001
Age	0.069	0.020	3.497	128.080	<0.001
Severity	0.307	0.173	1.773	135.973	0.039
TXGRP	−1.640	0.427	−3.842	181.090	<0.001

8. Analysis example 2

In this section I reanalyze the regression model from Equation (1) after applying a categorical imputation routine to pain severity ratings and the incomplete questionnaire items. The methodological literature convincingly demonstrates the benefits of item-level imputation, but it does not offer a clear prescription for dealing with the combination of intermittent missingness and attrition. For example, at Wave 2, 41 cases have one or two missing items, and 26 individuals have no questionnaire data. At Wave 3, attrition is the largest source of missingness, with 23 participants missing three or fewer items, and 71 cases missing the entire wave. Item-level imputation improves power because it borrows information from highly correlated items measured at the same wave, but attrition forces the procedure to impute the entire Wave 3 questionnaire from weaker correlations with Wave 1 and Wave 2 items. In some situations, it may be useful to reduce the number of variables in the imputation model by using a combination of item- and scale-level imputation. For example, I could impute the Wave 1 items and the Wave 3 scale score, perhaps using a subset of the Wave 3 items as auxiliary variables.

Appendices C and D give the imputation and analysis scripts from Blimp and R, respectively. Because the benefit of item-imputation is so substantial and because the number of questionnaire items to be imputed is relatively small, I applied item-level imputation to the Wave 1 and Wave 3 questionnaire data, and I used the three Wave 2 items with the lowest missingness rates as auxiliary variables. Listing the questionnaire items on the ORDINAL command line triggers categorical imputation based on the latent variable formulation. Although there were no incomplete nominal variables, the imputation script shows the NOMINAL command line so that readers are aware of this option. The NOMINAL line identifies incomplete nominal variables, but it also converts complete nominal variables to dummy codes for use in the imputation regression model. Finally, based on convergence diagnostics, I instructed the iterative algorithm to save a data set after every 1000 computational cycles. A larger interval is usually necessary when imputing ordinal variables because the threshold parameters that slice the latent variable distributions into segments are relatively unstable across iterations. Table 6 gives the pooled estimates and standard errors from the regression analysis. As before, these quantities are interpreted in the same fashion as those from a complete-data analysis (e.g., $\beta_4 = -1.64$ indicates that the intervention group mean was 1.64 points lower than that of the control group, controlling for the covariates).

9. Significance testing and model fit

Although multiple imputation offers no particular advantages to maximum likelihood estimation when it comes to significance testing — in fact, there is some evidence to suggest that imputation-based significance tests may require a slightly larger sample size to reach optimal performance (Enders & Mansolf, in press) — this important topic warrants a brief discussion, as the development and evaluation of multiple imputation test statistics is an active area of methodological research (Enders & Mansolf, 2016; Liu & Enders, 2016; Grund, Lüdtke, and Robitzsch, 2016; Licht, 2010).

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

Returning to the regression analysis from the previous section, the primary interest is whether the intervention influences the outcome above and beyond the covariates. Evaluating a single estimate is done using the familiar t (or z) ratio.

$$t = \frac{\left(\widehat{\theta} - \theta_0\right)}{SE} \tag{4}$$

where $\widehat{\theta}$ is the multiple imputation point estimate (the arithmetic average of the imputation-specific estimates), θ_0 is the hypothesized value (typically zero), and the denominator is the pooled standard error. The test statistics in Tables 5 and 6 uses a small-sample adjustment based on a fractional degrees of freedom value (Barnard & Rubin, 1999), but the test statistic can also be referred to a standard normal z distribution (Asparouhov and Muthen, 2010). Small-sample adjustments to the t statistic appear to offer better Type I error control in very small samples (Barnard & Rubin, 1999; Reiter, 2007), but the choice of reference distribution probably makes little difference in most cases (Liu & Enders, 2016).

