Psychological Methods

A Practical Guide to Selecting and Blending Approaches for Clustered Data: Clustered Errors, Multilevel Models, and Fixed-Effect Models

Daniel McNeish

Online First Publication, November 13, 2023. https://dx.doi.org/10.1037/met0000620

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000620

A Practical Guide to Selecting and Blending Approaches for Clustered Data: Clustered Errors, Multilevel Models, and Fixed-Effect Models

Daniel McNeish
Department of Psychology, Arizona State University

Abstract

Psychological data are often clustered within organizational units, which violates the independence assumption in standard regression models. Clustered errors, multilevel models, and fixed-effects models all address this issue, but in different ways. Disciplinary preferences for approaching clustered data are strong, which can restrict questions researchers ask because certain approaches are better equipped to handle particular types of questions. Resources comparing approaches to facilitate broader understanding of clustered data approaches exist for economists, political scientists, and biostatisticians. These existing resources use concepts and terminology consistent with statistical training in other disciplines, so this article provides a resource using language and principles familiar to psychologists. The article starts by walking through the origin and importance of the independence assumption to motivate the problem and emergence of different solutions in different fields. Then, information on clustered errors, multilevel models, and fixed-effect models is provided, including (a) how each approach addresses independence violations, (b) research questions ideally suited for each approach, and (c) example analyses highlighting advantages and disadvantages. The article then discusses how these approaches are not mutually exclusive but instead can be blended together to create tailor-made models that flexibly accommodate idiosyncrasies in research questions and are robust to nuances of a particular data set. The broader theme is that there is no one-size-fits-all approach to clustered data. The research question—not disciplinary preferences—should inform the statistical approach. Wider appreciation of the landscape of clustered data approaches can expand the questions researchers ask and improve the theoretical foundation of statistical models.

Translational Abstract

It is common for behavioral data to be structured such that people are members of some larger organizational unit (e.g., schools, companies, hospitals), which violates assumptions of basic statistical methods and can affect the accuracy of conclusions drawn from statistical models. A few advanced statistical methods have been developed to address this issue, but different academic disciplines have strong preferences for which method is commonly used in research studies. That is, even though the underlying nature of the problem is the same in different disciplines, the approach taken by economists tends to look different from a medical researcher or epidemiologist which tends to be different from a psychologist or educational researcher. This article starts by walking through the fundamental statistical problem that organizationally clustered data structures present and how different advanced statistical methods work to address the problem in different ways. Specifically, the article focuses on how each method can have advantages and disadvantages for answering certain types of questions through statistical analysis, which contributes to the disparity in preferences among different disciplines because different types of questions are asked in different disciplines. However, as the article discusses, aspects of these three separate methods can be integrated to engineer statistical models that combine advantages of multiple methods simultaneously within a single model. The general theme of the article is that scientific research disciplines tend to be insular but that much can be gained from looking outward to other disciplines that may approach the same problem from a unique perspective.

Keywords: hierarchical linear model, mixed effect model, linear mixed model, cluster robust errors, sandwich estimator

Supplemental materials: https://doi.org/10.1037/met0000620.supp

Clustered data are typical in psychology with common examples being students clustered within schools, people clustered within neighborhoods, patients clustered within treatment centers, or employees clustered within teams (e.g., Hox et al., 2017; Raudenbush & Bryk, 2002). Clustered data violate independence assumptions made by traditional methods within the family of generalized linear models (e.g., regression and analysis of variance [ANOVA]), and specialized methods are required to accommodate clustered data. Multilevel models

Daniel McNeish https://orcid.org/0000-0003-1643-9408
The Open Science Framework link is available at https://osf.io/w4x9n/.

Correspondence concerning this article should be addressed to Daniel McNeish, Department of Psychology, Arizona State University, PO Box 871104, Tempe, AZ 85287, United States. Email: dmcneish@asu.edu

have historically been the most frequent approach in psychology to account for unique aspects of clustered data (e.g., Bauer & Sterba, 2011; Huang, 2016), although increases in multidisciplinary research have increased exposure to methods popular in other fields.

Alternatives to multilevel models like clustered errors or fixed-effect models were developed outside of psychology, so many existing resources for these approaches appear in resources targeting economists (Cameron & Miller, 2015; Cameron et al., 2008; Primo et al., 2007), political scientists (A. Bell & Jones, 2015; A. Bell et al., 2019; Clark & Linzer, 2015), or biostatisticians (Dieleman & Templin, 2014; Gardiner et al., 2009; Hubbard et al., 2010). Treatments targeting psychologists have emerged (Huang, 2016, 2018; McNeish et al., 2017; McNeish & Kelley, 2019), although these resources tend to serve as introductions to each method rather than addressing broader practical questions of how to decide among different methods based on the research question of interest.

This article therefore has two primary goals. The first is to provide researchers in psychology and adjacent areas with a practical guide on how different methods for clustered data map onto different research interests. The intention is to minimize reliance on equations and mathematical exposition given that technical comparative treatments exist (e.g., Gardiner et al., 2009). Where technical detail is provided, it is intended to be supplemental so that broader conceptual ideas remain comprehensible. The second goal is to show how these three methods comprise a broader framework for modeling clustered data and how aspects of different methods can be blended to create more flexible, robust, and comprehensive models that directly and efficiently address the researchers' specific questions. To keep the manuscript streamlined, the focus is primarily on clustering due to shared organizational membership, but special considerations for longitudinal data and complex data structures (e.g., three-level or cross-classified hierarchies) are included at the end of the article.

To outline the structure of the manuscript, the origin of the independence assumption in linear regression is reviewed followed by a discussion of how clustered data violate the independence assumption and the associated statistical ramifications. Three different approaches for clustered data that address independence assumption violations are discussed: clustered errors (a.k.a. cluster robust errors or sandwich estimators), multilevel modeling (a.k.a. mixed-effect models, random effect models, hierarchical linear models), and fixed-effect models. Sections on each model discuss (a) how the model addresses independence violations, (b) what types of research questions are best addressed by the method, (c) an example empirical analysis addressing a relevant hypothetical research question, and (d) a brief overview of advanced topics to consider with each method. Subsequently, strategies for combining aspects of these three methods into one model are discussed to create models that are more robust and tailor-made to suit the needs of the data structure and the research questions.

Origin of Regression Assumptions

Consider a standard linear regression for a single continuous outcome

$$\overset{\text{Outcome}}{\mathbf{y}} = \overset{\text{Predictors}}{\mathbf{X}} \times \overset{\text{RegressionCoefficients}}{\boldsymbol{\beta}} + \overset{\text{Errors}}{\mathbf{e}}, \tag{1}$$

where $\mathbf{y}$ is a vector of the outcome variable (each row is the outcome for a different person), $\mathbf{X}$ is a matrix of predictor variables (each row is a person, each column is a predictor variable), $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\mathbf{e}$ is a vector of errors representing the difference between observed ($\mathbf{y}$) and predicted ($\hat{\mathbf{y}}$) values. There are three assumptions associated with a standard linear regression model: (a) normality of the errors, (b) independence of the errors, and (c) homoskedasticity of the errors. These three assumptions can be written concisely as $\mathbf{e} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ where "i.i.d" stands for "independently and identically distributed" and σ^2 is the variance of the error distribution.

Note that all three of these assumptions concern the *error* term. An underappreciated fact among psychologists is that these assumptions are only required to obtain accurate standard errors and to make inferences about regression coefficients (e.g., testing whether the population value equals 0). They are *not* required to estimate the regression coefficients accurately.

Specifically, using ordinary least squares, regression coefficients that minimize the vertical distance between data points and the regression line are estimated by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The mechanics of this equation are less important than the fact that the errors ($\mathbf{e}$) are not present; only the matrix of predictors ($\mathbf{X}$) and the vector of the outcome variable ($\mathbf{y}$) are needed to estimate the regression coefficients. Normality, homoskedasticity, and independence are not required to if inference or hypothesis testing are not of interest.

Assumptions are necessary for computing standard errors due to an incongruence with what the standard errors are trying to quantify and how studies are typically conducted. To explain, imagine an example where the goal is to predict test scores from SES. If a researcher conducted their study one time and fit a linear regression to their data, they would get a single coefficient for the effect of SES on test scores. Now, imagine that they collect test score and SES values from a new random sample from the same population and fit the same linear regression model to the new data. The regression coefficient estimate for SES may now be different because there is sampling variability due to imprecision from using a sample of the broader population. Finally, consider that this researcher is independently wealthy or has generous grant funding and that they can conduct their study on 100 different random samples from the target population, yielding 100 different estimated values for the SES effect. Hypothetical visual representations of these scenarios are shown in Figure 1a.

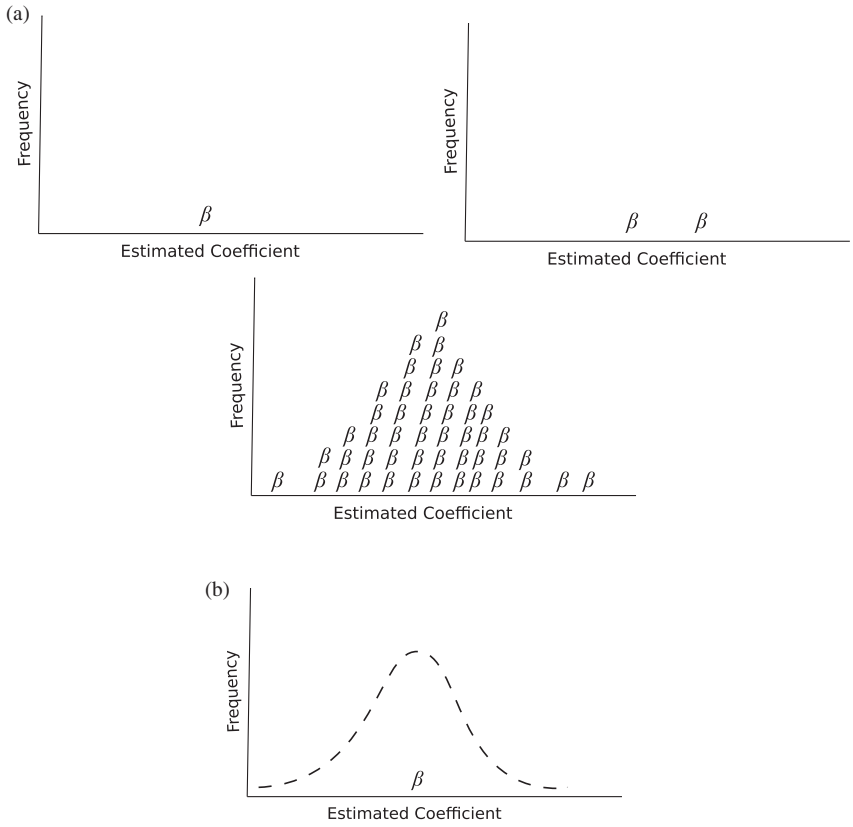
With coefficient estimates from many samples, it is possible to form a sampling distribution of estimated regression coefficients (as in the bottom panel of Figure 1a). To evaluate variability of the coefficient estimates across samples, the descriptive variance or standard deviation of all these coefficients could be taken to quantify the precision of coefficient estimates between samples. To make inferences about the coefficients, the probability of obtaining a particular estimate from a population whose true value were 0 based on this sampling distribution could be calculated (e.g., is 0 a plausible or unrealistic population value?).

