American Psychological Association

Psychological Methods

1082-989X/19/$12.00

2019, Vol. 24, No. 1, 20–35
http://dx.doi.org/10.1037/met0000182

Fixed Effects Models Versus Mixed Effects Models for Clustered Data: Reviewing the Approaches, Disentangling the Differences, and Making Recommendations

Daniel McNeish
Arizona State University

Ken Kelley
University of Notre Dame

Abstract

Clustered data are common in many fields. Some prominent examples of clustering are employees clustered within supervisors, students within classrooms, and clients within therapists. Many methods exist that explicitly consider the dependency introduced by a clustered data structure, but the multitude of available options has resulted in rigid disciplinary preferences. For example, those working in the psychological, organizational behavior, medical, and educational fields generally prefer mixed effects models, whereas those working in economics, behavioral finance, and strategic management generally prefer fixed effects models. However, increasingly interdisciplinary research has caused lines that separate the fields grounded in psychology and those grounded in economics to blur, leading to researchers encountering unfamiliar statistical methods commonly found in other disciplines. Persistent discipline-specific preferences can be particularly problematic because (a) each approach has certain limitations that can restrict the types of research questions that can be appropriately addressed, and (b) analyses based on the statistical modeling decisions common in one discipline can be difficult to understand for researchers trained in alternative disciplines. This can impede cross-disciplinary collaboration and limit the ability of scientists to make appropriate use of research from adjacent fields. This article discusses the differences between mixed effects and fixed effects models for clustered data, reviews each approach, and helps to identify when each approach is optimal. We then discuss the within–between specification, which blends advantageous properties of each framework into a single model.

Translational Abstract

Even though many different fields encounter data with similar structures, the preferred method for modeling such data can be vastly different from discipline to discipline. This is especially true in the case of clustered data where in subsets of observations belong to the same higher order unit, as is common in organizational science, education, or longitudinal studies. To model such data, researchers trained in the economic tradition primarily rely on fixed effects models, whereas researchers trained in the psychological tradition employ mixed effects models. As the disciplinary lines between these economics and psychology continue to be blurred (e.g., in fields such as behavioral economics or strategic management), the disparity in approaches to statistical modeling can prevent dissemination and proper interpretation of results. Additionally, each of these statistical methods has certain limitations that can prevent answering particular research questions, limiting the scope of hypotheses that can be tested. The goal of this article is to compare and contrast the fixed effect and mixed effect modeling frameworks to overview the general idea behind each and when employing each method may be most advantageous. We also discuss ways in which aspects of both models can be blended into a single framework to maximally benefit from what each method can provide.

Keywords: fixed effect model, multilevel model, HLM, panel data, random coefficients model

Clustered data are ubiquitous in many contexts; classical cross-sectional examples include students nested within schools in educational contexts, clients nested within therapists in clinical psychology, patients nested within doctors in medicine, persons clustered within neighborhoods in epidemiology, employees nested within firms in business, and hospitals within systems in health care (Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Clustering also occurs by repeated measures on multiple units, often deemed longitudinal or panel data, in which the same information is tracked across time for the same units. In this context, the nesting structure is in terms of the multiple observations being nested within a single entity (Bollen & Curran, 2006; Greene, 2003; Rabe-Hesketh & Skrondal, 2008; Singer & Willett, 2003; Wooldridge, 2002). When data are cross-sectional, panel data, or a combination of the two, the nested structure of the data must be accommodated in the analysis. A failure to do so violates the independence assumption postulated by many popular

This article was published Online First June 4, 2018.

Daniel McNeish, Department of Psychology, Arizona State University; Ken Kelley, Department of Information Technology, Analytics, and Operations, and Department of Psychology, University of Notre Dame.

Correspondence concerning this article should be addressed to Daniel McNeish, Department of Psychology, Arizona State University, P.O. Box 871104, Psychology Building, Tempe, AZ 85287, or to Ken Kelley, Department of Management–Analytics, University of Notre Dame, Mendoza College of Business, Notre Dame, IN 46556. E-mail: dmcneish@asu.edu or KKelley@ND.edu

models, which leads to biased standard error estimates (often downwardly biased) and improper statistical inferences (e.g., Liang & Zeger, 1986; Maas & Hox, 2005; Moulton, 1986, 1990; Raudenbush & Bryk, 2002; Wooldridge, 2003).

Fortunately, several methods exist to account for clustering so that inferences made with clustered data are appropriate and yield test statistics that align with the theoretical properties (e.g., well-behaved Type I error rates and confidence interval coverage). There are strong disciplinary preferences by which clustered data are modeled (McNeish, Stapleton, & Silverman, 2017). For example, a review performed by Bauer and Sterba (2011) found that about 94% of psychology studies between 2006 and 2011 accounted for clustering via mixed effects models (MEMs). Conversely, a review by Petersen (2009) showed almost the exact *opposite* trend in econometrics—only 3% of studies reviewed used MEMs. As noted in McNeish et al. (2017), MEMs are so frequently used in psychological research that it has, for all intents and purposes, become synonymous with modeling clustered data. However, as noted by recent articles (Huang, 2016; McNeish & Stapleton, 2016), several alternative methods exist for modeling clustered data. In addition to classical approaches, such as within-subjects analysis of variance and multivariate analysis of variance for modeling clustered data (e.g., Maxwell, Delaney, & Kelley, 2018), more modern approaches have included the use of MEMs, which have become popular in psychology, or fixed effects models (FEMs), which have become popular in fields grounded in economics. Recent studies have contrasted the typical approach for modeling clustered data in psychology to methods common in biostatistics, such as generalized estimating equations (McNeish et al., 2017), and survey methodology methods, such as Taylor series linearization (Huang, 2016). However, there has been less attention paid to differences between MEMs that are widely adopted by psychologists and FEMs that tend to be preferred by researchers working in the econometric tradition. Given that there can be considerable disparities in training within psychology-adjacent fields (e.g., industrial-organizational psychology, management studies, education, and policy studies), researchers trained in one tradition may have difficulty comprehending the rationale behind the analytic methods chosen by their colleagues within their own field but trained in another statistical tradition. This may result in difficulty interpreting statistical results in such fields, which can cause confusion and inhibit the pace of development as such fields move forward.

This article reviews FEMs, which are popular in fields grounded in an econometrics tradition (A. Bell & Jones, 2015) but not well known in psychology. We provide comparisons between the FEMs and MEMs, making recommendations for when each should be used. Generally, FEMs account for a clustered data structure by including the cluster affiliation information directly into the model as a predictor (i.e., a fixed effect) rather than treating cluster-specific quantities as random effects. In other words, FEMs are equivalent to adding categorical predictors to the model representing the cluster variable (e.g., including an intercept for each person in a longitudinal model). Some researchers in fields related to economics have gone so far as to refer to the FEM framework as the "gold standard" for modeling clustered data (Schurer & Yong, 2012). In the political science literature, in which the econometric tradition has a strong presence, a recent article by A. Bell and Jones (2015) tried to dissuade researchers from using FEMs, arguing that they are overused much in the same way that McNeish et al. (2017) argued that MEMs are overused in psychology. Clearly, some clarification and recommendations are needed to help clarify when each method is appropriate.

Despite the high praise for, and near ubiquity of, FEMs in the econometrics and related literatures, psychological researchers seem largely indifferent with this modeling framework. As evidence for this claim, we searched two flagship empirical psychology journals for references to FEMs and MEMs: *Journal of Personality and Social Psychology* and *Journal of Applied Psychology*, during the publication years 2015 and 2016. In the *Journal of Personality and Social Psychology*, 0 of 247 studies contain the phrase "fixed effects model" or "fixed-effects model." When searching for "fixed effect" instead, we found three studies, but each of these referred to fixed factors in a fixed effects ANOVA context or a fixed effect in a MEM context. In the *Journal of Applied Psychology,* we found three studies from a possible 399 articles (0.8%) containing the phrase "fixed effects model" or "fixed-effects model" as used in an FEM context. The more relaxed search term "fixed effect" yielded two additional studies, but both of these additional used the phrase in the context of ANOVA or MEMs. In a pair of searches performed in leading economic journals, 33 of 79 articles (42%) published in the *Quarterly Journal of Economics* during the same time period featured FEMs, and 131 of 248 articles published in the *Journal of Financial Economics* (53%) featured FEMs. There were also major discrepancies between statistical texts in their discussion of analyzing clustered data. Texts commonly used in psychology, such as Raudenbush and Bryk (2002), Maxwell et al. (2018), and Hox (2010), do not feature any coverage of FEMs, whereas Snijders and Bosker (2012) discuss FEMs only briefly (pp. 43–48). Conversely, the Greene (2003)[1] text on econometric analysis briefly discusses MEMs (pp. 293–295, 318–320), with additional discussion of the different estimation procedures (pp. 295–301).

This vast disconnect between fields with a psychological foundation and fields with an economic foundation begs the question, are psychology and related fields at a disadvantage by not considering an entire modeling framework that is the principal approach to modeling clustered data in economics? Conversely, are economics and related fields at a disadvantage by predominantly focusing their modeling efforts on FEMs? These questions have particularly strong implications as disciplinary boundaries between business, psychology, and economics continue to be blurred.

To address the chasm between MEMs and FEMs, we provide an overview of both the MEM and FEM frameworks to help readers understand alternative frameworks with which they may not be familiar. We then discuss strengths and limitations of each model for types of analyses common in psychology and provide empirical examples. Particular emphasis is placed on discussion of the exogeneity assumption of MEMs, whose tenability has led to different methods being preferred in different fields. We conclude with the within–between specification of a MEM, a method that incorporates advantages of both FEMs and MEMs, especially with respect to treatment of the exogeneity assumption. We discuss how this

[1] Page numbers refer to the fifth edition of this text.

specification flexibly applies elements of both methods, making analyses stronger across both traditions.[2]

Overview of Mixed Effects Models

In MEMs, the clustered structure of the data is accounted for by including random effects in the model (Laird & Ware, 1982; Stiratelli, Laird, & Ware, 1984). Coefficients in MEMs represent two possible types of effects: fixed effects or random effects. Fixed effects are estimated to represent relations between predictors and the outcome irrespective to which cluster observations belong, similar to a standard single-level multiple linear regression model (Raudenbush & Bryk, 2002). Random effects, on the other hand, capture how much the relation between the predictor and the outcome differs from the fixed effect estimate for a specific cluster, essentially capturing the unique effect of the predictor in the cluster of interest. Put another way, the effects within each cluster form a distribution of effects in which the fixed effect is the mean and the random effects are unique data points.