Multiparameter significance tests can be obtained using either a Wald or likelihood ratio statistic. When applied to a single parameter, the Wald test is just the square of the statistic from Equation (4).

$$W = \frac{\left(\widehat{\theta} - \theta_0\right)^2}{SE^2} \tag{5}$$

However, the Wald formulation above readily extends to more than one parameter (the multiparameter variant is often referred to as D_1 in the literature; Enders, 2010; Schafer, 1997). For example, consider an omnibus test that evaluates the four regression slopes from Equation (1). In this case, $\widehat{\theta}$ is a vector containing the four estimates, θ_0 is a corresponding vector of zeros, and the denominator is a 4 by 4 matrix (a submatrix from the variance-covariance matrix of the estimates), the diagonal of which contains squared standard errors.[4] The R syntax in Appendix D illustrates the Wald test, the value of which indicates that the set of predictors explains variation in Wave 3 depression scores, $F(4, 246.58) = 80.13$, $p < 0.001$. As an aside, the corresponding test statistic from the first analysis example was substantially lower, $F(4, 172.89) = 60.87$, $p < 0.001$, presumably due to the loss of power that results from scale-level imputation. Finally, note that either an F or chi-square distribution can generate the p-value for the Wald test (Asparouhov and Muthen, 2010; Li, Raghunathan, & Rubin, 1991).

Whereas the Wald test requires only one analysis, the likelihood ratio statistic (sometimes referred to as D_3 in the literature; Enders, 2010; Schafer, 1997) compares the relative fit of two models. The first is a general model that includes all parameters of interest, and the second is a nested model that imposes constraints on a subset of parameters. Returning to the omnibus test example, the analysis model from Equation (1) is the general model, and the nested model constraints the four regression slopes to zero during estimation. The R syntax in Appendix D also illustrates the likelihood ratio statistic, the value of which is $F(4, 1047.16) = 48.06$, $p < 0.001$. The likelihood ratio statistic is generally more complicated to implement because it requires multiple passes through the data (the first pass computes the average likelihood ratio test from the imputed data sets, and the second computes the average a second time while constraining all parameters to their pooled values), but its derivation implies equivalence to the Wald test, at least in large samples (Meng & Rubin, 1992).

Structural equation modeling is an exceedingly popular analytic tool that warrants a brief discussion. The Wald statistic for multiply imputed data is applicable to a wide variety of multiparameter significance tests (e.g., testing whether a group of path coefficients is different from zero; testing between-group equality constraints in multiple group invariance models), but it does not function as a global test of model fit. The familiar test of model fit for complete data is a likelihood ratio statistic comparing the relative fit of the researcher's model (the nested model) to that of a best-fitting saturated model (the general model). The multiple imputation variant of the likelihood ratio statistic outlined by Meng and Rubin (1992) can serve the same purpose with missing data, and it is currently available in some software packages (e.g., Mplus and the lavaan package in R; Asparouhov and Muthen, 2010; Contributors, 2014). The scant research on this topic suggests that, relative to its maximum likelihood counterpart, the multiple imputation test statistic may require somewhat larger sample sizes to reach its optimal performance (e.g., $N > 200$ in relatively simple models), and it may have less power to reject a false model (Enders & Mansolf, in press). Importantly, the chi-square version of Meng and Rubin's (1992) test can be used to construct imputation-based fit indices such as the CFI, TLI, and RMSEA (Enders & Mansolf, in press) by substituting the pooled statistic into standard complete-data expressions for the fit measures. Limited research suggests that these measures are comparable to those of maximum likelihood.

10. Interaction effects

Interaction (moderation) effects are ubiquitous throughout psychology and pose interesting methodological challenges for missing data handling. As its currently implemented in popular software packages, multiple imputation holds no particular advantage over maximum likelihood estimation, but newly developed imputation routines that are not yet implemented in mainstream software programs appear to offer a substantial improvement (Bartlett, Seaman, White, & Carpenter, 2015). Interactive effects are important to consider here because they generally require specialized imputation procedures. To motivate the ensuing discussion, I expand the regression analysis from Equation (1) to include a product term where pain severity moderates the effect of the intervention on depression scores.