No part of this process requires assumptions about the errors. If the study could be repeatedly conducted, the sampling variability could be quantified using only the repeatedly estimated regression coefficients across different samples, which rely solely on information from the data ($\mathbf{X}$ and $\mathbf{y}$) without any need to assume a particular distribution for or independence of the errors.

However, in practice, most analyses and inferences are conducted on data from a single sample, so the situation most closely resembles the upper left panel of Figure 1a. The challenge is to estimate sampling variability from a single value that has no descriptive variance. That is, it is hard to make inferences about whether a coefficient is plausibly 0 without an idea of the precision of the estimate.

Figure 1

Regression Coefficient Sampling Variability via Repeated Sampling or a Single Sample With Assumptions



Note. (a) Hypothetical histogram of estimated regression coefficient for one sample (upper left), two samples (upper right), and many samples (bottom). (b) Estimated coefficient from a single sample with the theoretical distribution derived conditional on assumptions shown as a dashed line.

This is where assumptions enter the picture. By assuming normality, homoskedasticity, and independence (of the *errors*), it is possible to construct a *theoretical* sampling distribution around a single coefficient estimate. This is shown in Figure 1b where the dashed line is the theoretical sampling distribution under particular assumptions. The standard deviation of the theoretical sampling distribution—referred to as the *standard error*—is used as an estimate of the sampling standard deviation. In other words, the standard error tries to capture what the standard deviation of the sampling distribution would have looked like if it were possible to repeatedly conduct the study. Essentially, the tradeoff is that assumptions are made to obtain information that could otherwise only be obtained by (cost-prohibitively) conducting the study repeatedly with many different random samples.

Issues With Clustered Data

Statistical Issues With Clustered Data

The independence assumption implies that each observation brings 100% unique information to the analysis, which is used (in part) to determine the width of the theoretical sampling distribution (the dashed distribution in Figure 1b). When data are clustered such that observations have some shared context through membership in the same organizational unit like a school or work team, independence can be a tenuous assumption. For example, test scores for students who attend the same school likely do not contain 100% unique information because students share experiences that influence test scores like teachers, curriculums, and peers. Observations in clustered data therefore contain a mix of information about the individual (the student) and the shared context (the school).

In clustered data, information about the shared context is common among all members of the organizational unit. However, assuming independence treats all information as unique such that any shared information is *double-counted* when computing standard errors, which makes the data appear to have more information than they actually contain. When computing standard errors, the precision of the theoretical sampling distribution will therefore be inaccurate. Violating independence typically overestimates precision such that standard errors are too small (i.e., the distribution in Figure 1b is too narrow). Wald tests used in inference divide a coefficient estimate by its standard error. So, if standard errors are too small, test statistics like t or Z will be too large and p-values will be too small. As a result, Type-I error rates are inflated and inference errors occur more often than the prescribed rate.

Diagnosing Independence Assumption Violations

The *intraclass correlation* (ICC) is a measure between 0 and 1 that quantifies what proportion of the outcome is attributable to the shared context, which helps quantify the extent of the double counting (0 means no information is double-counted and observations are independent; 1 means that all information is double-counted and observations are entirely dependent). Figure 2 shows a graphical depiction of this idea for two individuals within the same cluster. Each circle represents sources of variance in the outcome variable for one person; each person brings some unique individual information (the white space inside each circle) but there is also some shared contextual information represented by the grey shaded area. As an example, if the outcome is a test score, there are some individual sources that contribute to scores (e.g., study habits, motivation to learn), but there are also school sources (e.g., teacher quality, course offerings) such that scores from students in the same school will be related to some degree. The ICC aims to determine the proportion occupied by the grey area in Figure 2.

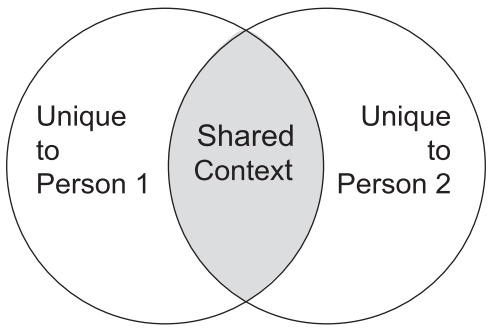
The ICC is often interpreted as a test of whether there will be statistical issues if clustering were ignored (i.e., if standard errors will be too small). However, the proportion of variance attributable to a shared context is distinct from whether standard errors will be accurate. The square root of the *design effect* (DEFT; Kish, 1965) is more appropriate for evaluating if clustering will affect standard errors. The design effect (DEFF; Hox et al., 2017, p. 5) is a ratio of the sampling variability when treating data as clustered versus treating data as independent. The DEFT takes the square root of this ratio to place the metric onto the same metric as the standard error, which can be more interpretable (Kish, 1995, p. 56). The DEFT is therefore a measure of how much larger standard errors will be when data are treated as clustered compared to treating the data as independent.

Specifically, DEFT $= \sqrt{1 + (m - 1) \times \text{ICC}}$ where m is the average number of observations in a cluster. For instance, a DEFT of 2 suggests that standard errors would be twice as large if clustering were acknowledged than if independence were assumed (i.e., the DEFT is a multiplicative term). If the total sample size is divided by the DEFT squared (i.e., N/DEFT^2), the result is the *effective sample size*, which is the number of "independent-equivalent" observations contained within a clustered sample (Kish, 1965). So a data set where $N = 200$ and DEFT $= 2$ contains the same amount of unique information as a data set with $(200/2^2) = 50$ independent observations. A historical threshold for when clustering becomes problematic is DEFT $> \sqrt{2}$ (Muthen & Satorra, 1995), although recent studies find that informative DEFT values can depend on several characteristics (Lai & Kwok, 2015).

The DEFT is related to the ICC but it also incorporates cluster size. The distinction is that the ICC is quantifying the average shared information (the *average* grey area in a Venn diagram) whereas the DEFT is quantifying the extent of double counting shared information (the *sum* of all grey areas in a Venn diagram). A small ICC does not necessarily mean that clustering is ignorable because double counting may add up and become problematic with large clusters.

This process is depicted in Figures 3 and 4. Double counting many small overlapping Venn diagram sections (as in Figure 3) is just as problematic for standard errors and p-values as double counting a few large overlapping Venn diagram sections (as in Figure 4) because the total grey area is equal in either case. To provide numerical information for the scenario in Figures 3 and 4, consider that data in Figure 3 have 60 observations per cluster and an ICC of .03 whereas data in Figure 4 have eight observations per cluster and an ICC of .25. The DEFT of both data sets is the same; the Figure 3 DEFT is $\sqrt{1 + (60 - 1) \times 0.03} = 1.66$ and the Figure 4 DEFT is $\sqrt{1 + (8 - 1) \times 0.25} = 1.66$, meaning that the standard errors are equally affected by independence violations despite divergent ICC values.

Figure 2
Conceptual Diagram of Shared Contextual Sources of Variability That the Intraclass Correlation Is Attempting to Quantify



Note. The grey area represents shared contextual variance, and the white areas represent individual variance.

Substantive Issues With Clustered Data

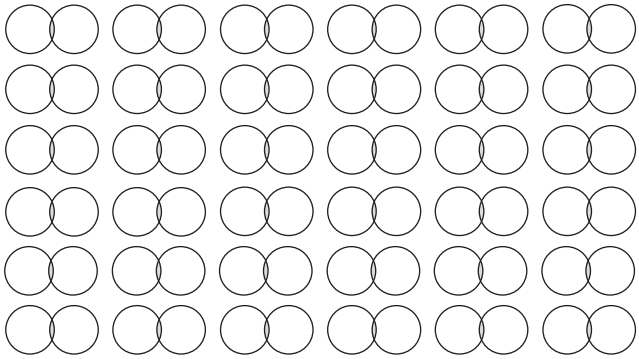
The previous section emphasized statistical issues with standard errors and p-values when modeling clustered data, but there are also substantive issues to consider. In independent data, the association between a predictor and an outcome is a single value such that the effect is the same for all people. In clustered data, there may be *heterogeneity* in the association between predictors and outcomes such that the effect has a different magnitude in different clusters. That is, effects may be better represented by a *distribution* rather than a single value.

For instance, if studying the effect of identifying as queer on self-esteem in high school students, the magnitude of the association may be different in different schools given that the characteristics of the school may differ in ways that affect this association (e.g., whether the school is religious or secular, the political leanings of where the school is located). Some methods for clustered data attempt to quantify heterogeneity in associations between variables because contextual differences may be seen as a moderating effect central to understanding behavioral processes rather than a statistical inconvenience that distorts standard errors and p-values.

The ICC is the best metric for identifying potential heterogeneity because it quantifies the proportion of variance in the outcome attributable to shared context. If the proportion is sufficiently high (.05 is a common threshold; Hox, 1998), explicitly modeling heterogeneity in associations between the outcome and predictors may be worthwhile. If the proportion is small (e.g., below .05), it may not be worth building a more complex model to explain a small source of variability. Nonetheless, insufficient shared contextual variability (measured by the ICC) is distinct from whether standard errors are accurately gauging sampling variability (measured by the DEFT).

Figure 3

Hypothetical Depiction of a Cluster With 60 People Where Only 3% of the Outcome Is Attributable to Shared Contextual Sources



Note. The total amount of grey shading is identical to Figure 4 but spread over more people. This figure is an approximation and shows shared context between pairs of observations. In reality, the context would be shared amongst all members of the cluster, but this was harder to legibly depict visually, so the figure is a simplification.

Additionally, the definition of "one-unit difference" in the interpretation of regression coefficients may become ambiguous in clustered data. Consider again the context of predicting test scores from SES. A "one-unit difference" could mean a student's SES increases, the school's SES increases, or some mix of the student and school SES increasing that sums to one-unit overall. Models can be specified in ways to produce effects that clarify the definition of "one-unit difference." More formal definitions for these types of disaggregated effects are (a) the *within* effect that quantifies the expected change in the outcome when an individual's value of a predictor differs but the context stays the same, (b) the *contextual* or *compositional* effect that quantifies the expected change in the outcome if an individual's value were the same but they were in a different context, or (c) the *between* effect that quantifies the expected change in the outcome if the cluster mean differed.