Mathematically, MEMs for continuous outcomes can be expressed as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\varepsilon}_j, \tag{1}$$

where $\mathbf{y}_j$ is an $m_j \times 1$ vector of responses for cluster j and m_j is the number of units within cluster j, $\mathbf{X}_j$ is an $m_j \times p$ design matrix for the fixed predictors in cluster j (at either level) and p is the number of predictors (including the intercept), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $\mathbf{Z}_j$ is an $m_j \times q$ design matrix for the random effects of cluster j, $\mathbf{u}_j$ is a $q \times 1$ vector of random effects for cluster j, q is the number of random effects (where $p \geq q$), $E(\mathbf{u}_j) = \mathbf{0}$, $Cov(\mathbf{u}_j) = \mathbf{G}$, $\mathbf{G}$ is $q \times q$, $\boldsymbol{\varepsilon}_j$ is an $m_j \times 1$ vector of residuals of the observations in cluster j, $E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$, and $Cov(\boldsymbol{\varepsilon}_j) = \mathbf{R}$. Although there are several estimators that can be used to estimate MEMs, either maximum likelihood or restricted maximum likelihood are typically used (e.g., SAS Proc Mixed, SPSS, Stata, R lme4). For more detail on the algorithms and methods typically employed to estimate these models, see Goldstein (1986, 1989). For foundational information on restricted maximum likelihood estimation, see Harville (1977) or Patterson and Thompson (1971). For a more general statistical overview of the mixed effects model, see Laird and Ware (1982). For a conceptual overview of the differences between maximum likelihood and restricted maximum likelihood estimation, see McNeish (2017b).

Assumptions

Although MEMs permit rich models to be fit to clustered data, the inclusion of random effects in these models require that several assumptions be made. An important assumption of MEMs is that the predictor variables included in the model do not covary with either the residuals or the random effects (i.e., $Cov[X, u] = Cov[X, r] = 0$), which is commonly referred to as the *exogeneity assumption* (e.g., Gardiner et al., 2009). When the exogeneity assumption is violated, the model is said to be endogenous, meaning that there is an unmodeled relation that establishes nonzero covariance between a predictor in the model and a random error term. That is, there are omitted confounders that threaten construct validity (e.g., see Maxwell et al., 2018, for a review). If the exogeneity assumption is violated, then the coefficient estimates will contain noticeable amounts of bias (Greene, 2003). We will return to this issue in subsequent sections, because differing disciplinary perspectives on this assumption are a primary reason for preferring either MEMs or FEMs.

Other assumptions of MEMs include that (a) all relevant random effects have been included in the model, (b) the covariance structures of the residuals and random effects have been properly specified, and (c) the residuals and the random effects follow multivariate normal distributions and do not covary across levels. Violating these assumptions can affect model selection as well as estimates and inferences made from the model. For more information on the ramifications of assumption violations, see Ebbes, Wedel, Böckenholt, and Steerneman (2005), Kim and Frees (2006, 2007), McNeish et al. (2017), and Raudenbush and Bryk (2002, Chapter 9).

Hypothetical Example

Although matrix notation can help facilitate mathematical details of MEMs, it is more common to see Raudenbush and Bryk (RB) notation in empirical applications. RB notation more clearly displays the location and relations between predictors and elucidates why these models are often referred to as hierarchical or multilevel models. To demonstrate RB notation, consider an example in which work motivation (at Level 1, the employee level) and the quality of the company's incentive program (at Level 2, the company level) predict employee productivity.[3] This model would be represented as

$$Productivity_{ij} = \beta_{0j} + \beta_{1j}\,Motivation_{ij} + r_{ij} \tag{2a}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Incentive_j + u_{0j} \tag{2b}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Incentive_j, \tag{2c}$$

where Productivity_{ij} is the productivity score for the ith employee in the jth company, β_{0j} is the company-specific intercept for the jth company, β_{1j} is the company-specific slope for motivation for the jth company, and r_{ij} is the residual for the ith person in the jth company.

[2] Comparisons of FEMs and MEMs have appeared in other literatures over the last decade (e.g., Clark & Linzer, 2015; Gardiner, Luo, & Roman, 2009; Greene, 2003; Halaby, 2004; Yang & Land, 2008) but with little relevance to psychology. These previous treatments have focused on intersections of areas outside of psychology (healthcare and economics, economic sociology) and cover analyses that are germane to those areas but that are not always common in psychology. The current article discusses these models in contexts directly relevant to the study of psychological phenomenon.

[3] The notation in Equation 1 does not differentiate at which level the predictor exists. This can be seen much more clearly in RB notation. Level 1 predictors are collected at the lowest level of the hierarchy (e.g., employees if data are collected in companies, time-varying covariates if data are collected longitudinally). These variables appear in the first equation in RB notation. Level 2 predictors are collected at the higher level of the hierarchy (e.g., company-level variables if data are collected on employees clustered within companies; time-invariant predictors if data are collected longitudinally). Level 2 predictors can either be properties of the Level 2 unit itself (e.g., school size) or an aggregation of the Level 1 units with the Level 2 unit (e.g., average salary at a company). Although other methods do not formally adopt the level-specific designation of predictors, researchers often use this terminology out of convenience.

Equation 2a then allows researchers to model the company-specific intercept (β_{0j}; i.e., the Level 1 coefficients in Equation 2a are the dependent variables in Equations 2b and 2c). In Equation 2b, γ_{00} is the overall intercept for productivity across all companies, γ_{01} is a regression coefficient that captures how much the intercept of productivity is expected to change for a one-unit change in incentive for company j, and u_{0j} is a random effect for the jth company that captures how much the intercept for the jth company differs from the overall intercept γ_{00} after accounting for incentive. The company-specific intercepts β_{0j} follow a normal distribution that is centered on $\gamma_{00} + \gamma_{01}Incentive_j$, with a variance denoted by τ_{00} in this notation. If u_{0j} is not included in Equation 2b, then this implies that τ_{00}, the variance of the intercepts between different clusters, is equal to zero. Equation 2c contains similar information as Equation 2b except that the outcome is now the company-specific slope (β_{1j}) rather than the company-specific intercept. In Equation 2c, the absence of a random effect for motivation means that, controlling for the incentive, the effect of motivation on productivity does not vary between companies. The three separate equations can be combined into one single equation such that

$$Productivity_{ij} = \gamma_{00} + \gamma_{01}Incentive_j + (\gamma_{10} + \gamma_{11}Incentive_j) \times Motivation_{ij} + u_{0j} + r_{ij}. \quad (3)$$

Note that Equation 3 contains a "mix" of the fixed effect γ parameters as well as the random effect u parameter, giving rise to the "mixed effect model" name.

In general, MEMs are a flexible method that allows researchers much freedom in building models that test their research questions and theories. However, researchers pay for this flexibility with assumptions.

Overview of Fixed Effects Models

As noted in Gelman and Hill (2006, p. 245) and Gardiner et al. (2009), the term "fixed effect" has several, nonoverlapping definitions when used in various branches of statistics. When psychologists hear or see "fixed effects," they tend to think of (a) parameters in MEMs that represent the association between a Level 1 predictor and the outcome across all clusters, or (b) levels of one or more factors (e.g., control group, Treatment A, Treatment B) that are purposely chosen in the context of analysis of variance.

In the econometric tradition, "fixed effects" does not refer to a particular component of a model for clustered data or a type of research design. Rather, "fixed effects" refers to an entire modeling framework (Allison, 2009). Similar to MEMs, FEMs explicitly model the clustered structure of the data. However, FEMs do not use random effects or random coefficients (i.e., the u parameter in Equation 2 and 3). Instead, cluster affiliation dummy variables are included directly in the model as predictor variables. A regression coefficient is then estimated for each cluster affiliation variable to yield cluster-specific estimates, much in the same way that each cluster has a unique random intercept estimate in MEMs (Gardiner et al., 2009). To make another connection, the FEM can be seen as an ANCOVA in which the cluster affiliation is a categorical factor predicting the outcome. All variability at Level 2 is explained in the FEM because the cluster affiliation variables are treated as predictors (unlike MEMs, in which Level 2 variability is a partitioned component of the residual variance). Unlike traditional ANCOVA, the focus of the FEM is on the covariates, with the categorical cluster factor serving as a control variable to account for the data structure.

Such a specification has noticeable ramifications for how variability at Level 2 is accounted for, which can be viewed as an advantage or disadvantage depending on the research question at hand. Though we will discuss this issue in much more detail in subsequent sections, the general consequence of the FEM is that the cluster affiliation variables account for all of the variability at Level 2. This means that researchers need not be concerned with including Level 2 predictors in the model, because variance attributable to all Level 2 variables (whether available in the data or not) is consumed by the cluster affiliation variables. As a result, all Level 1 coefficients are conditional on all possible Level 2 effects. As a cost of such a strategy, researchers lose the ability to estimate coefficients for any individual Level 2 predictor in the model—once the cluster affiliation variables are included in the model, only coefficients for Level 1 predictors can be estimated.

MEMs assume that random effects are drawn from a particular distribution (often a normal distribution); however, FEMs do not assume that clusters are a random sample from the possible population of clusters. This represents one reason why FEMs tend to be popular in economics research—clustering units are frequently selected for a specific purpose (e.g., European countries, chief executive officers from Fortune 100 companies, top-rated universities) instead of being randomly selected.