$$DEP_3 = \beta_0 + \beta_1(DEP_1) + \beta_2(AGE) + \beta_3(SEVERITY) + \beta_4(TXGRP) + \beta_5(SEVERITY)(TXGRP) + \varepsilon \tag{6}$$

Recall that pain severity ratings are incomplete, which means that the product term is as well. At first glance, it may seem reasonable to simply compute the interaction as the product of the treatment indicator and the imputed severity scores. However, this so-called impute-then-transform approach (von Hippel, 2009) pushes the product term coefficient toward zero because the imputation phase fails to model the interactive effect (i.e., an imputation model based solely on the lower-order terms explicitly assumes that β_5 equals zero).

When one component of the product is categorical and complete – as it is here – the best strategy is to perform imputation within subgroups defined by the categorical variable (Enders & Gottschall, 2011). The procedure requires the following steps: (a) create separate data sets for the intervention and control groups, (b) generate imputations for each group (the intervention dummy

[4] Technically, the multivariate version of the Wald test does not use a matrix in the denominator, as there is no division operator in matrix algebra. Rather, the test statistic employs an inverse, which is the matrix analog of a reciprocal from scalar algebra.

code cannot be included because it is constant within group), (c) merge the subgroup data sets, (d) compute the product term by multiplying the dummy code and the imputed severity ratings, and (e) estimate the regression model from Equation (6). Notice that the product term is not an explicit part of the imputation model and must be computed post-imputation. Although this step resembles the problematic impute-then-transform approach, separate-group imputation allows the covariance structure to differ across groups, thereby preserving all possible two-way interactions between the grouping variable and the imputation model variables (Enders & Gottschall, 2011).

Separate-group imputation is limited in practice because it requires a categorical predictor with no missing values, and group sizes must be sufficiently large to support imputation (Graham, 2009) – at a minimum, all variables in a group's imputation model should have more observations than the number of variables in the model (Asparouhov and Muthen, 2010). A second option is to include the incomplete product term in the imputation model, treating it just like any other incomplete variable (Enders, Baraldi, & Cham, 2014; van Buuren, 2012; von Hippel, 2009). Unfortunately, this so-called transform-then-impute strategy (von Hippel, 2009) generally requires an MCAR mechanism and produces biased estimates when data are systematically missing (Carpenter & Kenward, 2013; Enders et al., 2014; Yuan & Savalei, 2014). Fortunately, promising new imputation procedures are currently under development and will likely become available in software packages in the near future (Bartlett et al., 2015). Until then, product term imputation may be the only viable option for preserving interaction effects with incomplete data. As an aside, maximum likelihood estimation suffers from the same limitations as product term imputation and thus does not provide a viable alternative.

Product term imputation has two noteworthy features. First, because the procedure treats the product as a distinct variable with unique moments, the resulting imputations do not equal the product of their component parts. It is tempting to remedy this issues by recomputing the product following imputation, but this too introduces bias and should be avoided (von Hippel, 2009). Second, the common practice of centering lower-order variables prior to computing their product (Aiken & West, 1991) is not applicable to multiply imputed data because the means are unknown. Consequently, the imputation and analysis stages must use raw score variables. Conditional (lower-order) effects that are consistent with a centered solution can then be obtained via algebraic transformation (Aiken & West, 1991; Bauer & Curran, 2005; Hayes & Matthes, 2009), or by centering the lower-order terms *and* the product variable post-imputation (Enders et al., 2014). It is important to emphasize that centering a product term is not as simple as subtracting its average because the product of two deviation score variables generally has a non-zero mean. Enders et al. (2014) provides computer code for product term imputation that illustrates the transformation and post-imputation centering methods.