Overview of Three Methods for Clustered Data

The remainder of this article discusses three approaches—clustered errors, multilevel models, and fixed-effect models—to modeling clustered data and the different ways in which they accommodate independence assumption violations. Specific details are discussed in a dedicated section for each method, which outline the basic idea of the method and a description of research questions for which the method is designed. Subsequent sections include example analyses demonstrating how particular research questions can be addressed with each method and more advanced topics to be considered with each method.

Table 1 provides a broad comparative overview of the three methods. Figure 5 shows the conceptual idea of each method using the Venn diagram idea in Figure 2 (each panel will be explained in the corresponding section). As a short, one-sentence overview of the idea of each method,

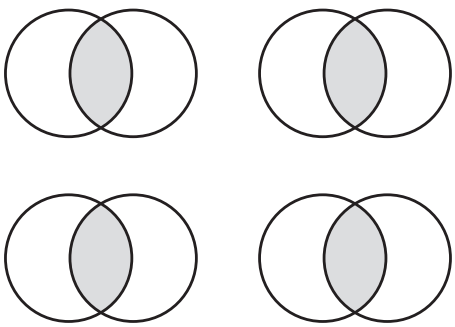
1. Clustered errors: Clustering impacts the statistical properties of inferences, so a correction is applied to produce standard errors (and *p*-values) that are robust to independence violations

2. Multilevel models: Clustering provides richer information that can be used to build a more complex model for the interplay between individuals and their environment in addition to producing standard errors and *p*-values that are robust to independence violations

3. Fixed-effect models: Clustering results in an interplay between individuals and their environment that impacts the ability to meaningfully compare individuals from different clusters, so contextual and environmental influences are statistically removed so they do not contaminate estimates of individual-level effects

Figure 4

Hypothetical Depiction of a Cluster With Eight People Where 25% of the Outcome Is Attributable to Shared Contextual Sources



Note. The total amount of grey shading is identical to Figure 3, but spread over fewer people.

Clustered Errors

How It Works

Clustered errors are not a multilevel model but are a technique to make single-level models robust to independence assumption violations through a statistical correction to the standard errors. Essentially, the width of the theoretical sampling distribution (the dashed line in Figure 1b) is altered to incorporate double counting of contextual information (e.g., the grey area in Figures 2–4) and to better reflect the effective sample size. Clustered errors

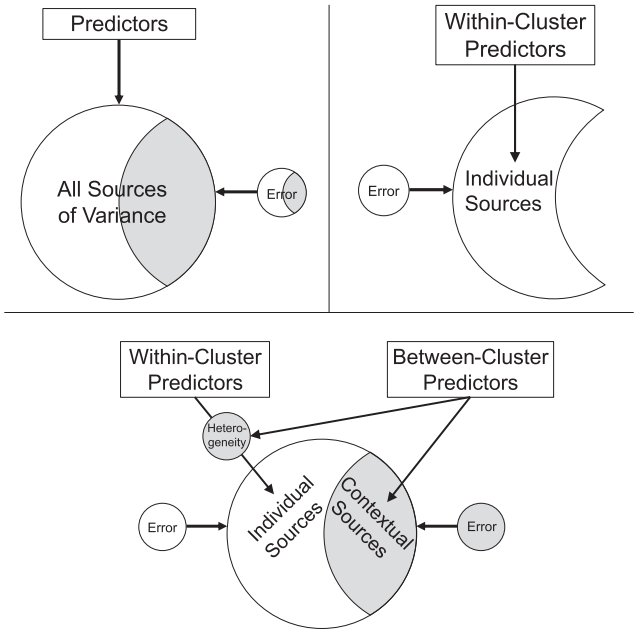
Table 1
Comparative Overview of Mechanism and Goals for Clustered Errors, Multilevel Models, and Fixed-Effect Models

Attribute	Clustered errors	Multilevel models	Fixed-effect models
Accommodating clustering	Corrects standard errors to reflect loss of information when data are clustered rather than independent	Partitions the variance by level and builds a model to explain sources of dependence among observations	Adds cluster affiliation dummies to remove all contextual variance so that errors are conditionally independent
Main interest	Clustering is a nuisance for which to correct so that standard errors, inferences, and *p*-values are accurate	Clustering permits quantifying and explaining heterogeneity in associations between predictors and the outcome	Controlling for shared context so effects of individual-level variables are free of contextual influences
Venn diagram description	One model for combined area that corrects for the presence of grey area	Separates the white area and grey areas and builds a submodel for each	Removes the grey area entirely and builds a model for the white area
Pooling	Complete pooling; effects represent the average across all clusters	Partial pooling; cluster-specific effects are blended with the average across all clusters to improve generalizability to broader population	No pooling; estimates are cluster-specific and not influenced by the average across clusters. Inference not intended beyond clusters in sample
Coefficient heterogeneity	Not directly supported	Random coefficients with an assumed distribution	Interactions with cluster affiliation dummies that are directly estimated
Sample size requirement	40 clusters, without corrections	30 clusters, without corrections	No limit
Distributional assumptions	Only for single error term	Error term at each level and each coefficient that is heterogeneous	Only for single error term

approximate the sampling distribution of the regression coefficients if the study were repeatedly done with differently *cluster* sampled data sets rather than *independently* sampled data sets.

Typically, standard errors become larger to reflect decreased imprecision when each observation does not contain 100% unique information (standard errors may decrease in the rare instance of negative ICCs, though). Test statistics (e.g., t or Z) for each parameter are updated with corrected standard errors, which yield revised p-values for inference. Importantly, clustered errors do not explicitly model or control for any of shared contextual information and the point estimates and interpretation of the regression coefficients are unchanged. Clustered errors address *statistical* issues arising from clustering (e.g., inflated Type-I error rates) but do not necessarily address *substantive* issues related to clustering (e.g., the context affects the relation between a predictor and the outcome).

Figure 5
Conceptual Diagrams of Different Clustering Methods



Note. The left panel shows clustered errors, the right panel shows fixed-effect models, and the bottom panel shows multilevel models. White areas indicate variance attribute to individual characteristics, and grey areas indicate variance attributable to shared contextual characteristics.

The left panel of Figure 5 shows a conceptual diagram of the idea of clustered errors. The outcome (represented by the circle) contains variability attributable to individual sources (in white) and contextual sources (in grey), but predictors do not differentiate between the white and grey areas and a single error term contains unexplained variance from all sources. Standard errors adjust for the presence of the grey area, but there are no steps to explicitly separate the white and grey areas.

Clustered errors operate in a single-level regression framework, so the interpretation is no different than any other single-level linear regression model. The coefficients still correspond to the expected change in the outcome for a one-unit increase in the predictor. Also similar to single-level regression, the coefficients are *completely pooled*, which means that the model does not differentiate within-cluster variance from between-cluster variance and the regression coefficients are estimated using data points from all clusters (e.g., if clusters are different sizes, bigger clusters will have more influence on the coefficients). Straightforward methods to address model fit like R^2 remain calculable and formulas for effect sizes are unchanged. The only aspect of the model that changes is the method by which the standard errors are computed to address independence assumption violations.

More Technical Explanation

Imagine a small data set with six observations and two clusters such that each cluster has three observations. If a regression model

is fit to data from these six people, each person would have their own error term such that the error vector $\mathbf{e}$ would have a length of 6. The covariance matrix of the errors (denoted as $\mathbf{\Omega}$) would be 6×6. If this matrix were unstructured, it would look like

$$\text{Cov}(\mathbf{e}) = \mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & & & & & \\ \sigma_{21} & \sigma_2^2 & & & & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & & & \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & & \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 & \\ \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{bmatrix}, \quad (2)$$

such that each element is potentially unique. If this were feasible, there would be no homoskedasticity assumption because the diagonal elements are unconstrained. The independence assumption would be relaxed also because the off-diagonal terms are all estimated, allowing for dependence among observations. Without these assumptions, the sampling covariance matrix of the regression coefficients (from which standard errors are taken by the square root of the diagonal terms) would be calculated as $\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. The mechanics of this expression are less important than the fact that $\mathbf{\Omega}$ (the error covariance matrix) is in the middle and is central to quantifying sampling variability of the regression coefficients.

Unfortunately, uniquely estimating each element of $\mathbf{\Omega}$ is not possible because there are $(6 \times 7)/2 = 21$ unique elements, which would exhaust the degrees of freedom with six observations (Goldfeld & Quandt, 1965). This holds generally because N parameters can be estimated but the number of nonredundant elements of $\mathbf{\Omega}$ is $[N \times (N+1)]/2$, which is always larger than N.

Assuming independence and homoskedasticity simplifies the error covariance matrix considerably to $\mathbf{\Omega} = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is an identity matrix with 1s on the diagonal and 0s in the off-diagonal. This makes the diagonal elements constant and constrains off-diagonal elements to 0,

$$\mathbf{\Omega} \overset{\text{i.i.d.}}{=} \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & & & & & \\ 0 & \sigma^2 & & & & \\ 0 & 0 & \sigma^2 & & & \\ 0 & 0 & 0 & \sigma^2 & & \\ 0 & 0 & 0 & 0 & \sigma^2 & \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}. \quad (3)$$

When $\mathbf{\Omega} = \sigma^2\mathbf{I}$, most of the terms in the sampling covariance matrix expression above for $\text{Cov}(\hat{\boldsymbol{\beta}})$ drop out because multiplying by an identity matrix is analogous by multiplying scalars by 1. Specifically, under independence and homoskedasticity such that $\mathbf{\Omega} = \sigma^2\mathbf{I}$, the sampling covariance of the regression coefficients reduces to $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which parsimoniously captures the entire error covariance matrix with a single estimated parameter σ^2, provided that assumptions are met. When assumptions are not met, a dilemma emerges—the simplified matrix provides inaccurate estimates of sampling variability, but the unstructured matrix has too many parameters to estimate.