In FEMs, the creation of cluster-specific affiliation variables can be done with two different dummy-coding methods. The first method of coding, which we will refer to as *absolute coding*, is to include as many cluster affiliation variables as there are clusters and to remove the intercept to prevent overparameterization of the model. In this case, estimates for the cluster affiliation variables represent the intercept value for each specific cluster, similar to how each cluster receives a cluster-specific intercept estimate in MEMs. Another method of coding, which we refer to as *reference coding*, is to retain the intercept estimate in the model and leave one of the cluster affiliation variables out as a reference cluster. Cluster affiliation variable estimates with this method represent the difference in the intercept between a specific cluster and the reference cluster whose cluster affiliation variable was omitted form the model. Thus, the first method gives the cluster affiliation estimates an absolute interpretation, whereas the second method gives a relative interpretation.

Notationally, using an absolute coding scheme, the FEM can be written as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{C}_j\boldsymbol{\alpha} + \mathbf{r}_j, \quad (4)$$

where $\mathbf{y}_j$ is an $m_j \times 1$ vector of responses for the jth cluster,[4] $\mathbf{X}_j$ is an $m_j \times p$ design matrix of substantive predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of substantive regression coefficients, $\mathbf{C}_j$ is an $N \times J$ matrix of cluster affiliation dummy codes, $\boldsymbol{\alpha}$ is a $J \times 1$ vector of

[4] Readers may note that the FEM is a single-level model, but we define the model using j subscripts to denote the cluster. Though not required, we use this model notation because it elucidates the role of the cluster affiliation dummy variables and the associated coefficients. Otherwise, this information would be appended to the end of the $\boldsymbol{\beta}$ matrix, where its function may be less clear.

cluster-specific intercepts, and $\mathbf{r}_j$ is an $m_j \times 1$ vector of residuals that is traditionally assumed to be distributed N ($\mathbf{0}$, $\sigma^2\mathbf{I}$). Because the cluster affiliation dummy variables account for all cluster-level variance, the FEM residual variance is equivalent to the Level 1 residual variance in MEMs, not the total residual variance across all levels. Because there are no random effects in the model,[5] FEMs with continuous outcomes are often estimated with ordinary least squares (OLS; the same as a single-level multiple linear regression model). The form of the model presented in Equation 4 assumes that coefficients are equal across clusters. As will be discussed later, this can be relaxed.

Assumptions

A primary benefit of the FEM framework is that there are far fewer assumptions compared with MEMs. In FEMs, the model only needs to be properly specified with respect to Level 1[6] predictors (time-varying predictors in panel data) for the parameter estimates and standard errors to have desirable statistical properties: The cluster affiliation dummy variables handle all possible Level 2 predictors (time-invariant predictors in panel data) regardless of whether the variable was collected in the data or not. Level 1 predictors must have variability within clusters because FEMs only explicitly model within-group variation (all variables without within-group variance are accounted for by the cluster affiliation dummy variables). This means that effects for Level 2 predictors cannot be directly estimated within a FEM (Baltagi, 2013; Hsiao, 2003; Kim & Frees, 2006), although they will be accounted for in the model via the cluster affiliation dummy variables (i.e., other effects are conditional on the Level 2 predictors, but effects of specific Level 2 variables are not estimable). If an FEM model is estimated with OLS, then it is conventional to assume that the residuals are distributed N ($\mathbf{0}$, $\sigma^2\mathbf{I}$). This assumption may be reasonable with cross-sectional clustering but could be violated with panel data because of the presence of autocorrelation (A. Bell & Jones, 2015), meaning that the residuals at different time points are not independent of one another. However, other estimation methods like generalized least squares or maximum likelihood can account for this dependence with different types of residual structures (e.g., compound symmetry, Toeplitz). Unlike MEMs, FEMs do not assume that clusters are sampled randomly.

Hypothetical Example

To see how FEMs differ from MEMs, consider the example from the previous section in which employee productivity is predicted by work motivation and the quality of the company's incentive program. If there are five companies in the data, the FEM would be written out completely as

$$\begin{aligned} Productivity_{ij} = {} & \beta_1 Motivation_{ij} + \beta_2(Motivation_{ij} \times Incentive_j) \\ & + \alpha_1 Company1 + \alpha_2 Company2 + \alpha_3 Company3 \\ & + \alpha_4 Company4 + \alpha_5 Company5 + r_{ij}, \end{aligned} \quad (5)$$

where β_1 captures the relation between motivation and productivity, β_2 captures how the effect of motivation changes depending on the values of incentive (i.e., the interaction of the two), and the α parameters represent the company-specific intercepts (we use the absolute coding here). Notice that the incentive predictor does not explicitly appear in the model as a separate term (although it does appear as part of an interaction) because the main effect of incentive is included in the variance accounted for by the company variables. The tests for motivation and the Motivation × Incentive interaction therefore control for incentive even though it is no longer possible to estimate the specific effect of incentive, because this variable is perfectly collinear with the company variables. This is an important distinction in FEMs—the cluster affiliation variables include all the measured and unmeasured variables at Level 2 even though the specific effects of Level 2 predictors are unobtainable. That is, all effects in the model are conditional on incentive (and all other Level 2 variables), but the main effect for incentive specifically is inestimable. The company variables subsume all Level 2 variables and explain all the variance at Level 2; therefore, there is no unique variance left for any single Level 2 variable to explain because any variance explained any Level 2 variables will necessarily overlap completely with variance explained by the company variables. The Motivation × Incentive interaction term is permissible even though it involves a Level 2 variable because there is still within-cluster variation once the multiplicative term is created. Table 1 shows how parameters from Equation 2 map onto parameters in Equation 5.

De-Meaning

The previous FEM description is commonly referred to as the least square dummy variable (LSDV) fixed effects model. There is another alternative specification that may be used to achieve the same goal without needing to create dummy variables, which can be useful when there are many clusters and estimation would be computationally burdensome with many dummy variables (e.g., in big data; Allison, 2009).[7] This alternative approach is referred to as *de-meaning*, in which the cluster means are subtracted from the observed values for all variables. The de-meaned version of Equation 5 without the Motivation × Incentive cross-level interaction would be

$$\begin{aligned} (Productivity_{ij} - \overline{Productivity_j}) = {} & \beta_1(Motivation_{ij} - \overline{Motivation_j}) \\ & + (r_{ij} - \bar{r}_j). \end{aligned} \quad (6)$$

To demonstrate how de-meaning works, consider Equation 5 as a random intercepts MEM such that

$$Productivity_{ij} = \beta_1 Motivation_{ij} + u_{0j} + r_{ij}. \quad (7)$$

[5] Technically, FEMs do have a random component in the model because the residuals are random effects and are assumed to be randomly drawn from a particular distribution. In this article, we reserve use of "random effects" for random effects at Level 2 and refer to what are technically Level 1 random effects as "residuals."

[6] Although we talk about "levels" in FEMs, note that FEMs do not consider different levels of analysis as in MEMs. The FEM model does not differentiate between Level 1 or Level 2 variables. The FEM is an inherently single-level regression model that uses clever coding to manipulate the single-level framework into estimating coefficients from clustered data. We use "Level 1" and "Level 2" to facilitate the discussion about parameters that are similarly estimated by MEMs and FEMs. In the FEM framework, Level 1 means that the variable has variability within clusters, whereas Level 2 means that that variable has no variability within a cluster.

[7] Such big data situations will likely become more common in psychology and related fields due to the instrumented world in which we now live and such methods can measure various aspects of behavior (e.g., Adjerid & Kelley, 2018).

Table 1
Comparison of MEM and FEM Model Parameters

From MEM	To FEM	From FEM	To MEM
γ_{00}	Weighted average of α_1 to α_5	β_1	γ_{10}
γ_{01}	None, included in α_1 to α_5	β_2	γ_{11}
γ_{10}	β_1	α_1	u_{01}
γ_{11}	β_2	α_2	u_{02}
u_{0j}	α_j^*	α_3	u_{03}
r_{ij}	r_{ij}	α_4	u_{04}
		α_5	u_{05}
		r_{ij}	r_{ij}

Note. MEM = mixed effects model; FEM = fixed effects model.
* α and u are conceptually related but are not the same. α_j are fixed effects, whereas u_j are random effects. Depending on how the FEM is specified, α_j is either the complete intercept for cluster j or the difference between the intercept of cluster j and a reference cluster. In MEM, u_j are the deviations from the overall mean and are defined to have a mean of 0. Even if all other aspects of the model are identical, the u_j and α_j estimates will differ because u_j will be shrunken via empirical Bayes estimation.

If Equation 7 were de-meaned (i.e., means are subtracted so that the variables are in terms of deviations from the mean), any variable with no variation within clusters (i.e., Level 2 predictors and random effects) will be reduced to a constant value of zero and will therefore drop out of the model because the cluster mean and the observed values will necessarily be identical:

$$\begin{aligned}&\left(Productivity_{ij} - \overline{Productivity_j}\right)\\ &\quad = \beta_1(Motivation_{ij} - \overline{Motivation_j}) + (u_{0j} - \bar{u}_{0j}) + (r_{ij} - \bar{r}_j)\\ &\quad = \beta_1(Motivation_{ij} - \overline{Motivation_j}) + (r_{ij} - \bar{r}_j). \qquad (8)\end{aligned}$$

Equation 6 and Equation 8 are the same and are also equivalent to Equation 5 without the Motivation × Incentive interaction term. However, there is a difference in implementation. Using the LSDV FEM, software will use the correct degrees of freedom for the standard error calculation, because the cluster affiliation dummy variables are directly included in the model and consume one degree of freedom each, as with any categorical predictor. In the de-meaned approach, degrees of freedom will not be computed properly because the subtraction of the means occurs as a data preprocessing step that precedes the estimation of the model. As a result, the p values and inferential tests will not be correct and must be adjusted to reflect the implicit inclusion of the cluster means in the model (Judge, Griffiths, Hill, & Lee, 1985).