11. Multilevel data

A great deal of recent methodological work has focused on missing data handling methods for multilevel data structures (Drechsler, 2015; Enders et al., 2016; Enders et al., 2016; Goldstein, Bonnet, & Rocher, 2007; Goldstein, Carpenter, Kenward, & Levin, 2009; Grund, Lü;dke, & Robitzsch, 2016; Schafer & Yucel, 2002; Shin & Raudenbush, 2007; Yucel, 2008, 2011). This work has important implications for psychological research because nested data structures are exceedingly common throughout our discipline (e.g., students nested within schools; repeated measures nested within individuals; individuals nested within dyads or families). Maximum likelihood estimation for multilevel models with incomplete data is often limited because software packages may restrict missing data handling to outcome variables, and programs that can handle incomplete predictors typically limit this functionality to random intercept models (Asparouhov and Muthen, 2010; Shin & Raudenbush, 2007, 2010). Multiple imputation, on the other hand, is well suited for a variety of multilevel analysis problems because it makes no distinction about a variable's role in the subsequent analysis model. The Blimp application was developed to handle a wide variety of common multilevel analysis problems, including models with random slopes, categorical variables, and up to three levels (Enders et al., 2016). Other software packages offer a more limited set of capabilities; some programs are restricted to random intercept analyses, others can impute only level-1 variables, while others are restricted to normal variables (Enders et al., 2016).

Single-level imputation procedures are inappropriate for multilevel data structures because they ignore between-cluster sources of variation (Enders et al., 2016; Reiter, Raghunathan, & Kinney, 2006; van Buuren, 2011). One method for addressing this problem is to code cluster membership with a set of dummy variables and include the codes as predictors in a single-level imputation scheme. This so-called fixed effect imputation strategy tends to distort standard errors in some cases (Andridge, 2011; Reiter et al., 2006; van Buuren, 2011) and can produce biased estimates when the intraclass correlation is low (Drechsler, 2015). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. From a practical perspective, fixed effect imputation is usually not an ideal option because it is limited to random intercept analyses, and it cumbersome to implement (Enders et al., 2016).

Earlier in the paper I established the idea that imputations are effectively the sum of a predicted value a random residual term. Multilevel imputation schemes apply the same logic, but the predicted values include one or more residual terms that capture between-cluster differences in the intercepts and possibly the slopes. To illustrate, consider a simple bivariate random intercept analysis.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij} \tag{7}$$

where Y_{ij} and X_{ij} are scores for observation i within cluster j, u_j is a between-cluster (i.e., level-2) residual that captures residual mean differences not explained by X, and ε_{ij} is a within-cluster (i.e., level-1) residual. Further, assume that X is missing and Y is complete. Like its single-level counterpart, multilevel imputation draws missing values from a normal distribution with a mean and variance equal to the predicted score and within-cluster residual variance, respectively.

$$X_{ij(mis)} \sim N\left(\left[\beta_0 + \beta_1 Y_{ij} + u_j\right], \sigma^2_\varepsilon\right) \tag{8}$$

Importantly, the predicted score enclosed in square brackets includes a level-2 residual term u_j that preserves between-cluster (i.e., random intercept) variation X. With more complicated analysis models, the predicted score can expand to include cluster means and random slope residuals, among other things (Carpenter & Kenward, 2013; Enders et al., 2016). It is worth noting that imputation for random slope models may not yield estimates that are bias free, particularly when predictor variables in the analysis model have missing data (Enders et al., 2016; Grund et al., 2016). Nevertheless, extending Equation (8) to include random slope residuals reflects the current implementation of multilevel FCS (Enders et al., 2016; Grund et al., 2016; van Buuren, 2011, 2012; van Buuren et al., 2014), although new approaches to this problem are

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

currently under development (Erler et al., 2016).