Clustered errors provide a compromise between these two extremes. If the data are clustered such that observations are related within a cluster, but observations are independent between clusters, then $\mathbf{\Omega}$ can be specified to be *block diagonal*. That is, observations in the same cluster have nonzero off-diagonal elements (i.e., observations from the same cluster covary due to their shared context) but observations in different clusters have off-diagonal elements constrained to 0 (i.e., they do not have a shared context and are therefore independent). For the hypothetical example with six people where Observations 1 through 3 are in Cluster 1 and Observations 4 through 6 are in Cluster 2, such a block diagonal error covariance matrix would be

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 & & & & & \\ \sigma_{21} & \sigma_1^2 & & & & \\ \sigma_{31} & \sigma_{32} & \sigma_1^2 & & & \\ 0 & 0 & 0 & \sigma_2^2 & & \\ 0 & 0 & 0 & \sigma_{54} & \sigma_2^2 & \\ 0 & 0 & 0 & \sigma_{64} & \sigma_{65} & \sigma_2^2 \end{bmatrix}. \quad (4)$$

If counting the number of unique elements in this matrix, readers may note that the number of unique terms still exceeds N. To work around this and preserve degrees of freedom, these elements are not parameters that are directly estimated in the model (i.e., they do not consume any degrees of freedom; White, 1980, Lemma 2, p. 820). Instead, the values substituted into this block diagonal matrix are determined using the *descriptive* covariance of the errors from a fitted model assuming independence and homoskedasticity.

That is, if these assumptions are violated, errors for observations in the same cluster will have some association. The logic is similar to using the errors in regression diagnostic tests where the errors are output for further analysis. Rather than estimating the association between errors in the same cluster as part of the model (and consuming degrees of freedom), the association is determined post hoc, preserving degrees of freedom.

This property is why clustered errors are sometimes called "empirical errors" because the standard error correction is based on the *empirical* association between errors within the same cluster. It is also why the regression coefficients do not change with clustered errors. Clustering is not acknowledged during estimation and instead only when quantifying sampling variability based on associations among empirical errors.

Specifically, clustered errors substitute $\mathbf{\Omega} = \sum_{j=1}^{J} \hat{\mathbf{e}}_j\hat{\mathbf{e}}'_j$ into the $\text{Cov}(\hat{\boldsymbol{\beta}})$ expression where j is an index for the cluster and J is the total number of clusters. In this formula, $\hat{\mathbf{e}}_j = \mathbf{y}_j - \hat{\mathbf{y}}_j$ such that the errors for observations in cluster j are equal to the observed values in cluster j minus predicted values in cluster j. $\hat{\mathbf{e}}_j\hat{\mathbf{e}}'_j$ is matrix shorthand for the descriptive covariance matrix of a vector, so $\hat{\mathbf{e}}_j\hat{\mathbf{e}}'_j$ calculates the empirical covariance among the errors in cluster j from a model assuming independence. The summation sign at the beginning dictates that this process is repeated for each cluster, which are appended to create an overall block diagonal structure.

Also note that the empirical error covariance does not necessarily require a constant error variance be (i.e., the diagonal of $\hat{\mathbf{e}}_j\hat{\mathbf{e}}'_j$ is not constrained within or between clusters), so clustered errors provide robustness to homoskedasticity violations in addition to independence violations. In fact, heteroskedasticity-consistent errors are just clustered errors with *all* off-diagonal terms constrained to 0 rather than just off-diagonal terms for observations from different clusters constrained to 0 (Cameron & Miller, 2015, p. 321). Chapter A in the online supplemental materials shows a small example to walk through details of this process.

At this point readers may ask why the block diagonal structure is necessary and why not just apply the empirical approach to the full matrix? The short answer is that the empirical approximation is rough, so applying one empirical approximation for the full matrix may not reasonably approximate $\mathbf{\Omega}$ and may not yield accurate

standard errors. However, applying an empirical approximation individually to each of the J clusters in a block diagonal matrix averages over J rough approximations (Cameron & Miller, 2015, p. 320, p. 324). As $J \to \infty$, the roughness of any single empirical approximation is less relevant through the law of large numbers, so accurate standard errors can be obtained. $J \geq 40$ is typically considered sufficiently large, but small sample corrections have also been devised to accommodate $J < 40$ (Fay & Graubard, 2001; Kauermann & Carroll, 2001; Mancl & DeRouen, 2001).

After clustering the errors, some ambiguity about the degrees of freedom emerges. Regression models base degrees of freedom on N, but that is not entirely appropriate given that the effective sample size in clustered data is not equal to N. One conservative (and infrequent) approach suggests that degrees of freedom are irrelevant because clustered errors are only accurate as $J \to \infty$, at which point degrees of freedom have no impact. A more common method is to base the degrees of freedom based on $J - 1$ or $J - 2$ (Cameron & Miller, 2015; Shah et al., 1977). A second method by Donald and Lang (2007) suggests using J minus the number of cluster-level predictors (i.e., variables that are constant for all members of the same cluster but that may differ for different clusters). More complex degrees of freedom based on the characteristics of the data have also been proposed (R. M. Bell & McCaffrey, 2002) and have been reported to perform better than heuristic approaches like Donald-Lang degrees of freedom with smaller sample sizes (Huang & Li, 2022). In general, degrees of freedom with clustered errors are an approximation and unlikely to be exact as with independent data.

What Clustered Errors Are Designed for

Clustered errors are ideal when clustering is a nuisance and not directly related to the research questions, such as when clustering was an incidental part of the data collection. As one example, consider a researcher who is interested in the effectiveness of a reading intervention designed to help students improve vocabulary with strategies for situations when they encounter a word whose meaning they do not know. For this type of study, it may not be possible to collect a large enough sample of students without recruiting from multiple classrooms.

Consequently, students may incidentally be clustered within classrooms in order to obtain a sufficient sample size. However, if there were some hypothetical, infinitely large classroom from which the entire sample could have been collected, the ability to assess the effectiveness of the intervention would not have been affected. It may still be necessary (depending on the DEFT) to apply a statistical correction for the lack of independence between observations within the same classroom; however, a single-level model with corrections for independence assumption violations may be sufficient and a more advanced model whose goal is to parse out individual and classroom influences may not be necessary to address the research question.

As another example, cluster sampling is sometimes used to reduce costs or increase sampling efficiency. Cluster sampling selects clusters to participate in a study and then includes all members of the selected cluster. For instance, a behavioral intervention to help children exercise more frequently might recruit all siblings within a family to participate in a study (where children are clustered in families). Or schools may be recruited to participate in an antibullying intervention and all consenting students may be included as participants (where students are clustered within schools).

Here, clustering is a feature of the data collection strategy, but clustering may not be central to the research question. For instance, including all siblings into a behavioral intervention promoting physical activity may not be done because there is an inherent interest in parental or home environment characteristics affecting exercise frequency or duration. This sampling approach may be enacted because it is just difficult to find participants and it is more efficient to include all children once a family consents (i.e., the burden to add a sibling from an already participating family is lower than finding a new family to consent). It is important to account for the clustering, but if the clustering is incidental and not directly tied to the research questions, a statistical correction like clustered errors may suffice to allow researchers to address their research questions using less complex methods.

Multilevel Models

How It Works

Multilevel models treat clustering as an opportunity to better understand and delineate the distinct contributions of the individual and the context. Multilevel models partition the variance in the outcome into individual sources (a.k.a. within-cluster variance) and contextual sources (a.k.a. between-cluster variance). Within-cluster and between-cluster submodels are then built to explain these different sources of variance, allowing a multilevel model to articulate which source of variance a predictor explains. Multilevel models produce correct standard errors as a byproduct, but correct standard errors themselves are not sole focus as with clustered errors.

The bottom panel of Figure 5 expresses the idea of multilevel models as a Venn diagram. The multilevel model separates the white area from the grey area and then builds *submodels* for each source of variance. The submodel for the white area features individual characteristics (variables that are potentially different for each person) to explain within-cluster variance and the submodel for the grey area features contextual characteristics (variables that are constant for people in the same cluster but that can differ across people in different clusters) to explain between-cluster variance. Each submodel has its own error term to capture unexplained variance from each respective source, meaning that the overall model has multiple error terms to quantify unexplained variance at different levels. Whereas clustered errors sought to correct for the presence of the grey area, a multilevel model uses the grey area to investigate the interplay of individuals and their environment to uncover mechanisms for how contexts moderate associations.

Importantly, multilevel models do not "control for clustering" in the sense that all contextual influences are accounted for.[1] Partitioning the variance quantifies the proportion of variance at each level with multiple error terms representing *unexplained* variance. To control for contextual characteristics, multilevel models must explicitly include predictors in the between-cluster submodel.

[1] Partitioning the variance may "explain clustering" because it clarifies how variance is allocated across levels (e.g., Rights & Sterba, 2020, p. 588) and "control for clustering" may accurately describe that multilevel model standard errors account for clustering. However, multilevel models do not partial out contextual characteristics when partitioning the variance, so a multilevel model does not "control for clustering" in the regression sense of "control" unless relevant contextual characteristics are explicitly included in the between-cluster submodel.

Random Coefficients

The within-cluster submodel may feature *random regression coefficients*. "Random" in this context does not mean "haphazard" as in everyday language. Instead, "random" comes from its use in probability to mean "follows a distribution." So "random regression coefficients" means that the regression coefficients do not have a single value but instead have a distribution of values. Conceptually, the association between the predictor and the outcome is different in each cluster, allowing heterogeneity in the association between the predictor and the outcome. In other words, there is not a single regression line but instead a unique regression line for each cluster. Researchers can acknowledge that associations may be moderated by the context surrounding each cluster and that the magnitude of the association between the outcome and a predictor may be contextually dependent.

As an example, students may be clustered in schools and the outcome of interest is self-esteem. A focal predictor is whether a student identifies as queer, which is a student characteristic that varies from person to person. The association between identifying as queer and self-esteem may be contextually dependent such that the magnitude (or even the direction) of the association varies across schools (i.e., there is heterogeneity in the association between identifying as queer and self-esteem). A multilevel model first quantifies this heterogeneity to inform a researcher how much the association varies across different schools. Subsequently, the heterogeneity can be predicted by school-level characteristics (e.g., the proportion of students identifying as queer, whether the school is religious) to better understand how the school context impacts or moderates the association between identifying as queer and self-esteem (e.g., the association between identifying as queer and self-esteem may be more negative in private schools than in public schools). This is referred to as a *cross-level interaction* because school characteristics predict or moderate the association among individual characteristics.

Random regression coefficients require that a particular distribution be assumed. Normality is a common choice and used by default in most statistical software (Littell et al., 2006), although other options are possible (Zhang & Davidian, 2001). When assuming a distribution for a coefficient across clusters, the model does not directly estimate the coefficient in each specific cluster. Instead, the model estimates properties of the assumed distribution. With a normal distribution, this corresponds to the mean and variance. The mean of this normal distribution is called the *fixed effect*, which represents the average association between the predictor and the outcome across all clusters. The *variance component* is the variance of the distribution, which represents the heterogeneity in the coefficient across clusters. The coefficient in a specific cluster can be recovered with empirical Bayes predictions, but cluster-specific coefficients are not directly featured in the model.