Comparing MEMs and FEMs

MEMs and FEMs each offer defensible ways to model clustered data and have the same general goal. However, there are considerable differences with regard to the type of analytic scenarios in which each can be optimally applied, or even applied at all. Differences between these two approaches across a range of scenarios are summarized in Table 2. In the subsequent sections, we provide additional discussion to highlight the more salient differences between the methods.

Sample Size

In many disciplines, the number of clusters in data sets tends to be small due to practical concerns, such as financial limitations, the use of extant data sets, or difficulties in recruiting large numbers of participants (Dedrick et al., 2009; McNeish & Stapleton, 2016). MEMs applied to data with fewer than 30 clusters have been shown to be at risk for downwardly biased estimates of the variance component and regression coefficient standard error estimates (e.g., B. Bell, Morgan, Schoeneberger, Kromrey, & Ferron, 2014; Maas & Hox, 2005). Several remedies have been proposed, including the Kenward-Roger correction (Kenward & Roger, 1997), Bayesian Markov Chain Monte Carlo (MCMC) estimation (Hox, van de Schoot, & Matthijsse, 2012), and Monte Carlo resampling methods (Bates, 2010). These methods are generally serviceable in practice although they are not without their faults (e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; McNeish, 2016, 2017b).

Because FEMs are routinely estimated with OLS, they tend to be much less susceptible to bias when data have very few clusters, because standard errors have a closed form and there is no need to estimate variance components. For example, for cross-sectional clustering and fewer than 15 clusters, McNeish and Stapleton (2016) found that FEMs performed the best overall with respect to minimizing bias, controlling Type I error rates, and maximizing power from among 12 competing small sample methods, including MCMC MEMs with the half-Cauchy prior recommended for small samples by Gelman (2006).

As noted previously, coefficient estimates of Level 2 predictors cannot be obtained from FEMs. This may lead some researchers to acknowledge the superior small sample performance, but note that FEMs ultimately do not estimate all the quantities that are often of interest in many contexts. Although true, note that these Level 2 coefficients are typically estimated rather poorly with fewer than 20 clusters (Baldwin & Fellingham, 2013; Hox et al., 2012; Stegmueller, 2013), so researchers should carefully consider the importance of Level 2 coefficients in such cases. Prioritizing the estimation of these Level 2 coefficients could lead to biased estimates throughout a MEM, whereas a FEM would limit the coefficient estimates to Level 1 coefficients. However, FEMs would safeguard against bias for the estimates the model does yield (i.e., Level 1 coefficients) while controlling for measured and unmeasured Level 2 effects, provided that FEM assumptions are met.

Cross-Level Interactions

Cross-level interactions refer to an interaction of two variables in which one variable is at Level 1 (e.g., the employee level) and the other is at Level 2 (e.g., the company level).[8] Cross-level interactions can be tested using either the FEM or MEM framework, with examples being presented in Equation 2 and Equation 5. In Equation 2, the cross-level interaction for a MEM is present by modeling the coefficient of a Level 1 predictor as a function of a Level 2 predictor (the γ_{11} specifically represents this cross-level interaction effect). In the FEM model in Equation 5, the same cross-level interaction effect is captured by the β_2 parameter.

[8] Although we talk about "levels" in FEMs, note that FEMs do not consider different levels of analysis as in MEMs. The FEM model does not differentiate between Level 1 or Level 2 variables. The FEM is an inherently single-level regression model that uses clever coding to manipulate the single-level framework into estimating coefficients from clustered data.

Table 2
Comparison of the Types of Modeling Questions That Can Be Assessed With MEMs and FEMs

Modeling problem	MEM	FEM
Accommodation of clustering	Random effects must be explicitly modeled by the user. The covariance matrix of the random effects also must be explicitly modeled.	Cluster affiliation dummy variables are included directly in the model.
Common estimation method	(Restricted) Maximum likelihood	Ordinary least squares
Predictors at Level 2	Allowed and coefficients are directly estimated. Proper specification is required, meaning that no relevant variables are omitted.	Generally not estimable (although there are proposed methods that claim to be able to provide estimates under particular circumstances). Omitted Level 2 variable bias is not a concern.
Omitted variable bias	A concern at all levels	Only a concern at Level 1
Accommodation of variability at Level 2	Predictor variables and random effects	Cluster affiliation dummy variables
Coefficient interpretation	Coefficients at either level are interpreted conditional on the variables explicitly included in the model.	Level 1 coefficients are conditional on all Level 2 variables (measured and unmeasured) being accounted for.
Level 2 sample size requirement	30 is the general recommendation; can be reduced (to about 10) if corrective procedures are used	Viable with very small Level 2 sample sizes
Cluster-varying slopes	Easily modeled with random effects	Must use interaction terms with cluster affiliation variables
Supports impure hierarchies	Yes	No
Efficiency	More efficient (smaller standard errors) but more likely to be biased if assumptions are violated or if the model is misspecified	Less efficient but less likely to produce biased estimates because there are fewer assumptions and fewer locations where misspecifications can occur
Mediation possible	Yes	Only if all variables are at Level 1
Analysis of contextual effects	Yes	No
Cross-level interactions	Yes	Yes
Extendable to three-level hierarchies (or more)	Yes	Not with a purely FEM
Assumption about sampling of clusters	Clusters are randomly sampled and representative of the population	Clusters need not be randomly sampled

Note. MEM = mixed effects model; FEM = fixed effects model.

Although FEMs do not allow for Level 2 predictors to be included in the model, cross-level interactions can be assessed with FEMs, because interactions of variables collected at different levels (e.g., employee and supervisor; client and therapist, student and teacher) will differ for each Level 1 unit and will therefore *not* be perfectly collinear with the cluster affiliation variables (i.e., there is within-cluster variation). To include these predictors in a FEM, similar to MEM, a product term is created and included in the model. Even though the Level 2 predictor cannot be included in the model, the Level 2 predictor main effect is still controlled for by the cluster affiliation dummy variables because these variables account for all variance attributable to measured and unmeasured Level 2 variables (Allison, 2009). Interactions of two Level 1 variables are also permissible in FEMs, but interaction effects of two Level 2 variables are not possible in FEMs because there will be no within-cluster variation.

Cluster-Varying Effects

A common research question is whether Level 1 effects vary in different clusters. For example, in an organizational setting, a researcher may want to determine if the effect of motivation on productivity is the same across supervisors. In MEMs, this is straightforward to test—one only needs to include a random slope for the Level 1 variable of interest (motivation, in this hypothetical example). If the associated random effect has a large variance relative to the magnitude of the coefficient, then it can be concluded that the effect differs depending to which cluster an employee belongs. Otherwise, the effect can be declared constant across clusters and the random effect removed. Using the hypothetical example from Equation 2, the effect of motivation on productivity can vary by clusters by adding a single random effect (u_{1j}) into the model, such that

$$Productivity_{ij} = \beta_{0j} + \beta_{1j} Motivation_{ij} + r_{ij} \quad (9a)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Incentive_j + u_{0j} \quad (9b)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Incentive_j + u_{1j} \quad (9c)$$

Similar to the interpretation of the random intercept (u_{0j}) above, u_{1j} allows the company-specific slope β_{0j} to change for each company beyond the change predicted by incentive. The u_{1j} term captures the difference in the effect of motivation from the overall slope γ_{10} for the jth cluster, conditional on incentive.

In FEMs, cluster-varying effects are possible but are more difficult to specify and interpret. As discussed so far, cluster affiliation coefficients in FEMs are similar to a random intercepts MEMs such that each cluster receives a unique intercept estimate. However, the Level 1 coefficients in FEMs do not vary across clusters as described so far, which would be equivalent to a MEM without random slopes. If cluster-varying effects are desired, this can be accomplished by including an interaction term between the

cluster affiliation variables and the Level 1 predictor into the model. In this way, the Level 1 predictor will be estimated to have a unique effect in each cluster.

To make this model more explicit, recall the hypothetical example of modeling productivity as a function of motivation and incentive. In Equation 5, the model posits that the effect of motivation is constant across all companies, conditional on the other effects that are in the model. If a researcher wanted to model unique effects for motivation in each company, the main effect for Motivation would be removed and Motivation × Company interactions would be included, such that

$$\begin{aligned} Productivity_{ij} = {} & \beta_1(Motivation_{ij} \times Company1) \\ & + \beta_2(Motivation_{ij} \times Company2) \\ & + \beta_3(Motivation_{ij} \times Company3) \\ & + \beta_4(Motivation_{ij} \times Company4) \\ & + \beta_5(Motivation_{ij} \times Company5) + \alpha_1 Company1 \\ & + \alpha_2 Company2 + \alpha_3 Company3 + \alpha_4 Company4 \\ & + \alpha_5 Company5 + r_{ij} \end{aligned} \quad (10)$$

Coefficients β_1 through β_5 now represent the effect of motivation on productivity for each respective cluster, similar to random effects in a MEM. There are also alternative ways in which this could be coded, such as leaving the main effect of motivation in the model but including $J - 1$ interaction terms (i.e., reference coding). The coefficients of these interaction terms would then represent the difference in the effect of motivation on productivity from the effect of motivation in the reference cluster. As can be seen from Equation 10, if there are many clusters in the model, estimation of the model with this approach can quickly become burdensome due to the large number of parameters.