Methodological literature makes it abundantly clear that multilevel data sets require specialized missing data handling routines, longitudinal analyses warrant discussion because they do not necessarily require a multilevel imputation scheme. In fact, single-level imputation may be preferable in some cases. In the context of longitudinal data, it is widely known that the multilevel growth model is sometimes equivalent to a single-level latent curve model from structural equation modeling (Chou, Bentler, & Pentz, 1998; Hox & Stoel, 2005; Mehta & West, 2000). The equivalence (or lack thereof) of the two modeling frameworks depends on whether the time between assessments is constant across individuals. Returning to the online chronic pain management program, suppose that depression scores were obtained from all participants at exactly 7 and 14 weeks post-baseline. Because a common covariance matrix is applicable to the entire sample, multilevel and structural equation modeling growth curve analyses are equivalent. In contrast, suppose that logistical constraints forced the researchers to stagger data collection, such that the initial follow-up was collected in the 5- to 9-week interval post-baseline, and the second follow-up was obtained between 12 and 16 weeks. The individually-varying assessment schedules require a unique covariance matrix for each participant (e.g., because the correlation between baseline and a 5-week follow-up would be stronger than the correlation between baseline and a 9-week follow-up), thereby necessitating a multilevel analysis.

In situations where a single-level analysis is appropriate (i.e., time between measurements is the same for all participants), arranging the data in wide format (i.e., columns represent repeated measurements) and applying a single-level imputation scheme is advantageous because the researcher need not specify the functional form of the growth trajectory. Returning to Equation (3), the β_0 terms from the first three imputation steps capture mean levels of depression, adjusted for the predictors in each equation. Importantly, these means freely vary with no restrictions, thereby producing imputations that are perfectly valid for exploring linear or quadratic growth, and it makes no difference whether the subsequent analysis employs a multilevel or latent curve model. In contrast, person-specific assessment schedules require a stacked data format and multilevel imputation. The multilevel framework treats the passage of time as an explicit predictor variable, with nonlinear trajectories often specified as polynomial functions of time (e.g., a quadratic growth curve includes a time variable and its square). In a similar vein, multilevel imputation requires the researcher to explicitly specify the growth curve's functional form (e.g., if nonlinear change is expected, the imputation model must include polynomial terms for the time variable). Because single-level imputation requires less a priori knowledge on the part of the researcher, it is probably preferable in situations where it's applicable.

As a brief aside, some disciplines still rely on a longitudinal imputation approach known as last observation carried forward. As its name implies, this method uses the last observed value for a participant to impute missing values at future measurements (e.g., if a participant drops out after the second wave, her Wave 2 score serves as an imputation for all future waves). Conventional wisdom suggests that last observation carried forward yields conservative estimates of intervention effects in longitudinal studies because it assumes no change after dropout. However, methodological research has shown that this imputation scheme can either exaggerate or attenuate group differences, even when the mechanism is MCAR (Cook, Zeng, & Yi, 2004; Liu & Gould, 2002; Mallinckrodt, Clark, & David, 2001; Molenberghs et al., 2004). Consequently, last observation carried forward is not a viable alternative to multiple imputation in longitudinal data sets.

12. Discussion

The last 20 years has seen a substantial increase in missing data research, and most software applications now implement one or more modern missing data handling routines. Despite their widespread availability and superior statistical properties (e.g., less stringent assumptions about the cause of missing data, greater accuracy and power), the adoption of these modern analytic methods is not uniform throughout psychology and related disciplines. Consequently, the primary goal of this manuscript is to promote the awareness and application of analytic methods that enjoy a strong foundation of support in the methodological literature. Broadly speaking, the missing data literature supports the use of maximum likelihood estimation and multiple imputation (and Bayesian analyses). Maximum likelihood estimation is perhaps the easiest method to use in practice because researchers need only specify their analyses in a capable software package (e.g., any one of several structural equation modeling packages). However, psychological data sets often feature complexities that are currently difficult to handle appropriately in the likelihood framework (e.g., mixtures of categorical and continuous variables). Multiple imputation is often a better tool for behavioral science data because it gives researchers the flexibility to tailor the missing data handling procedure to match a particular set of analysis goals. Throughout the paper, I discussed a number of practical issues that clinical researchers are likely to encounter when applying multiple imputation, including mixtures of categorical and continuous variables, item-level missing data in questionnaires, significance testing, interaction effects, and multilevel missing data. Two analysis examples illustrated the application of multiple imputation in freely available software, in hopes of encouraging the adoption of this method in applied research.

Acknowledgements

This work was supported by Institute of Educational Sciences award R305D150056.