Assuming a distribution of coefficients across all clusters rather than directly estimating the specific effects in each cluster has four notable benefits.

1. The target of inference is the entire population from which clusters were sampled rather than just the clusters that were included in the analysis.

2. The model directly quantifies the heterogeneity in the regression coefficients.

3. The model scales easily because estimating the mean and variance of a distribution costs few degrees of freedom regardless of the number of clusters in the data.

4. The cluster-specific coefficients are *partially pooled*, meaning that they are a weighted average of the fixed effect and the data specifically from the particular cluster (Gelman, 2006). This has advantages when some clusters are small and generally leads to better predictions

What Multilevel Models Are Designed for

Multilevel models are ideal when clustering is purposeful and fundamental to the research question itself such that the research question could not be answered if the data were not clustered. This includes questions about how context may affect individuals or when a primary objective is to explain heterogeneous associations between predictors and outcomes. To multilevel models, clustering is an asset that expands the research questions that can be explored and provides opportunities to clarify among competing mechanisms and investigate questions that cannot be answered from single-level data. There is no inherent problem with using multilevel models to produce standard errors and *p*-values that reflect clustering; however, doing so fails to realize the full potential of multilevel models and can overcomplicate analyses that could be handled more simply with other methods.

In school research, for example, there may be questions about how different types of teaching styles affect learning or about how different classroom environments affect students' motivation to learn. The classroom context in which the student finds themselves is central to these questions and students being clustered in different classrooms permits a multilevel model to examine (a) how much variability in the outcome is attributable to student characteristics and teacher characteristics, (b) if student-level characteristics have different associations with the outcome in different classrooms or pedagogical environments, and (c) which teacher-level characteristics might explain heterogeneity. If data were only collected from a single teacher, then all students would share the same context and questions about the differential contribution of students and teachers could not be addressed.

Disaggregated Effects

Multilevel models are well-suited to disaggregate a single predictor to differentiate how the outcome is anticipated to change when an *individual's* value of the variable increases by one unit and when a *cluster's* value of the variable increases by one unit. Using students clustered in classrooms as an example, it could be important to differentiate among the effect of a student's motivation increasing by one unit (a *within* effect), the classroom's average motivation increasing by one unit (a *between* effect), and a student's motivation staying the same but the classroom's average motivation increasing by one unit (a *contextual* effect). A multilevel model can differentiate among these effects based on how the predictors in the within-cluster submodel are centered and whether the cluster means of within-cluster predictors are included as between-cluster predictors.

Uncentered predictors in the within-cluster submodel and excluding cluster means as predictors in the between-cluster submodel misses opportunities to differentiate between changes in

individual characteristics and changes in contextual characteristics. That is, raw individual characteristics are partially informed by the individual's context, so uncentered individual-level variables contain a mix of individual and contextual information. Centering isolates pure individual information by subtracting contextual information contained in cluster mean. Otherwise, coefficients may be conflated (a.k.a. blended, composite, total, or smushed; Burstein, 1980; Hoffman, 2015; Preacher et al., 2010; Wang & Maxwell, 2015) and yield an uninterpretable blend of individual and contextual information (Curran & Bauer, 2011; Enders & Tofighi, 2007; Hamaker & Grasman, 2015).

Fixed-Effect models

How It Works

Fixed-effect models partition the variance in the outcome into different sources, similar to multilevel models. However, fixed-effect models do not build a between-cluster submodel for contextual sources of variance. Instead, they completely remove influences of all possible contextual variables, regardless of whether they were collected or appear in the data. The goal is to create a within-cluster model that cannot be affected by the between-cluster information. Whereas multilevel models partition the variance to eventually build a submodel to explain different sources variance, fixed-effect models partition the variance in order to completely factor out all contextual sources. Multilevel models treat contextual variance as an equally important source deserving its own submodel, but fixed-effect models treat contextual variance as a potential confound for which to control. The within-cluster model has clear primacy in a fixed-effect model whereas the within-cluster and between-cluster submodels have equal status in a multilevel model.

Like clustered errors, fixed-effect models view the clustering as unrelated to the research questions and as an aspect to merely be accommodated. However, fixed-effect models are motivated by substantive rather than statistical issues. In clustered errors, the concern is that the clustering will adversely affect inferences, so the emphasis is correcting standard errors and *p*-values. The concern in fixed-effect models is that individuals from different clusters will be incomparable due to different contexts and that clustering will distort interpretation of individual characteristics on the outcome. Fixed-effect models therefore aim to control for all contextual characteristics so that associations in a within-cluster model can be estimated as if the data were independent. This is distinct from clustered errors, which only correct the standard errors but do not alter the regression coefficients to reflect possible differences in contextual characteristics.

The right panel of Figure 5 shows the conceptual idea of fixed-effect models . The model partitions the variance into individual and contextual sources. However, fixed-effect models discard the grey area from the analysis entirely after partitioning such that the model proceeds only with within-cluster variance associated with the white area. This is opposed to a multilevel model where the grey area is substantively interesting and separately modeled. Fixed-effect models also feature a single error term that contains unexplained variance solely from individual sources. This differs from clustered errors where the error term is composed of unexplained variance from all sources.

To accomplish these goals, the most common approach is to create cluster affiliation dummy variables for each cluster. For instance, if there are 50 clusters, 50 binary variables are created whose value equals 0 if the observation is not a member of the cluster or equals 1 if the observation is a member of the cluster. The set of all cluster affiliation dummies are then directly included in a single-level regression model as predictors (if all are included, the intercept must be suppressed; if the intercept is included, then one cluster affiliation dummy is omitted as a reference cluster). The set of cluster affiliation dummies *saturates* the between-cluster submodel, essentially acting as a black box that nondescriptly absorbs all possible contextual sources of variance. The tradeoff is that all between-cluster variance will be explained and factored out, but effects for specific contextual characteristic cannot be estimated because it will necessarily be perfectly collinear with the cluster affiliation dummies. That is, because the cluster affiliation dummies explain all sources of contextual variance, any variance explained by a specific between-cluster predictor will be completely redundant with variance explained by the cluster affiliation dummies.

A fixed-effect model can be considered an analysis of covariance (ANCOVA) where the cluster affiliation is a fixed categorical factor and substantive within-cluster predictors are covariates. In traditional ANCOVA, the categorical factor is the focus and covariates are controls. In a fixed-effect model, the roles are reversed and the categorical factor controls for clustering and the covariates are the focus. Because fixed-effect models statistically remove variance attributable to contextual sources, they truly "control" for clustering in a regression sense because all contextual sources of variance are partialed out.

The coefficients associated with each cluster affiliation variable represent cluster-specific intercepts. This is related to—but distinct from—the approach used in a random intercepts multilevel model. The difference is that a multilevel model assumes a distribution for all the intercepts whereas the fixed-effect model directly estimates the intercept for each cluster, which is the origin of "fixed-effect model" because cluster-specific intercepts are directly estimated. This means that

1. Results from a fixed-effect model only generalize to the clusters that were included in the data.

2. The clusters are not assumed to be randomly sampled from the broader population of clusters.

3. There is no pooling of regression coefficients (i.e., only the observations within each cluster contribute to the cluster-specific intercept; the estimates are not blended with the overall mean).

4. All sources of contextual variances for the intercept are accounted for, regardless of whether the relevant contextual characteristics were collected and included in the data.

Multilevel and fixed-effect model models are sometimes written equivalently with the distinction only hinging on assumptions about the cluster-specific intercepts.

Standard errors will accurately reflect clustering when cluster affiliation dummies are included as predictors. To understand why,

recall that (a) the independence assumption is about errors and (b) errors are conditional on predictors in the model. All contextual variance is explained by the cluster affiliation dummies, so the errors will be free of contextual influences. Once the contextual variance is factored out, the (conditional) independence assumption is upheld because any source of covariance between errors from observations in the same cluster has been explained by the cluster affiliation dummies (assuming only one source of clustering, violations of which are discussed later).

What Fixed-Effect Models Are Designed for

As instances where fixed-effect models may be helpful, perhaps relevant contextual characteristics were not collected such that it is not possible to build an appropriate between-cluster submodel. As a hypothetical example, consider a secondary data analysis using a large public data set to investigate cannabis use where people are clustered within U.S. states. The data may not include the legal status of cannabis in each state at the time of data collection (cannabis laws differ from state to state and many have changed recently). Cannabis legality likely affects aspects related to cannabis use (e.g., how easy it is to acquire, attitudes towards use), but this variable cannot be incorporated in the model if it were never collected. A fixed-effect model could control for cannabis legality (and all other state-level variables) despite the variable not being collected. The effect of cannabis legality could not be directly estimated, however.

This type of situation may also occur when clustering is incidental. For instance, a study on anxiety may collect data from multiple treatment centers. Perhaps these treatment centers differ in relevant ways (e.g., different approaches to treatment, geographical differences, types of therapists) but differences are not related to the research questions, so variables about the treatment centers are either not collected (e.g., limited researcher time, high expense of data collection) or the researchers do not have a theory about how center-level variables affect anxiety. In this case, fixed-effect models could factor out all center characteristics and allow the researchers to fit models that focus on patient characteristics, controlling for any center-level differences. Fixed-effects models also do not require randomly sampled clusters, which can be useful if centers are incidental such that researchers did not expend effort to randomly sample centers when centers are not be related to the research questions.

Data with few clusters can also present similar issues such that there may not be sufficient information or variability to confidently build a between-cluster submodel. Using the same treatment center example, there may have only been seven centers in the data because centers were not a research focus. Due to distributional assumptions, multilevel models are susceptible to estimation issues with few clusters, usually defined as 30 or fewer (Hox & McNeish, 2020; Maas & Hox, 2005).[2] There also may not be sufficient variability in between-cluster variables. For instance, if six centers are affiliated with a hospital and only one center is not, it would be difficult to reliably estimate the effect of hospital affiliation with so little variability in this predictor.

In such cases, a fixed-effect model may be the preferred choice even if researchers may have a reasonable between-cluster submodel in mind. When the data may not be sufficiently rich to support a between-cluster submodel, potential insights about specific between-cluster predictors may be sacrificed to fortify the integrity of within-cluster estimates where the sample size is larger and modeling can be conducted more assuredly.

Lastly, fixed-effect models are well-suited for instances where the primary motivation is inference to a specific set of clusters rather than a broader population. Because fixed-effects models do not make distributional assumptions, their inferences generalize only to the clusters included in the data (A. Bell et al., 2019). This lack of generalizability is sometimes cited as a weakness, but it can be a strength in situations where the target of inference is restricted to clusters specifically in the data. If clusters represent all U.S. states, all countries in the European Union, schools in a particular district, or all hospitals in a county; generalizing beyond the sample may not be relevant and distributional assumptions may not be helpful or necessary (Clark & Linzer, 2015). This maps onto the distinction of fixed versus random ANOVA where fixed ANOVA generalizes only to the conditions included in the study whereas random ANOVA generalizes to the population of possible conditions to which participants could have been exposed (Hedges & Vevea, 1998).