Although this type of specification can achieve the goal of modeling a cluster-varying effect within the FEM framework, such an approach does have disadvantages. One disadvantage is that each cluster affiliation by Level 1 predictor interaction coefficient will consume a degree of freedom, meaning that researchers may not have enough degrees of freedom to estimate multiple cluster-varying effects or even a single cluster-varying effect if there are many predictors in the model. Another disadvantage is that, unlike MEMs, which estimate a variance component for the random slopes, which can inferentially tested (although specialized 50:50 mixture chi-square tests are usually needed for this test; see Stram & Lee, 1994), FEMs have less intuitive methods to infer whether a cluster-varying effect is needed in the model. If reference coding is used such that the interaction terms are compared with a reference cluster, a multiparameter test for the interaction term (e.g., omnibus F test, Wald multiparameter Type III test) may provide helpful inferential information. With absolute coding, such a multiparameter test would likely be of little use, because it would be testing whether the cluster slopes are collectively equal to zero rather than testing if they are different from each other (with reference coding, these two questions are synonymous). The variance of the cluster affiliation effects can also be calculated by saving all of the estimates and running descriptive statistics, although this will not be equal to a MEM variance component estimate because it will not account for aspects such as unequal cluster sizes or differential reliability in each cluster.

Mediation

Using our running example, suppose that a researcher's interest is, again, the effect of motivation on productivity; however, now it is of interest to test whether the effect of job satisfaction on productivity is mediated by motivation. Mediation analysis is one of the most important methods for explaining causal pathways. Mediation analyses with clustered data can also be modeled in either the MEM or FEM framework as well as other frameworks (e.g., structural equation modeling, linear regression). A path diagram for this type of mediation model in the FEM framework is displayed in Figure 1. FEMs can only model mediation if all the variables are at Level 1. This would correspond to the so-called 1–1–1 mediation model, in which all variables of interest are at the lowest level of the hierarchy but are clustered within higher level units. Level 2 variables cannot be included in a FEM mediation model because they are collinear with the cluster affiliation variables (Hayes, 2013).

In Figure 1, the cluster affiliation dummy variables are predictors of both the mediator (motivation) and the outcome (productivity) because both serve as dependent variables in the system. The independent variable (job satisfaction) is an exogenous variable in the model (in that no arrows point to job satisfaction) and therefore need not be regressed on the cluster affiliation dummy variables. As noted in Hayes (2013), the FEM specification assumes that the effects of the coefficients do not vary across clusters. It is possible to create Cluster Affiliation × Predictor Variable interaction terms, as was discussed in the previous section, to overcome this, although the interpretation of the model would be similarly difficult. The intercepts of motivation and productivity must also be constrained to zero if using absolute coding in order to avoid overparameterizing the model. Alternatively, the intercepts could be retained if reference coding were used.

If relations beyond what can be tested in a 1–1–1 model are of interest, then one can use either a MEM framework (Krull & MacKinnon, 1999) or the closely related multilevel structural equation modeling (ML-SEM) framework (Kenny, Korchmaros, & Bolger, 2003; Preacher, Zyphur, & Zhang, 2010; Zhang, Zyphur, & Preacher, 2009). These frameworks allow researchers to test mediation involving Level 2 predictors, with the ML-SEM allowing for the most general models (Preacher et al., 2010). Importantly, it is easy to allow coefficients to differ across the

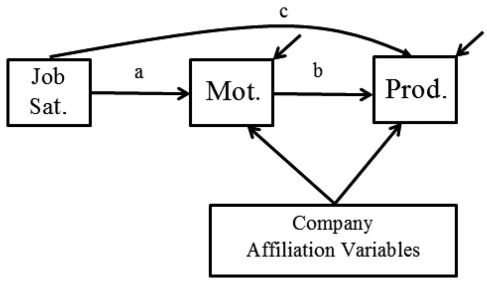

Figure 1. Path diagram for 1–1–1 mediation models using a fixed effects approach. The cluster affiliation variables are depicted as a single box—in the actual path diagram each cluster affiliation variable would be represented by its own box. a, b, and c = mediation paths; Job Sat. = Job Satisfaction; Mot. = Motivation; Prod. = Productivity.

different clusters in either framework by including random effects (Bauer, Preacher, & Gil, 2006; Preacher, 2011; Preacher, Curran, & Bauer, 2006). Using the hypothetical example from Figure 1, this means that the relation from job satisfaction to motivation could vary across clusters or that job satisfaction could be measured at the company level instead of the employee level.

Contextual Effects

The goal of a contextual effect analysis is to determine the extent to which a Level 1 effect differs from a Level 2 effect for the cluster mean of the same variable (Raudenbush & Bryk, 1986). From the definition alone, it may be clear that FEMs cannot accommodate this type of question because FEMs remove consideration of all specific Level 2 information from the model (Beck & Katz, 2001). In MEMs, contextual effects are estimated by including a Level 1 predictor in the model and then including the cluster mean for that same variable as a Level 2 predictor (e.g., Raudenbush & Bryk, 2002, p. 139). Depending on how the Level 1 variable is centered, the contextual effect is either captured by the Level 2 regression coefficient (with grand-mean centering) or the difference of the Level 2 regression coefficient and the Level 1 regression coefficient (with group-mean centering; e.g., Enders & Tofighi, 2007; Hofmann & Gavin, 1998; Kreft, de Leeuw, & Aiken, 1995). More information about contextual effect analysis can be found in Raudenbush and Bryk (2002, pp. 139–141), or in Feaster, Brincks, Robbins, and Szapocznik (2011) for a less technical introduction.

Nonpure Hierarchies and Multiple Levels of Clustering

In the MEM framework, nonpure hierarchies are straightforward to model with cross-classified models (see Grady & Beretvas, 2010, or Meyers & Beretvas, 2006, for more detail on cross-classified models). Because FEMs use dummy-coded cluster affiliation variables to account for clustering, cross-classification cannot be accommodated because cluster affiliation is not mutually exclusive (A. Bell & Jones, 2015).

Data with multiple levels of clustering (e.g., three-level models) are also common, including hierarchies such as employees/department/companies, people/neighborhoods/cities, and students/teachers/schools. In the MEM framework, additional levels can be added without much effort through the use of nested random effects (Bryk & Raudenbush, 1988). In a pure FEM, three or more levels of clustering cannot be included in a single model because there will not be sufficient degrees of freedom. In several contexts, the lack of generalizability of the FEMs to more than two levels can be problematic. In panel data, for instance, if employees are followed over time but are also clustered within companies or departments (e.g., repeated-measures/person/department), FEMs are not equipped to accommodate this three-level data structure.

However, FEMs and MEMs can be combined in the case of three or more levels of clustering (McNeish & Wentzel, 2017). This is most likely to be used when a researcher does not have theoretically meaningful predictors at the third level. In this case, a fixed effects approach can be used to account for the third level, whereas the first two levels are modeled with a MEM. This way, the Level 1 and Level 2 coefficients are estimated such that they control for all measured and unmeasured Level 3 variables without needing the less interesting level to be explicitly modeled. Higher levels of the hierarchy also tend to have smaller samples, which are less problematic for FEMs compared with MEMs (McNeish & Stapleton, 2016).

Endogeneity and Omitted Confounders

In econometrics, the primary argument for FEMs stems from concerns about the correct attribution of causality and the additional assumptions imposed by MEMs (Kim & Frees, 2006, 2007; Plümper & Troeger, 2007). Specifically, violations of the exogeneity assumption are the primary concern. Under such a violation, the model is said to be *endogenous.* Endogeneity refers to a non-null correlation between the errors (either the Level 2 random effects or the Level 1 residuals) and the outcome variable, which often arises from omitted confounders (e.g., Allison, 2009; Kim & Frees, 2006, 2007). More formally, endogeneity occurs when $Cov(X, u) \neq 0$ and/or $Cov(X, r) \neq 0$ using the notation defined earlier. In the MEM framework, endogeneity most often results from the omission of relevant Level 2 variables from the model (although endogeneity can be caused by other design aspects such as self-selection into treatment groups; e.g., Tofighi & Kelley, 2016). In such a case of omitted confounders, the errors in the model contain information from the omitted predictor. However, because both the omitted variable and the other predictors in the model are related to the outcome, the included predictors are no longer independent of the error term. More formally, for an omitted variable O, if $Cov(O,Y) \neq 0$ and $Cov(X, Y) \neq 0$, then $Cov(O, X) \neq 0$. Because the covariance of omitted variables is relegated to an error term, omission of a meaningful variable implies that $Cov(X, u) \neq 0$ or $Cov(X, r) \neq 0$, which is the definition of endogeneity. In the presence of endogeneity, the estimates of a MEM are no longer statistically consistent and coefficients can exhibit bias, even as sample size approaches infinity (Greene, 2003).

Although many studies in psychology are nonexperimental or quasi-experimental and strict causal claims may not be desired, statistical consistency of estimates remains a relevant consideration (Antonakis, 2017; Antonakis, Bendahan, Jacquart, & Lalive, 2010). Even in correlational research, reporting that X is related to Y contains some degree of causal implication (Antonakis et al., 2010), and threats to validity need to at least be considered. Antonakis, Bendahan, Jacquart, and Lalive (2014) go as far as to say "nonexperimental designs that do not address problems of endogeneity are pretty much useless for understanding a phenomenon" (p. 94). Although the merits of the aforementioned statement can be debated, the issue of endogeneity is important to all users of MEM, even when causal claims are not the primary interest.

Despite the vast emphasis placed on endogeneity in econometrics, endogeneity does not receive as much attention in the psychological tradition. To provide some anecdotal evidence, for publication years 2014 through 2016, 41 of the 199 articles (21%) that appeared in *Econometrica* discussed issues related to endogeneity. By contrast, during the same time frame in *Psychological Methods*, only one out of 97 articles (1%) discussed issues related to endogeneity (this lone article is DeMaris, 2014). In addition,

popular checklists written for researchers in the psychological tradition for conducting analyses with a MEM by Ferron et al. (2008), Dedrick et al. (2009), and Hox (2010) do not mention checking or accommodating issues related to exogeneity assumptions or omitted variables. Granted, many studies in psychology were historically based on controlled experiments, in which the risk of confounding variables is minimized. However, in areas of psychology that are more observational in nature, we believe that the researchers need to consider the deleterious effect that endogeneity can have on the validity of causal interpretations—an issue that has historically plagued econometricians who have a difficult time performing experimental studies in macro situations.