Appendix A

Blimp imputation syntax for Analysis Example 1.

```
data: /users/craig/desktop/example/painstudy.dat;
varnames: txgrp male age educ exercise interf severity
  t1dep1 t1dep2 t1dep3 t1dep4 t1dep5 t1dep6
  t2dep1 t2dep2 t2dep3 t2dep4 t2dep5 t2dep6
  t3dep1 t3dep2 t3dep3 t3dep4 t3dep5 t3dep6
  dep1 dep2 dep3;
missing: 999;
model: ~ txgrp age interf severity dep1 dep2 dep3;
seed: 90095;
burn: 200;
thin: 200;
nimps: 20;
chains: 2;
outfile: /users/craig/desktop/example/ex1imps.csv;
options: stacked psr;
```

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

Appendix B

R Analysis and Pooling Syntax for Analysis Example 1.

```
# load libraries
library(mitml)
# read imputations in stacked format
impdata <- read.csv(file = "~/desktop/example/ex1imps.csv",
  head = FALSE, sep = ",")
names(impdata) = c("imp", "txgrp", "male", "age",
  "educ", "exercise", "interf", "severity",
  "t1dep1", "t1dep2", "t1dep3", "t1dep4", "t1dep5", "t1dep6",
  "t2dep1", "t2dep2", "t2dep3", "t2dep4", "t2dep5", "t2dep6",
  "t3dep1", "t3dep2", "t3dep3", "t3dep4", "t3dep5", "t3dep6",
  "dep1", "dep2", "dep3")
# split stacked data into separate files
implist <- split(impdata, impdata$imp)
implist <- as.mitml.list(implist)
# estimate models
model <- with(implist, lm(dep3 ~ dep1 + age + severity + txgrp))
dfdenom <- 300 - 4 - 1
testEstimates(model, df.com = dfdenom)
```

Appendix C

Blimp imputation syntax for Analysis Example 1.

```
data: /users/craig/desktop/example/painstudy.dat;
varnames: txgrp male age educ exercise interf severity
  t1dep1 t1dep2 t1dep3 t1dep4 t1dep5 t1dep6
  t2dep1 t2dep2 t2dep3 t2dep4 t2dep5 t2dep6
  t3dep1 t3dep2 t3dep3 t3dep4 t3dep5 t3dep6
  dep1 dep2 dep3;
missing: 999;
ordinal: severity t1dep1 t1dep2 t1dep3 t1dep4 t1dep5 t1dep6
  t2dep3 t2dep4 t2dep5 t3dep1 t3dep2 t3dep3 t3dep4 t3dep5 t3dep6;
nominal: ;
model: ~ txgrp age interf severity t1dep1 t1dep2 t1dep3 t1dep4 t1dep5
t1dep6
  t2dep3 t2dep4 t2dep5 t3dep1 t3dep2 t3dep3 t3dep4 t3dep5 t3dep6;
seed: 90291;
burn: 1000;
thin: 1000;
nimps: 20;
chains: 2;
outfile: /users/craig/desktop/example/ex2imps.csv;
options: stacked psr;
```

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

Appendix D

R Analysis and Pooling Syntax for Analysis Example 1.

```
# load libraries
library(mitml)
# read imputations in stacked format
impdata <- read.csv(file = "~/desktop/example/ex2imps.csv",
  head = FALSE, sep = ",")
names(impdata) = c("imp", "txgrp", "male", "age",
  "educ", "exercise", "interf", "severity",
  "t1dep1", "t1dep2", "t1dep3", "t1dep4", "t1dep5", "t1dep6",
  "t2dep1", "t2dep2", "t2dep3", "t2dep4", "t2dep5", "t2dep6",
  "t3dep1", "t3dep2", "t3dep3", "t3dep4", "t3dep5", "t3dep6",
  "dep1", "dep2", "dep3")
# compute scales
dep1items <- c("t1dep1", "t1dep2", "t1dep3", "t1dep4", "t1dep5",
"t1dep6")
dep3items <- c("t3dep1", "t3dep2", "t3dep3", "t3dep4", "t3dep5",
"t3dep6")
impdata$dep1 <- rowSums(impdata[, dep1items])
impdata$dep3 <- rowSums(impdata[, dep3items])
# split stacked data into separate files
implist <- split(impdata, impdata$imp)
implist <- as.mitml.list(implist)
# estimate model
model <- with(implist, lm(dep3 ~ dep1 + age + severity + txgrp))
dfdenom <- 300 - 4 - 1
testEstimates(model, df.com = dfdenom)
# estimate restricted model and compute wald (d1) and likelihood ratio
(d3) tests
restricted <- with(implist, lm(dep3 ~ 1))
testModels(model, restricted, method = c("D1"), df.com = dfdenom)
testModels(model, restricted, method = c("D3"), df.com = NULL)
```