Hypothetical Research Question and Example Analysis

With the general idea of each method covered, this section provides example analyses for each method to answer different hypothetical research questions from a single data set. Data and code for fitting all example models in SAS and R is provided on the Open Science Framework page associated with this article, https://osf.io/w4x9n/.

Example Data

This section uses the 1982 High School Beyond data set that appears in the Raudenbush and Bryk (2002) multilevel modeling textbook. The data contain 7,185 U.S. high school students clustered within 160 schools and each school has a different number of students (range = 14–67). The main outcome is math achievement scores ($M = 12.75$, $SD = 6.88$). There are two student-level predictors in the data: student SES ($M = 0$, $SD = 0.78$) and an indicator for whether the student's racial identity is non-White (yes = 28%). There are also two school-level predictors: whether the school is public or private (44% private) and the number of students in the school ($M = 1{,}097$, $SD = 629$).

The original focus of predicting achievement scores based on racial and socioeconomic differences may not feel maximally inclusive to some readers 40 years later. Nonetheless, there is an expanding literature in educational and developmental psychology on achievement gaps based on student racial identification and socioeconomic background (e.g., Assari et al., 2021; Howard, 2019; Merolla & Jackson, 2019) and how differences persist throughout the lifespan (Henry et al., 2020). This work became especially prevalent after Covid-19 as gaps are forecasted to widen (Bailey et al., 2021) despite previously shrinking (albeit slowly) for decades (Hanushek et al., 2022; Hashim et al., 2020). This research emphasizes contextual characteristics in relation to the formation and reduction of gaps such as instructor mindset

[2] There are small sample corrections (Kenward & Roger, 1997, 2009) and Bayesian methods to address small sample issues in multilevel models (e.g., Baldwin & Fellingham, 2013; Stegmueller, 2013; van de Schoot et al., 2015).

(Canning et al., 2019), teaching style (Theobald et al., 2020), and school climate (Berkowitz, 2021), which maps well onto the different approaches for clustered data. In sum, these data and the hypothetical research questions in subsequent example analyses are consistent with ongoing research to quantify and reduce achievement gaps and promote equitable outcomes rather than being an antiquated example that uses outdated terminology and ideas.

Clustered Errors

The hypothetical research question is whether reporting a non-White racial identity moderates SES achievement gaps in math scores (i.e., whether the SES achievement gap is the same size for students identifying as White or non-White). This research question does not involve exploring school influences on this association and there is no explicit interest in quantifying whether achievement gaps are heterogeneous across different schools. The clustered nature of the data is not necessary to answer this question and the ability to address this question would be unaffected if it were feasible to collect data from a single school, so a statistical correction for an independence violation may be sufficient.

The regression model to address this question could be written as

$$\begin{aligned}\text{Math}_i = \beta_0 + \beta_1 \text{SES}_i + \beta_2 \text{Non-White}_i \\ + \beta_3(\text{SES}_i \times \text{Non-White}_i) + e_i \end{aligned} \quad (5)$$

and contains an intercept (β_0), main effects for SES (β_1) and non-White identity (β_2), and one interaction to test moderation (β_3). Estimates assuming independence are shown in the top of Table 2; both main effects and the interaction are significant at the .01 level. However, students are clustered within schools and the standard errors may be incorrect because the independence assumption upon which they rely is likely violated given the data structure. Calculating the DEFT can approximate the magnitude of this independence violation. The ICC is .18 and the average cluster contains $7{,}185/160 = 44.90$ students, so $\text{DEFT} = \sqrt{1 + (44.9 - 1) \times 0.18} = \sqrt{8.90} = 2.98$. This DEFT value is too large to reasonably ignore because the standard errors are estimated to be about three times larger if clustering were accounted than if independence were assumed. It therefore seems prudent to correct the standard errors to reflect the clustering of observations so that the inferences are more accurate.

Table 2

Comparison of Estimates for a Linear Regression Model Assuming Independence (Top) and a Linear Regression Model With Clustered Errors (Bottom)

Effect	Estimate	*SE*	*df*	*t*	*p*
Assuming independence					
Intercept	13.50	0.09	7,181	152.28	<.001
SES	2.94	0.12	7,181	24.28	<.001
Non-White identity	−2.94	0.18	7,181	−16.59	<.001
SES × Non-White Identity	−0.60	0.21	7,181	−2.84	.005
Clustered errors					
Intercept	13.50	0.17	159	80.56	<.001
SES	2.94	0.15	159	19.49	<.001
Non-White identity	−2.94	0.33	159	−8.91	<.001
SES × Non-White Identity	−0.60	0.32	159	−1.87	.063

Note. Degrees of freedom with clustered errors are calculated by $J - 1$. SES = socioeconomic status.

The results from the same model with clustered errors are shown at the bottom of Table 2. Clustered errors do not impact the regression coefficients given that they do not rely on the independence assumption. Standard errors, *t*-statistics, degrees of freedom, and *p*-values are all adjusted because these values are used for inference and are affected when independence is violated.

Clustered standard errors are about 1.5 to two times larger than the standard errors from the model assuming independence rather than the 2.98 times larger suggested by the DEFT. The DEFT is an estimated value and tends to be conservative. Clustered errors correct the standard errors of each coefficient individually whereas the DEFT is an estimate for the entire design. In this way, DEFT is somewhat analogous to a Dunn–Bonferroni correction—it is easy to calculate and provides a conservative approximation, but more sophisticated approaches provide refined adjustments (i.e., clustered errors are more efficient than multipliying standard errors assuming independence by the DEFT).

The conclusion for the interaction changes after clustering the errors. When assuming independence, there was evidence for moderation ($t[7, 181] = -2.81$, $p < .01$) such that the SES achievement gap in math scores for students identifying as non-White is smaller. Conversely, after clustering the errors, there was insufficient evidence for moderation ($t[159] = -1.87$, $p = .06$) such that the SES achievement gap in math scores is indistinguishable for different racial identifications. Of course, this elicits sentiments that "God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277). Nonetheless, the point is that ignoring clustering can produce incorrect standard errors and change inferences, even in cases where the ICC and DEFT are modest and the *p*-value in the model assuming independence is nowhere near the .05 threshold.

Note that estimates in Table 2 are not disaggregated into within-cluster and between-cluster effects, so they may be considered conflated. There are differing viewpoints about disaggregating effects with clustered errors, which is discussed in detail in the Advanced Considerations section.

Multilevel Models

Using the same data, imagine that the clustering is seen as providing an opportunity to explore possible contributions of school characteristics to achievement gaps. This could include whether there is heterogeneity in the SES and non-White achievement gaps across schools and, if so, whether a school being private explains some or all of the heterogeneity. This research question is not merely about correcting for the presence of clustering, so clustering the errors would be less useful. Instead, the clustered nature of the data permits exploration of heterogeneity and assessment of why achievement gaps may be stronger or weaker in certain kinds of schools. Therefore, building a multilevel model to quantify the heterogeneity and explain heterogeneity is a more suitable option. The main text focuses on the final model, but Chapter B in the online supplemental materials provides a more detailed account of the modeling building process.

Full Model

The full model to explore whether private school status (0 = *public*, 1 = *private*) explains slope heterogeneity in SES and non-White

achievement gaps for math scores can be written as

$$\begin{cases} \text{Within-School Submodel} \begin{cases} \text{Math}_{ij} = \beta_{0j} + \beta_{1j}\text{SES}_{ij}^{(\text{CMC})} \\ \qquad + \beta_{2j}\text{Non-White}_{ij}^{(\text{CMC})} \\ \qquad + \beta_{3j}(\text{SES}_{ij}^{(\text{CMC})} \\ \qquad\quad \times \text{Non-White}_{ij}^{(\text{CMC})}) + r_{ij} \\ r_{ij} \sim N(0, \sigma^2) \end{cases} \\ \text{Between-School Submodel} \begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}\overline{\text{SES}}_j + \gamma_{02}\overline{\text{Non-White}}_j^{(\text{GMC})} \\ \qquad + \gamma_{03}(\overline{\text{SES}}_j \times \overline{\text{Non-White}}_j^{(\text{GMC})}) \\ \qquad + \gamma_{04}\text{Private}_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}\text{Private}_j + u_{1j} \\ \beta_{2j} = \gamma_{20} + \gamma_{21}\text{Private}_j + u_{2j} \\ \beta_{3j} = \gamma_{30} \\ \mathbf{u}_j \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix}\right). \end{cases} \end{cases} \tag{6}$$

There are two different subscripts in Equation 6, *i* indexes students and *j* indexes schools. Any term with *i* and *j* subscripts differs for each student in each school whereas terms with only *j* subscripts are the same for all students within a school but may differ across schools. The within-school submodel is simply a linear regression similar to Equation 5 and includes student characteristics (i.e., variables have *i* and *j* subscripts). The difference in Equation 6 is that the within-school coefficients are heterogeneous and potentially differ in each school (i.e., β coefficients have *j* subscripts). These coefficients therefore become outcomes in the between-school submodel such that each β has a model to explain why its values vary in different schools. This is what makes the model "multilevel" or "hierarchical," coefficients in one equation simultaneously serve out outcomes in equations at the next level (i.e., within-school coefficients are between-school outcomes).

Between-school submodel equations include school characteristics (i.e., variables only have *j* subscripts) to model contextual circumstances responsible heterogeneity in β. The main school characteristic in Equation 6 is private school status, which appears in the β_{0j}, β_{1j}, and β_{2j} equations to assess whether the intercept, SES achievement gap, or non-White achievement gap systematically differ for public or private schools. Each β equation is not required to have the same predictors and different variables can be used in different equations. Between-school coefficients are interpreted similarly to a typical regression. The γ terms whose second subscript is 0 represent the expected value of β when all predictors in the equation equal 0. The γ terms next to variables indicate the predicted change in β when the predictor increases by one unit.

Equation 6 disaggregates all student-level predictors with cluster mean centering such that $\text{SES}_{ij}^{(\text{CMC})} = (\text{SES}_{ij} - \overline{\text{SES}_j})$ and $\text{Non-White}_{ij}^{(\text{CMC})} = (\text{Non-White}_{ij} - \overline{\text{Non-White}_j})$, meaning a value of "0" indicates that the student is at their school's mean of the predictor. The school means of non-White identity (i.e., the proportion of students identifying as non-White in the school) and SES predict the intercept β_{0j} to fully disaggregate effects of these predictors. The school mean of non-White identity is further grand-mean centered such that $\overline{\text{Non-White}}_j^{(\text{GMC})} = \overline{\text{Non-White}}_j - \overline{\text{Non-White}}$ to improve interpretation such that 0 refers to a school with the sample *average* proportion of students identifying as non-White (27.5% in this data) rather than *zero* students identifying as non-White.