Inattention to the exogeneity assumption of MEMs, and the lack of familiarity with FEMs, may contribute to the pervasive use of MEMs in psychology and related fields (Kim & Frees, 2006, 2007). In studies grounded in psychology, for example, it is common to see researchers report that random intercepts included in a MEM account for all the variability at Level 2 (McNeish et al., 2017). However, random intercepts included in MEM account for all of the variability at Level 2 only when the exogeneity assumption is met. As aptly stated in Allison (2009),

> the key point here is that, contrary to popular belief, estimating a [mixed] effects model does not really "control" for unobserved heterogeneity. That's because the conventional [mixed] effects model assumes no correlation between the unobserved variables and the observed variables. (p. 22)

However, as pointed out in A. Bell and Jones (2015), there is a similar misconception about FEMs made by econometricians, namely, that if one wishes to protect against endogeneity from omitted variables at Level 2, then one must employ FEMs and thus lose the ability to estimate Level 2 coefficients in the process. As a further consequence, researchers lose the ability to address research questions and advanced modeling techniques that require these coefficients (note that FEMs are robust to endogeneity produced by omitted Level 2 variables but not necessarily endogeneity attributable to design issues). However, FEMs are not the only modeling option one can employ to address potential endogeneity attributable to Level 2 variables. As will be discussed in the next section, there are modeling strategies that provide the benefit of FEMs in accounting for possible endogeneity at Level 2 and also provide the benefit of MEMs in that Level 2 coefficients can be estimated.

Combining the Benefits of the Econometric and Psychological Traditions

The distinction between FEMs and MEMs sets up what is commonly referred to in econometrics as "the all or nothing" effect (e.g., Baltagi, 2013; Kim & Swoboda, 2010). MEMs allow researchers to flexibly model all the effects in which they are interested, but all relevant predictors must be included in the model to avoid endogeneity—a potentially daunting task in observational behavioral science and economic research. Conversely, FEMs are inflexible in that they do not allow for Level 2 predictors to be estimated, yet endogeneity at Level 2 will not be problematic. Researchers in either discipline may be slow to appreciate the advantages of the modeling approach taken by the other. Domain-specific preferences suggest that psychologists want Level 2 coefficient estimates and econometricians want protection from endogeneity. However, thinking of clustered data as a binary choice between MEMs or FEMs imposes a false dichotomy: There are gradations that exist between two extremes. We now present a model specification that addresses endogeneity while also allowing for Level 2 coefficients to be estimated, allowing researchers to break free from the false dichotomy imposed by "the all or nothing" effect.

To simultaneously model Level 2 coefficients and successfully address issues of endogeneity, we recommend that researchers use a within–between specification of a MEM (WB-MEM), an extension of Mundlak's (1978) specification. Although similar suggestions have recently been provided (e.g., A. Bell & Jones, 2015; Dieleman & Templin, 2014), the method has not been prominently utilized in analyses of empirical data. In the WB-MEM specification, the Level 1 predictors are group mean centered and the cluster mean of the Level 1 predictor is also included as a Level 2 predictor. Researchers in psychological or organizational research may recognize that this approach as an extension the process used to investigate contextual effects. Consider, again, the example of predicting productivity from motivation that has been used throughout this article. The WB-MEM specification for this model to protect against Level 2 endogeneity would be written

$$Productivity_{ij} = \beta_{0j} + \beta_{1j}(Motivation_{ij} - \overline{Motivation_j}) + r_{ij} \tag{11a}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\overline{Motivation_j} + u_{0j} \tag{11b}$$

$$\beta_{1j} = \gamma_{10} \tag{11c}$$

The Level 1 motivation predictor is included in the model as usual but the Level 2 cluster mean for motivation is then included as a Level 2 predictor. This specification results in the Level 1 coefficient estimates in which omitted variables at Level 2 are not a concern, provided that the model is properly specified at Level 1 (an assumption also made by FEMs). If multiple Level 1 predictors are of interest, then the corresponding cluster mean would be included as a Level 2 predictor of the intercept for *each* predictor. For example, if compensation were added to the model in Equation 11,

$$Productivity_{ij} = \beta_{0j} + \beta_{1j}(Motivation_{ij} - \overline{Motivation_j}) + \beta_{2j}(Compensation_{ij} - \overline{Compensation_j}) + r_{ij} \tag{12a}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\overline{Motivation_j} + \gamma_{02}\overline{Compensation_j} + u_{0j} \tag{12b}$$

$$\beta_{1j} = \gamma_{10} \tag{12c}$$

$$\beta_{2j} = \gamma_{20} \tag{12d}$$

Alhough the WB-MEM specification appears to be only a minor tweak to a standard MEM, setting the model up as a WB-MEM can have a profound impact on meeting the requirements of the exogeneity assumption.

In essence, the WB-MEM specification completely separates the estimation of within-cluster effects from between-cluster effects, a notable advantage, as this allows for the Level 2 effects to be modeled. This follows from the properties of the de-meaned FEM.

The logic of the WB specification is similar to the de-meaned specification in Equation 6—group-mean centering creates a within-cluster estimate of motivation that does not depend of between-cluster information which is absorbed in the Level 2 cluster mean predictor. Unlike the de-meaned model, the WB-MEM specification allows for estimates of coefficients for Level 2 variables as well as random effects on Level 1 coefficients. Put another way, effects of endogeneity manifest in MEMs because two processes—the effect being explicitly modeled and the implicit effect of omitted variables—are relegated to one parameter in the model (A. Bell & Jones, 2015). Splitting each Level 1 effect into within and between components allows the within component to be estimated irrespective of possible omitted Level 2 variables. The effect of Level 2 variables are completely consolidated into the between component (detailed statistical arguments for the effectiveness of this general strategy can be found in Mundlak (1978).

Provided that the model is properly specified at Level 1, the WB-MEM specification protects against bias from potential omitted Level 2 variables while also allowing for coefficients of Level 2 predictors to be directly estimated. That is, Level 2 predictors are no longer perfectly collinear with the mechanism that guards against omitted variable bias at Level 2. Thus, if researchers want to estimate effects of incentive (at the company level) and motivation (at the employee level) while also accounting for possibly omitted company-level variables (as in the FEM in Equation 4), this could be done with a WB-MEM specification as

$$Productivity_{ij} = \beta_{0j} + \beta_{1j}(Motivation_{ij} - \overline{Motivation_j}) + r_{ij} \tag{13a}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Incentive_j + \gamma_{02}\overline{Motivation_j} + u_{0j} \tag{13b}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Incentive_j \tag{13c}$$

Even though group-mean centering is widely used in psychology studies, Allison (2009) notes,

> Although it is well-known that group mean centering can produce substantially different results, [the mixed-effects model] literature has not made the connection to fixed effects models, nor has it been recognized that group mean centering controls for all time-invariant predictors. (p. 25)

Thus, even though the spirit of the WB-MEM specification is commonplace in contextual analyses, its utility is much broader in scope, such that it can be employed to address endogeneity issues attributable to omitted Level 2 variables.

Illustrative Example

To demonstrate the practical differences between the types of models we discussed, we present two examples. The first example features a panel analysis from Holzer, Block, Cheatham, and Knott (1993) examining the effect of training grants on worker efficacy across 54 manufacturing firms. This analysis will assume the effect for all time-varying covariates is the same across all firms (e.g., none of the Level 1 effects vary across the firms). The second example will use the same data set but will allow one of the covariates to vary randomly for each firm to explore how different methods are influenced by cluster-varying effects.

Holzer et al. (1993) followed 54 manufacturing firms in Michigan to investigate whether one-time training grants improved worker performance as defined by the scrap rates of products produced by the firms (scrap rate is measured once per year, resulting in three measures per firm). The outcome variable is the log of the scrap rate per 100 units, which is then modeled by whether the firm received a training grant in the current year (grant), whether the firm received a training grant the previous year (grant last year), the percentage of the firm's employees that received training (percent), and whether the firm's employees are unionized (union). Grant, grant last year, and percent are time-varying (Level 1) covariates, and union is a firm covariate (Level 2). Table 3 shows the model specification for three models used in each of these examples. In all models, the residual error structure is a homogeneous diagonal (i.e., $\sigma^2\mathbf{I}$), and for the firm-varying slope models, the random effects did not covary with the random intercepts for the MEM or WB-MEM specification.

No Firm-Varying Slopes

Table 4 shows the results for the three different model types with the predictors not being allowed to vary across firms. Generally, the results show that receiving a grant in the previous year and having a higher percentage of workers receiving training decreased the number of scrapped items. Although not egregious, the results do show some possible effects of endogeneity based on a comparison of coefficients across models. If all relevant Level 2 predictors were included, the effects of the traditional MEM would align with the WB-MEM specifications because the alternative specification would have no advantages, as the standard MEM would meet all requirements of the exogeneity assumption. In the MEM, the effect of union is positive but not significant. However, in the WB specification, the effect of union is notably larger (0.58 vs. 0.76, a 30% increase), though the effect remains nonsignificant and may be due to sampling error. As expected, the time-varying effects are identical between the FEM and WB-MEM specifications, demonstrating that both specifications account for endogeneity. As an added benefit, the WB-MEM specification allows researchers to estimate and test the union effect, which is not possible in the FEM (although the FEM does control for union).