References

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Allison, P. D. (2002). *Missing data*. Newbury Park, CA: Sage.

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. In *Paper presented at the SAS users group international*.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom J, 53*(1), 57–74. http://dx.doi.org/10.1002/bimj.201000140.

Asparouhov, T., & Muthen, B. (2010). *Multiple imputation with Mplus*. Retrieved from: http://www.statmodel.com/download/Imputations7.pdf.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*, 948–955.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research, 24*, 462–487.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40*, 373–400.

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine, 26*, 1368–1382.

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. West Sussex, UK: Wiley.

Chou, C. P., Bentler, P. M., & Pentz, M. A. (1998). Comparison of two statistical approaches to study growth curves. *Structural Equation Modeling: A Multidisciplinary Journal, 5*, 247–266.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330–351.

Contributors, s (2014). *semTools: Useful tools for structural equation modeling. R package version 0.4-9*. Retrieved from: http://cran.r-project.org/web/packages/semTools/index.html.

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on LOCF imputation. *Biometrics, 60*, 820–828.

Drechsler, J. (2015). Multiple imputation of multilevel missing data–rigor versus

simplicity. *Journal of Educational and Behavioral Statistics, 40*(1), 69–95. http://dx.doi.org/10.3102/1076998614563393.

Eekhout, I., Enders, C. K., Twisk, J. W. R., de Boer, M. R., de Vet, H. C. W., et al. (2015). Analyzing incomplete item scores in longitudinal data by Including item score information as auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 1–15. http://dx.doi.org/10.1080/10705511.2014.937670.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 434–448.

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

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

Enders, C. K. (Ed.). (2013). *Analyzing structural equation models with missing data* (2nd ed.). Greenwich, CT: Information Age.

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

Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(1), 35–54. http://dx.doi.org/10.1080/10705511.2011.532695.

Enders, C. K., Keller, B. T., & Levy, R. (2016). *A fully conditional specification approach to multilevel imputation of categorical and continuous variables* (Manuscript submitted for publication).

Enders, C. K., & Mansolf, M. (2016). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods* (in press).

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*, 222–240.

Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W. V., Franco, O. H., & Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine, 35*, 2955–2974.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics, 32*(3), 252–286. http://dx.doi.org/10.3102/1076998606298042.

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling, 9*(3), 173–197. http://dx.doi.org/10.1177/1471082x0800900301.

Gottschall, A. C., West, S. G., & Enders, C. K. (2012). A comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behavioral Research, 47*(1), 1–25. http://dx.doi.org/10.1080/00273171.2012.640589.

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10*, 80–100.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530.

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

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206–213.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2008). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323–343. http://dx.doi.org/10.1037/1082-989X.11.4.323.supp.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology, 12*, 75–88.

Grund, S., Lüdke, O., & Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavioral Research Methods, 48*, 640–649.

Grund, S., Robitzsch, A., & Lüdke, O. (2016). *Package 'mitml'*. Retrieved from: https://cran.r-project.org/web/packages/mitml/.

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implemen- tations. *Behavioral Research Methods, 41*, 924–936.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology, 39*, 265–291.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician, 57*, 229–232.

Hox, J., & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Chichester, UK.: John Wiley & Sons.

Jelicic, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45*, 1195–1199.