Including school means as a school-level predictor allows effects for a one-unit increase at different levels to be separately estimated because there are two coefficients associated with each student-level predictor. For instance, γ_{10} and γ_{01} in Equation 6 both capture effects of SES, the former is the effect of a one-unit SES increase in the *student* and the latter is the effect for a one-unit SES increase in the *school*. The two cluster means also interact in Equation 6 to differentiate within-school (γ_{30}) and between-school (γ_{03}) moderation to assess whether moderation is driven by the SES and racial identity of the student or the SES and racial composition of the school.

The *u* terms are random effects capturing the sources of variance in β not explained by the predictors. In other words, the *u* terms are between-school error terms. For instance, u_1 represents reasons why the SES achievement gap varies across schools beyond private school status. Which coefficients randomly vary across clusters is up to the researcher's discretion (and is sometimes limited by the complexity that the data can accommodate). In Equation 6, within-school moderation is modeled as a constant across schools with no random effect, so the β_{3j} equation has no *u* term.

Because there are multiple random effects, the distributional assumption for the *u* terms is a *multivariate* normal distribution defined by a mean vector and a covariance matrix. The mean vector is comprised of three zeroes, one for each random effect. The τ terms on the diagonal of the covariance matrix are the random effect variances capturing heterogeneity in each β not explained by the predictors in each respective β equation. τ_{00} corresponds to unexplained heterogeneity in the intercept, τ_{11} captures unexplained heterogeneity in the SES achievement gap, and τ_{22} captures the unexplained heterogeneity in the non-White achievement gap. Off-diagonal terms of the covariance matrix capture systematic relations between the coefficients (i.e., the random effect covariances). For instance, if τ_{21} were positive, it would mean that schools with higher values of β_{1j} tend to have higher values of β_{2j} (i.e., different achievement gaps are systematically related within the same school).

Results and Interpretation

Table 3 contains estimates from SAS `PROC MIXED` with restricted maximum likelihood and Satterthwaite degrees of freedom. As an example of how to interpret a disaggregated effect, a one-unit difference in SES for students attending the same public school with an average proportion of students identifying as non-White (i.e., the within effect; changing the student characteristic while keeping the surrounding context constant) is estimated to result in a 2.39 point gap ($t[152] = 15.28$, $p < .01$) whereas a one-unit difference in the school mean of SES at a public school with an average proportion of students identifying as non-White (the between effect) is estimated to lead to a 4.41 point gap ($t[147] = 41.47$, $p < .01$). Effects refer to public schools with an average proportion of students identifying as non-White because effects are conditional on *Private* = 0 and $\overline{\text{Non-White}}_j^{\text{GMC}} = 0$.

Table 3
Estimates From Multilevel Model Examining Whether School Characteristics Explain the Slope Heterogeneity in Effects Across Schools

Effect	Notation	Estimate	*SE*
Student characteristics			
Intercept	γ_{00}	11.85	0.20
SES	γ_{10}	2.39	0.16
Non-White identity	γ_{20}	−3.98	0.36
SES × Non-White Identity	γ_{30}	−0.67	0.31
School characteristics			
SES school mean	γ_{01}	4.41	0.42
Non-White school mean	γ_{02}	−2.61	0.57
SES School Mean × Non-White School Mean	γ_{03}	(ns) −1.02	1.03
Private	γ_{04}	1.64	0.30
Private × SES	γ_{11}	−1.03	0.23
Private × Non-White	γ_{21}	2.12	0.50
Variances			
Intercept	τ_{00}	2.05	—
SES slope	τ_{11}	0.23	—
Non-White identity slope	τ_{22}	1.23	—
Residual	σ^2	35.63	—
Correlations			
Intercept, SES slope		(ns) 0.02	0.32
Intercept, non-White slope		(ns) 0.02	0.26
SES slope, non-White slope		(ns) −0.65	0.79

Note. SES = socioeconomic status; *ns* = not significant as the .05 level.

Holding a public-school student's SES constant, the effect of attending a public school with an average proportion of students identifying as non-White whose mean school mean of SES is one point higher (i.e., the contextual effect; holding the student characteristic constant while changing the context) would be calculated by the between effect minus the within effect: $\gamma_{01} - \gamma_{10} = 4.41 - 2.39 = 2.02$ points, $t(187) = 4.50$, $p < .01$.

The gap resulting from student-level differences in SES within the same public school is about equal to the gap resulting from holding a public-school student's SES constant but moving them to a higher SES public school, so the SES achievement gap in Math Scores appears to have both individual and contextual sources of influence.

Regarding school characteristics, for average SES schools with an average proportion of students identifying as non-White, the non-White achievement gap in private schools is 2.12 points less negative ($-3.98 + 2.12 = -1.86$; which still indicates a nonnull gap $t(77.6) = -5.37$, $p < .01$) than the non-White achievement gap in public schools (−3.89 points).

The public–private difference in the non-White achievement gap is significant, $t(101) = 4.28$, $p < .01$. The SES achievement gap is also reduced and significant in private schools—2.39 for public vs. $2.39 - 1.03 = 1.36$ for private; $t(153) = 7.86$, $p < .01$. Non-White and SES achievement gaps are both roughly halved in private schools compared to public schools.

The random effect variances of non-White identity and SES are $\tau_{11} = 1.23$ and $\tau_{22} = 0.23$, respectively, indicating that there is achievement gap heterogeneity not explained by private school status. However, the random effect variances are reduced from a model that does not include private as a predictor where they were 2.14 and 0.45, respectively (see Chapter B in the online supplemental materials), so private school status appears to at least partially explain differences in achievement gaps across schools.

Regarding moderation, the model suggests marginal evidence for student-level moderation, $\gamma_{30} = -0.67$, $t(5187) = -2.17$, $p = .03$, such that the SES achievement gap is reduced for students who identify as non-White. However, there was no evidence to support between-school moderation, $\gamma_{03} = -1.02$, $t(145) = -0.98$, $p = .33$. These results are not directly comparable to moderation in the model with clustered errors because the multilevel model disaggregates within-school and between-school moderation whereas the model with clustered errors estimated a single aggregated moderation effect.

This example highlights the unique capabilities of multilevel models. They can quantify heterogeneity in effects across clusters and then build a model to identify potential reasons why heterogeneity exists by featuring coefficients as between-cluster outcomes. Multilevel models can be powerful when their full potential is realized and extend far beyond simply correcting standard errors.

Fixed-Effect Models

Similar to previous subsections, the goal of this subsection is to test whether non-White identity moderates the SES achievement gap in Math Scores. The difference in this subsection is that there is concern that students from different schools may not be comparable and that there are insufficient school characteristic variables in the data to build a proper between-school submodel. For instance, the only school characteristics in these data are whether the school is private and how many students attend each school, but perhaps there are other relevant school-level characteristics that affect Math Scores (e.g., what proportion of teachers have advanced degrees, whether calculus is offered) that are not accessible in the data but for which the model should ideally account.

Whereas clustered errors address the statistical issue presented by clustering, a fixed-effect model addresses the substantive issues presented by clustering whereby different contextual circumstances may affect the associations between predictors and Math Scores. That is, the interest is estimating the moderation effect while controlling for all school differences. This is opposed to clustered errors, which correct standard errors but do not control for school differences unless predictors responsible for school differences are directly included in the model.

The traditional fixed-effect model for this situation would be

$$\begin{aligned} \text{Math}_i &= \sum_{j=1}^{J} C_j\alpha_j + \beta_1 \text{SES}_i + \beta_2 \text{Non - White}_i \\ &\quad + \beta_3(\text{SES}_i \times \text{Non - White}_i) + e_i \\ e_i &\sim N(0, \sigma^2) \end{aligned} \quad . \qquad (7a)$$

C_j represents the school affiliation dummy for school j and α_j is the directly estimated school-specific intercept in school j. The summation sign ranges from 1 to J to indicate that there is a school affiliation dummy and school-specific intercept for each of the J schools. There is no β_0 intercept term because there are j unique intercepts (α_j). Otherwise, the model is a standard single-level regression.

Equation 7a disaggregates main effects but will not properly disaggregate interactions (Balli & Sørensen, 2013; Giesselmann & Schmidt-Catran, 2022). Chapter D in the online supplemental materials covers this in more detail, but the general idea is that Equation 7a is

equivalent multiplying first then centering $(\text{SES}_{ij} \times \text{Non-White}_{ij}) - (\overline{\text{SES}_j \times \text{Non-White}_j})$ rather than centering first then multiplying $(\text{SES}_{ij} - \overline{\text{SES}_j}) \times (\text{Non-White}_{ij} - \overline{\text{Non-White}_j})$ where only the latter term disaggregates (Giesselmann & Schmidt-Catran, 2022, p. 1103). If within-cluster interactions are present in a fixed-effect model, the product must be manually cluster-mean centered to ensure proper disaggregation (the main effects can still be entered in their raw form because this only affects multiplicative terms). Equation 7b shows a properly disaggregated fixed-effect model with an interaction. Table 4 reports estimates from this model.[3]

$$\begin{aligned} \text{Math}_i = &\sum_{j=1}^{J} C_j\alpha_j + \beta_1\text{SES}_i + \beta_2\text{Non-White}_i + \\ &\beta_3(\text{SES}_{ij}^{\text{CMC}} \times \text{Non-White}_{ij}^{\text{CMC}}) + e_i \\ e_i \sim &N(0, \sigma^2) \end{aligned} \quad . \tag{7b}$$

In Table 4, the within-school SES achievement gap, $t(7022) = 17.35$, $p < .01$, and within-school non-White achievement gap, $t(7022) = -13.58$, $p < .01$, were significant but the within-school moderation effect was not, $t(7022) = -1.54$, $p = .12$. This indicates that—controlling for all school influences—the SES achievement gap is unaffected by whether the student reports a non-White identity. Table 4 also shows that school characteristics collectively explain 19.1% of the variance in math scores, but the model does not enumerate which school characteristics explain the most variance nor does it estimate the effects for any specific school characteristic. Between-school variance is simply factored out of the model.[4] Individual characteristics explain an additional 6.3% of the variance.