With Firm-Varying Slopes

Table 5 shows the results for the models allowing the effect of percent to vary across firms. In the MEM and WB specifications, the variance for this random effect was statistically significant using a 50:50 mixture chi-square test, as recommended by Verbeke and Molenberghs (2003). The omnibus test for the Percent × Firm interaction was also significant in the FEM. In Table 5, the comparison of the results for MEM and WB specifications follows a similar pattern to Table 4: The union effect is noticeably different between specifications. The union effect in the MEM is smaller and not statistically significant, whereas the effect is notably larger and statistically significant in the WB model. A similar pattern is found with the percent predictor. In this model, the FEM is estimating 110 parameters (two time-varying covariates, 54 firm intercepts, 54 percent slopes) from 157 total observations. The

Table 3
Model Equations for the Holzer et al. (1993) Example Analyses

Model type	No firm-varying slopes	With firm-varying slopes
MEM	$Log(Scrap\ Rate) = \beta_0 + \beta_1 Grant_{ij} + \beta_2 Grant\ LY_{ij} + \beta_3 Percent_{ij} + r_{ij}$ $\beta_0 = \gamma_{00} + \gamma_{01} Union_j + u_{0j}$ $\beta_1 = \gamma_{10}$ $\beta_2 = \gamma_{20}$ $\beta_3 = \gamma_{30}$	$Log(Scrap\ Rate) = \beta_0 + \beta_1 Grant_{ij} + \beta_2 Grant\ LY_{ij} + \beta_3 Percent_{ij} + r_{ij}$ $\beta_0 = \gamma_{00} + \gamma_{01} Union_j + u_{0j}$ $\beta_1 = \gamma_{10}$ $\beta_2 = \gamma_{20}$ $\beta_3 = \gamma_{30} + u_{3j}$
FEM	$Log(Scrap\ Rate) = \beta_1(Grant_{ij}) + \beta_2(Grant\ LY_{ij}) + \beta_3(Percent_{ij}) + \alpha_1 Firm_1 + \ldots + \alpha_{54} Firm_{54} + r_{ij}$	$Log(Scrap\ Rate) = \beta_1(Grant_{ij}) + \beta_2(Grant\ LY_{ij}) + \alpha_1 Firm_1 + \ldots + \alpha_{54} Firm_{54} + \gamma_1(Firm_1 \times Percent_{ij}) + \ldots + \gamma_{54}(Firm_{54} \times Percent_{ij}) + r_{ij}$
WB-MEM	$Log(Scrap\ Rate) = \beta_0 + \beta_1(Grant_{ij} - \overline{Grant_j}) + \beta_2(Grant\ LY_{ij} - \overline{Grant\ LY_j}) + \beta_3(Percent_{ij} - \overline{Percent_j}) + r_{ij}$ $\beta_0 = \gamma_{00} + \gamma_{01} Union_j + \gamma_{02} Grant_j + \gamma_{03} \overline{Grant\ LY_j} + \gamma_{04} Percent_j + u_{0j}$ $\beta_1 = \gamma_{10}$ $\beta_2 = \gamma_{20}$ $\beta_3 = \gamma_{30}$	$Log(Scrap\ Rate) = \beta_0 + \beta_1(Grant_{ij} - \overline{Grant_j}) + \beta_2(Grant\ LY_{ij} - \overline{Grant\ LY_j}) + \beta_3(Percent_{ij} - \overline{Percent_j}) + r_{ij}$ $\beta_0 = \gamma_{00} + \gamma_{01} Union_j + \gamma_{02} \overline{Grant_j} + \gamma_{03} Grant\ LY_j + \gamma_{04} Percent_j + u_{0j}$ $\beta_1 = \gamma_{10}$ $\beta_2 = \gamma_{20}$ $\beta_3 = \gamma_{30} + u_{3j}$

Note. MEM = mixed effects model; FEM = fixed effects model; WB-MEM = MEM with within-between specification; LY = last year.

model can be fit, but the model is heavily parameterized (only 47 degrees of freedom remain from 157 observations), which is reflected in the instability of the estimates, which depart from the estimates of the other models.

Discussion

Readers may note that the within–between specification possesses the advantages of FEMs while maintaining the flexibility of MEMs. This begs the question, when should researchers use standard FEMs or MEMs? A. Bell and Jones (2015) addressed this via theoretical and simulation-based arguments and found that the WB specification is at least as good as the FEM with respect to estimating coefficients in the presence of endogeneity due to omitted variables. Bell and Jones also found that the WB specification often outperformed the FEM in other scenarios such as unbalanced clusters. As noted earlier in this article, FEMs can be more cumbersome to specify for more complex analyses for which automated procedures do not always exist (e.g., FEM regressions in Stata are straightforward to specify but a fixed effects mediation model in M*plus* is not) or are unable to address common types of data structures and research questions that MEMs can handle with relative ease. By using a WB specification, the flexibility of MEMs can be realized while

Table 4
Holzer et al. (1993) Results for Model With No Firm-Varying Slopes

	MEM		FEM		WB-MEM	
Predictor	Est.	*p*	Est.	*p*	Est.	*p*
Time-varying predictors						
Intercept	.58	—	—	—	−.27	—
Grant	−.08	.65	−.07	.66	−.07	.66
Grant LY	−.63	<.01	−.65	<.01	−.65	<.01
Percent	−.56	<.01	−.60	<.01	−.60	<.01
Firm-level predictors						
Unionized	.58	.18	—	—	.76	.07
Grant mean	—	—	—	—	2.24	.16
Grant LY mean	—	—	—	—	−1.14	.49
Percent mean	—	—	—	—	.37	.54
Variance components						
Var(Int)	2.02		—		1.76	
Var(Residual)	.23		.23		.23	

Note. MEM = mixed effects model; FEM = fixed effects model; WB-MEM = MEM with within-between specification; Est. = estimate; LY = last year.

Table 5
Holzer et al. (1993) Results for Model With Firm-Varying Slopes

	MEM		FEM		WB-MEM	
Predictor	Est.	*p*	Est.	*p*	Est.	*p*
Time-varying predictors						
Intercept	.52	—	—	—	−.22	—
Grant	−.29	.06	−.41	.58	−.28	.07
Grant LY	−.72	<.01	−.86	<.01	−.74	<.01
Percent	−.41	.07	−.50	.04	−.46	.05
Firm-level predictors						
Unionized	.63	.13	—	—	.78	.05
Grant mean	—	—	—	—	1.98	.20
Grant LY mean	—	—	—	—	−1.03	.52
Percent mean	—	—	—	—	.28	.65
Variance components						
Var(Int)	1.71		—		1.55	
Var(Percent)	.64		—		.62	
Var(Residual)	.16		.16		.16	

Note. MEM = mixed effects model; FEM = fixed effects model; WB = MEM with within-between specification; Est. = estimate; LY = last year.

simultaneously estimating Level 1 effects that are protected from threats of omitted variables, at least to the extent that such protection is afforded by FEMs. With the WB specification, researchers do not have to consciously disregard aspects like cluster-varying effects or mediation hypotheses—these can be framed within the WB specification. More directly, MEMs can very often be specified in a way that can encompass the benefits of FEMs while also maintaining the flexibility to assess more types of and more complex research interests.

If researchers are weary of the assumptions made by including random effects, Dieleman and Templin (2014) also note that the WB specification does not necessarily require that clustering be accommodated with MEM—the WB specification can be adapted for use with cluster-robust errors or generalized estimating equations, such that omitted confounder bias can be similarly guarded against if clustering is accommodated with design-based methods. As has been argued (e.g., Beck & Katz, 2001; A. Bell & Jones, 2015; Castellano, Rabe-Hesketh, & Skrondal, 2014), the information that is being discarded in FEMs can often explain phenomena of interest; indeed, an MEM approach would allow a richer set of questions to be addressed. It is unwise to merely control for the entirety of Level 2 when there are straightforward ways to incorporate such information into the model and preserve access to research questions addressing Level 2 effects.

As studies continue to stray from the classic controlled experiment and as nonrandom samples continue to be collected, endogeneity increasingly threatens the validity of claims made by researchers. Given the increasingly interwoven and complex processes that are modeled in modern psychological research, omitted confounders are a genuine concern and researchers from the psychological tradition would be wise to adopt some econometric tools and concerns. Altering MEMs ever so slightly so that they form a WB-MEM specification could go a long way to address the impact of endogeneity issues have had on models for psychological data. Additionally, such a specification would allow more flexibility for econometrically grounded researchers accustomed to employing the FEM framework. This type of cross-fertilization of methods presents advantages for a wide variety of researchers, similar to the emergence of structural equation modeling framework that arose from blending structural models from economics with factor models from psychology, forming a stronger overall modeling framework.

Despite the high praise given to the WB-MEM specification, it is not a universal solution for modeling clustered data. As noted in McNeish and Stapleton (2016), the performance of MEMs, including the WB specification, are adversely affected by small samples. Both A. Bell and Jones (2015) and McNeish and Stapleton (2016) recommended a FEM when the number of Level 2 units is small. In such cases, the data are often not sufficiently rich to support the more ambitious WB-MEM specification (Maas & Hox, 2005; Stegmueller, 2013). The random effects in a WB-MEM also assume random sampling of clusters. Should this assumption be inappropriate, the assumption could be circumvented with an FEM because no such assumption is required in the FEM framework. Alternatively, design-based methods like generalized estimating equations can accommodate a WB specification that preserves Level 2 predictors without modeling with random effects. The WB-MEM specification also does not solve *all* issues with endogeneity. If researchers are trying to estimate treatment effects, endogeneity can occur when treatment and control groups are nonrandom or self-selected (e.g., only poorly performing employees are eligible for selection into the treatment group). Endogeneity, in this case, is attributable to properties of the research design, measurement errors, and/or the data collection, not due to the way the data are modeled (A. Bell & Jones, 2015; Kennedy, 2008; Li, 2011). Neither the WB specification nor the FEM address endogeneity attributable to the research design, and any estimated treatment effects using these methods would still be likely to contain bias—WB-MEM and FEMs primarily address endogeneity introduced from omission of Level 2 variables. For endogeneity of this type, researchers could condition on covariates (e.g., ANCOVA), use *propensity score methods*, or use *instrumental variables* to accommodate the endogeneity present in the variables themselves. Consistent with the overall theme of this article, the instrumental variables approach is another method that is well known by researchers trained in the econometric tradition but is not used by many researchers with backgrounds in psychology. The details concomitant with instrumental variables is beyond the scope of this article, so we will not address this method here—for more information, see DeMaris (2014) or Foster and McLanahan (1996) for studies grounded in psychology. For more general treatments, see Angrist and Krueger (2001), Greenland (2000), or Heckman (1997).