Keller, B. T., & Enders, C. K. (2014). A latent variable chained equations approach for multilevel multiple imputation. In *Paper presented at the modern modeling methods Conference, Storrs, Connecticut*.

Licht, C. (2010). *New methods for generating significance levels from multiply-imputed data* (Doctoral dissertation). Universität Bamberg. Retrieved from: http://d-nb.info/101104966X/34.

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association, 86*, 1065–1073.

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

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7*, 199–204.

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

Liu, Y., & Enders, C. K. (2016). *Evaluation of multi-parameter test statistics for multiple imputation*.

Liu, G., & Gould, A. L. (2002). Comparison of alternative strategies for analysis of longitudinal trials. *Journal of Biopharmaceutical Statistics, 12*, 207–226.

Mallinckrodt, C. H., Clark, W. S., & David, S. R. (2001). Accounting for dropout bias using mixed effects models. *Journal of Biopharmaceutical Statistics, 11*, 9–21.

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. *Multivariate Behavioral Research, 50*, 504–519.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*, 23–43.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*, 538–558.

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*, 103–111.

Mistler, S. A., & Enders, C. K. (2011). An introduction to planned missing data designs for developmental research. In B. Laursen, T. Little, & N. Card (Eds.), *Handbook of developmental research methods* (pp. 742–754). New York: Guilford.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., et al. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics, 5*, 445–464.

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*(1), 17–33. http://dx.doi.org/10.1037/a0022634.

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*, 525–556.

Quartagno, M., & Carpenter, J. R. (2016). *jomo: A package for multilevel joint modelling multiple imputation (Version 2.1-2)*. Retrieved from http://CRAN.R-project.org/package=jomo.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*, 85–95.

Raykov, T. (2011). On Testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(3), 419–429. http://dx.doi.org/10.1080/10705511.2011.582396.

Raykov, T., & West, B. T. (2015). On enhancing plausibility of the missing at random assumption in incomplete data analyses via evaluation of response-auxiliary variable correlations. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–9. http://dx.doi.org/10.1080/10705511.2014.937848.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika, 94*(2), 502–508. http://dx.doi.org/10.1093/biomet/asmO28.

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology, 32*, 143–150.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, New Jersey: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489.

Ruehlman, L. S., Karoly, P., & Enders, C. (2012). A randomized controlled evaluation of an online chronic pain self management program. *Pain, 153*, 319–330.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica, 57*, 19–35.

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

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571.

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics, 11*(2), 437–457. http://dx.doi.org/10.1198/106186002760180608.

Shin, Y., & Raudenbush, S. W. (2007). Just-identified versus overidentified two-level hierarchical linear models with missing data. *Biometrics, 63*(4), 1262–1268. http://dx.doi.org/10.1111/j.1541-0420.2007.00818.x.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35*, 26–53.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008

analysis of missing data. *Psychological Methods, 6*, 317–329.
van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*, 219–242.
van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox, & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). New York: Routledge.
van Buuren, S. (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.
van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*, 1049–1064.
van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2014). *Package 'mice'*. Retrieved from: cran.r-project.org/web/packages/mice/mice.pdf.
Widaman, K. F. (2006). Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development, 71*, 42–64.
Wilkinson, L., & Inference, T. F.o. S. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.
Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials, 1*, 368–376.
Yuan, K.-H., & Savalei, V. (2014). Consistency, bias and efficiency of the normal-distribution-based MLE: The role of auxiliary variables. *Journal of Multivariate Analysis, 124*, 353–370. http://dx.doi.org/10.1016/j.jmva.2013.11.006.
Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philos Trans A Math Phys Eng Sci, 366*(1874), 2389–2403. http://dx.doi.org/10.1098/rsta.2008.0038.
Yucel, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling, 11*, 351–370.

Please cite this article in press as: Enders, C. K., Multiple imputation as a flexible tool for missing data handling in clinical research, *Behaviour Research and Therapy* (2016), http://dx.doi.org/10.1016/j.brat.2016.11.008