In conclusion, both MEMs and FEMs have advantages that can provide unique benefits in the context of modeling clustered data, whereas the WB-MEM specification is an option that can often capitalize on advantages of both methods. Our goal has not been to show either method in a negative light but rather to bridge the two methods together to show how they can complement each other. As the line distinguishing psychology from economics continues to be blurred, it is beneficial for both types of researchers to leverage the best methodologies available across both fields. Different disciplines often emphasize different aspects of the same type of analyses, so integrating these different perspectives can only serve to strengthen analyses in all disciplines. Sometimes, the lack of awareness to a certain approach is not because it is not appropriate but simply because it is not discussed in key training materials. We hope this article helps to facilitate more understanding and awareness of the available models.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*. Advance online publication. http://dx.doi.org/10.1037/amp0000190

Allison, P. D. (2009). *Fixed effects regression models*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412993869

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives, 15,* 69–85.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly, 28,* 5–21.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21,* 1086–1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford handbook of leadership and organizations* (pp. 93–117). New York, NY: Oxford University Press.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18,* 151–164. http://dx.doi.org/10.1037/a0030642

Baltagi, B. (2013). *Econometric analysis of panel data*. New York, NY: Wiley.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/book/

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11,* 142–163. http://dx.doi.org/10.1037/1082-989X.11.2.142

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods, 16,* 373–390. http://dx.doi.org/10.1037/a0025813

Beck, N., & Katz, J. N. (2001). Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon. *International Organization, 55,* 487–495. http://dx.doi.org/10.1162/00208180151140658

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods, 3,* 133–153. http://dx.doi.org/10.1017/psrm.2014.7

Bell, B., Morgan, G., Schoeneberger, J. B., Kromrey, J., & Ferron, J. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10,* 1–11. http://dx.doi.org/10.1027/1614-2241/a000062

Bollen, K., & Curran, P. (2006). *Latent curve models: A structural equation perspective*. New York, NY: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97,* 65–108. http://dx.doi.org/10.1086/443913

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39,* 333–367. http://dx.doi.org/10.3102/1076998614547576

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods, 3,* 399–408.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., . . . Lee, R. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79,* 69–102. http://dx.doi.org/10.3102/0034654308325581

DeMaris, A. (2014). Combating unmeasured confounding in cross-sectional studies: Evaluating instrumental-variable and Heckman selection models. *Psychological Methods, 19,* 380–397. http://dx.doi.org/10.1037/a0037416

Dieleman, J. L., & Templin, T. (2014). Random-effects, fixed-effects and the within-between specification for clustered data in observational health studies: A simulation study. *PLoS ONE, 9,* e110257. http://dx.doi.org/10.1371/journal.pone.0110257

Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics, 3,* 365–392. http://dx.doi.org/10.1007/s11129-005-1177-6

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Feaster, D., Brincks, A., Robbins, M., & Szapocznik, J. (2011). Multilevel models to identify contextual effects on individual group member outcomes: A family example. *Family Process, 50,* 167–183. http://dx.doi.org/10.1111/j.1545-5300.2011.01353.x

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41,* 372–384. http://dx.doi.org/10.3758/BRM.41.2.372

Ferron, J., Hogarty, K., Dedrick, R., Hess, M., Niles, J., & Kromrey, J. (2008). Reporting results from multilevel analyses. In A. O'Connell & B. McCoach (Eds.), *Multilevel analysis of educational data* (pp. 391–426). Charlotte, NC: Information Age Publishing.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods, 1,* 249–260. http://dx.doi.org/10.1037/1082-989X.1.3.249

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine, 28,* 221–239. http://dx.doi.org/10.1002/sim.3478

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1,* 515–534. http://dx.doi.org/10.1214/06-BA117A

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, MA: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511790942

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73,* 43–56. http://dx.doi.org/10.1093/biomet/73.1.43

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika, 76,* 622–623. http://dx.doi.org/10.1093/biomet/76.3.622

Grady, M. W., & Beretvas, S. N. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivariate Behavioral Research, 45,* 393–419. http://dx.doi.org/10.1080/00273171.2010.483390

Greene, W. (2003). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology, 29,* 722–729. http://dx.doi.org/10.1093/ije/29.4.722

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30,* 507–544.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320–338. http://dx.doi.org/10.1080/01621459.1977.10480998

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources, 32,* 441–462. http://dx.doi.org/10.2307/146178

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24,* 623–641. http://dx.doi.org/10.1177/014920639802400504

Holzer, H. J., Block, R. N., Cheatham, M., & Knott, J. H. (1993). Are training subsidies for firms effective? The Michigan experience. *Indus-*

trial & Labor Relations Review, 46, 625–636. http://dx.doi.org/10.1177/001979399304600403

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6,* 87–93.

Hsiao, C. (2003). *Analysis of panel data*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511754203

Huang, F. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *Journal of Experimental Education, 84,* 175–196. http://dx.doi.org/10.1080/00220973.2014.952397

Judge, G. G., Griffiths, W. E., Hill, R. C., & Lee, T. C. (1985). *The theory and practice of econometrics*. New York, NY: Wiley.

Kennedy, P. (2008). *A guide to econometrics*. Malden, MA: Blackwell.

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods, 8,* 115–128. http://dx.doi.org/10.1037/1082-989X.8.2.115

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53,* 983–997. http://dx.doi.org/10.2307/2533558

Kim, J., & Frees, E. W. (2006). Omitted variables in multilevel models. *Psychometrika, 71,* 659–690. http://dx.doi.org/10.1007/s11336-005-1283-0

Kim, J., & Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika, 72,* 505–533. http://dx.doi.org/10.1007/s11336-007-9008-1

Kim, J. S., & Swoboda, C. M. (2010). Handling omitted variable bias in multilevel models: Model specification tests and robust estimation. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 197–217). New York, NY: Routledge.

Kreft, I. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21. http://dx.doi.org/10.1207/s15327906mbr3001_1

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review, 23,* 418–444. http://dx.doi.org/10.1177/0193841X9902300404

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963–974. http://dx.doi.org/10.2307/2529876

Li, X. (2011). *Approaches to modelling heterogeneity in longitudinal studies*. Wellington, New Zealand: Victoria University.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13–22. http://dx.doi.org/10.1093/biomet/73.1.13

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 86–92. http://dx.doi.org/10.1027/1614-2241.1.3.86

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective*. New York, NY: Routledge.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23,* 750–773. http://dx.doi.org/10.1080/10705511.2016.1186549

McNeish, D. (2017a). Challenging conventional wisdom for multivariate statistical models with small samples. *Review of Educational Research, 87,* 1117–1151. http://dx.doi.org/10.3102/0034654317727727

McNeish, D. (2017b). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52,* 661–670. http://dx.doi.org/10.1080/00273171.2017.1344538

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51,* 495–518. http://dx.doi.org/10.1080/00273171.2016.1167008

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22,* 114–140. http://dx.doi.org/10.1037/met0000078

McNeish, D., & Wentzel, K. R. (2017). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research, 52,* 200–215. http://dx.doi.org/10.1080/00273171.2016.1262236

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41,* 473–497. http://dx.doi.org/10.1207/s15327906mbr4104_3

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics, 32,* 385–397. http://dx.doi.org/10.1016/0304-4076(86)90021-7

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics, 72,* 334–338. http://dx.doi.org/10.2307/2109724

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46,* 69–85. http://dx.doi.org/10.2307/1913646

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58,* 545–554. http://dx.doi.org/10.1093/biomet/58.3.545

Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies, 22,* 435–480. http://dx.doi.org/10.1093/rfs/hhn053

Plümper, T., & Troeger, V. E. (2007). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis, 15,* 124–139. http://dx.doi.org/10.1093/pan/mpm002

Preacher, K. J. (2011). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research, 46,* 691–731. http://dx.doi.org/10.1080/00273171.2011.589280

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31,* 437–448. http://dx.doi.org/10.3102/10769986031004437

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15,* 209–233. http://dx.doi.org/10.1037/a0020141

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. Berkeley, CA: Stata Press.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1–17. http://dx.doi.org/10.2307/2112482

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Schurer, S., & Yong, J. (2012). *Personality, well-being and the marginal utility of income: What can we learn from random coefficient models?* Health, Econometrics and Data Group (HEDG) Working Papers. Department of Economics, University of York, York, United Kingdom.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195152968.001.0001

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis*. London, UK: Sage.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science, 57,* 748–761. http://dx.doi.org/10.1111/ajps.12001

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40,* 961–971. http://dx.doi.org/10.2307/2531147

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50,* 1171–1177. http://dx.doi.org/10.2307/2533455

Tofighi, D., & Kelley, K. (2016). Assessing omitted confounder bias in multilevel mediation models. *Multivariate Behavioral Research, 51,* 86–105. http://dx.doi.org/10.1080/00273171.2015.1105736

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics, 59,* 254–262. http://dx.doi.org/10.1111/1541-0420.00032

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *The American Economic Review, 93,* 133–138. http://dx.doi.org/10.1257/000282803321946930

Yang, Y., & Land, K. C. (2008). Age -period -cohort analysis of repeated cross-section surveys: fixed or random effects? *Sociological Methods & Research, 36,* 297–326.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models problems and solutions. *Organizational Research Methods, 12,* 695–719. http://dx.doi.org/10.1177/1094428108327450

Received January 17, 2017
Revision received January 5, 2018
Accepted January 25, 2018 ■