Estimates are not identical to earlier models because estimated the quantity being estimated is slightly different. The model with clustered errors estimated an aggregated moderation effect rather than within-school moderation in Table 4. The multilevel model (a) was conditional on school characteristics and (b) retained some unexplained school-level variance (i.e., the random intercept variance was nonzero) rather than controlling for all school characteristics as in the fixed-effect model.

Advanced Considerations

Clustered Errors: Conflated Coefficients

In the clustered error example, predictor variables were included in their raw form without centering. Uncentered predictors will not cause any statistical issues with clustered errors (i.e., standard errors will correctly quantify the sampling variability of the coefficients) but there may be substantive issues with clustering and uncentered predictors (i.e., the coefficients may be conflated).

For instance, with a predictor like SES, there may be a separate effect of an individual having high SES (a within-cluster effect) compared to the effect of attending a school comprised of mostly high SES students (a between-cluster effect). The predictor variable is the same (in this example, socioeconomic status [SES]), but a one-unit increase may have a different effect on the outcome variable depending on whether the characteristic of the person or the characteristic of the school increases.

Whereas conflated coefficients are problematic in multilevel models because there is an explicit interest in separating different levels of the hierarchy (e.g., Hoffman & Walters, 2022), conflated coefficients with clustered errors are more of a grey area because the analysis is neither strictly single-level nor multilevel. That is, the regression model itself is single-level whereas the data are multilevel. The idea of clustered errors is to remedy the discrepancy between model and the data structure, so there can be conflicting perspectives on conflated coefficients with clustered errors.

Historically, clustered errors are motivated by the statistical issue of correctly quantifying sampling variability in regression coefficients (Sanders & Konold, 2023, p. 7), so there may be instances where the aggregated effect remains the theoretical interest (Preacher et al., 2016, p. 190). This viewpoint corresponds to clustered errors being well-suited for a research question that could be addressed if one infinitely large classroom existed and clustering is a nuisance introduced by sampling limitations. In such a case, a researcher may not be interested in differentiating among different types of effects because this distinction may be theoretically irrelevant given that the potential existence of these within and between effects is a byproduct of the sampling design and unrelated to theory.

Nonetheless, this does not absolve the presence of conflated coefficients because—whether meaningful or a nuisance—clustering can create contextual influences that affect the interpretation of coefficients and it may be relevant to disaggregate these effects even if doing so is not a strict interest (A. Bell et al., 2019, pp. 1058–1059; Raudenbush & Bryk, 2002, p. 141).

Although multilevel models and fixed-effect models are historically more sensitive to substantive issues around conflated coefficients and more thought has been dedicated to tackling this issue

Table 4
Estimates for a Fixed-effect model Applied to the High School Beyond Data

Effect	Notation	Estimate	*SE*	*df*	*t*	*p*
SES	β_1	1.97	0.11	7,022	18.00	<.01
Non-White identity	β_2	−2.93	0.22	7,022	−13.24	<.01
SES × Non-White Identity	β_3	−0.47	0.31	7,022	−1.54	.12
Residual variance	σ^2	36.11				
R^2 school characteristics		19.1%				
R^2 student characteristics		6.3%				

Note. The model includes 160 school affiliation dummy variables and 160 corresponding school-specific intercepts are not reported in this table. SES = socioeconomic status

[3] Software output differs based on whether cluster affiliation coefficients are absorbed. Absorption removes the variance explained by a factor variable prior to estimating other coefficients. Computing standard errors requires inverting a matrix and, if there are many school affiliation coefficients, this inversion can be computationally demanding. Absorption retains the variance explained by all cluster affiliation dummies, but specific effects for each cluster and their standard errors are omitted to facilitate computation. The `lm` R function and SAS `PROC REG` do not absorb. SAS `PROC GLM` depends on whether the cluster ID variable is included in a `CLASS` statement (which does not absorb) or in an `ABSORB` statement (which does absorb). The `plimm` R package absorbs.

[4] Software output displays a single R^2 value and does not split the variance explained into cluster affiliation substantive components as in Table 4. This split can be calculated manually by subtracting the R^2 of a model with only the cluster affiliation dummies from the R^2 of the full model. In this example, the full model R^2 was 25.5% and the cluster affiliation-only model R^2 was 19.1%. The `lm` function in R does not correct sums of squares when the intercept in suppressed (e.g., Kvålseth, 1985), so R^2 should use a reference cluster specification in the `lm` function.

in those modeling frameworks, it is possible to disaggregate a single-level model with clustered errors. The "Blending Methods Together" section described shortly will show how to import ideas from multilevel models into analyses with clustered errors to demonstrate that clustered errors do not necessarily lock researchers into aggregated coefficients if their theoretical interest calls for separating within and between effects.

Clustered Errors: Generalized Estimating Equations and Generalized Least Squares

The classical approach to clustered errors starts from a model where the errors are assumed to be independent (i.e., $\mathbf{\Omega} = \sigma^2\mathbf{I}$) prior to correcting the standard errors based on dependencies in the empirical errors. However, this approach can encounter difficulties when the ICC is high (e.g., above .30; Zeger et al., 1988) or if there is slope heterogeneity because the off-diagonal terms will be quite different from 0, so the correction will be rather large (Huang, 2022). Clustered errors can be augmented to adopt a different baseline assumption to (a) build in intermediate steps, (b) reduce the reliance on the empirical covariance matrix, and (c) reduce the size of the correction.

This approach is referred to as *generalized estimating equations* (GEE) in biostatistics (Liang & Zeger, 1986; Zeger & Liang, 1986) or *feasible generalized least squares* (FGLS) in econometrics (e.g., Hansen, 2007). These methods are conceptually similar (e.g., Cameron & Miller, 2015, p. 355; McNeish, 2019, Footnote 2), but not always identical (e.g., GEE accommodates discrete outcomes; Ziegler & Vens, 2014).

The idea is to begin by assuming nonzero correlations between observations in the same cluster. For instance, instead of beginning from the premise that the errors are independent, one could begin by assuming that errors from all observations in a cluster are equally correlated (i.e., an exchangeable or compound symmetric structure). The baseline assumption is called the working correlation (in GEE) or covariance (in FGLS) structure. The model is estimated using the working structure for how errors covary for observations in the same cluster. Then, clustered errors are applied to protect inferences such that standard errors are accurate even if the working structure is incorrect (Diggle et al., 2002). Classical clustered errors are a special case where the working structure is a scalar matrix that is incorrect by implying independence among the errors.

An exchangeable working structure is typically suitable for organizationally clustered data with large ICCs (Ballinger, 2004). For instance, for small six-person example discussed earlier, an exchangeable working correlation would look like

$$\begin{bmatrix} 1 & & & & & \\ \rho & 1 & & & & \\ \rho & \rho & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & \rho & 1 & \\ 0 & 0 & 0 & \rho & \rho & 1 \end{bmatrix}. \tag{8}$$

The structure is block diagonal where observations in the same cluster are related but observations in different clusters are independent. The magnitude of the dependence within clusters is constrained to be equal (i.e., there is no subscript on ρ). This equality applies to all observations in a cluster (e.g., errors for persons 1, 2, and 3 are equally correlated) and across all the clusters (e.g., the within-cluster correlations in Cluster 1 is the same as the within-cluster correlations in Cluster 2).

The value of ρ is determined empirically from the errors of a model assuming independence (Liang & Zeger, 1986, pp. 17–18). Then the model is reestimated assuming that observations within the same cluster have a correlation of $\hat{\rho}$ rather than being independent. Standard errors would reflect the fact that data are clustered but will not necessarily be correct unless the exchangeable correlation matrix is exactly correct. The clustered error process can then be applied such that empirical errors ($\hat{\mathbf{e}}_j$) are taken from the model assuming observations have a correlation of $\hat{\rho}$ (rather than 0 as required by the independence assumption).[5] This will produce standard errors that account for clustering and that are not reliant on the selected working structure being exactly correct.

As a summary of the estimation process using an exchangeable working correlation structure,

1. Estimate the regression coefficients assuming independence.
2. Use the errors in Step 1 to estimate the correlation (ρ) between observations in the same cluster.
3. Reestimate the model assuming that observations in the same cluster have a correlation of $\hat{\rho}$.
4. Use the empirical covariance of the errors from Step 3 to calculate standard errors that are robust to incorrect selection of the working structure.

The end goal of GEE, FGLS, and the classical clustered error approach is the same—produce standard errors that accurately quantify sampling variability of the regression coefficients. Classical clustered errors go from Step 1 directly to Step 4 because the assumption is that observations are independent, so there are no off-diagonal terms to estimate. This also means that the classical approach entrusts the entire correction to the empirical covariance matrix. GEE and FGLS add Steps 2 and 3 to help the correction be more efficient (Cui & Qian, 2007). These additional steps build some aspects of clustering into the standard errors in Step 3 so that the empirical covariance matrix in Step 4 does not bear the full brunt of the correction.

Multilevel Models: Heterogenous Variance Models

Regression modeling in psychology revolves around the mean and coefficients correspond to the expected change in the mean of the outcome when a predictor increases by one unit. Multilevel models partition the variance into different variance components, so it is possible to build models that predict how the *variance* changes as a function of predictors (Hedeker et al., 2008; Hoffman, 2007). These models recently have been proposed for (a) modeling cohesion of work teams (Lang et al., 2018, 2019; McNeish, 2021), (b) meta-analysis (Viechtbauer & López-López, 2022), (c) repeated trials in cognitive psychology (Williams et al., 2019, 2021), or (d) ecological momentary assessment data (Hedeker et al., 2012; McNeish & Hamaker, 2020; Rast & Ferrer, 2018).

The models are referred to as *heterogeneous variance* or *location-scale* models because they simultaneously model aspects of mean change (the location) and variance change (the scale). The basic idea

[5] Clustered errors are applied by default with GEE but are a secondary step with FGLS. So, "FGLS with clustered errors" makes sense, but "GEE with clustered errors" does not because clustered errors are implied GEE.

is that—in a traditional multilevel model—any parameter in the within-cluster submodel can be an outcome in the between-cluster submodel. This was seen directly in Equation 6 where the regression coefficients (β) in the within-cluster submodel became outcomes in the between-cluster submodel. However, the residual variance (σ^2) is also a parameter in the within-cluster submodel, so it too could take a *j* subscript and be modeled to vary across clusters as a function of predictors.

This type of model could feature σ_j^2 as an between-cluster outcome and then predict differences in, for instance, variability across teams to assess hypotheses team cohesion (where "cohesion" is represented by small within-team variance). With ecological momentary assessment data, the same idea can be applied to assess what predictors make a time series more stable or volatile. A similar process can also be conducted to capture heterogeneous between-cluster variances (the τ terms; Hedeker et al., 2008, 2012). Fitting location-scale models can require extra care because they often are nonlinear given that variances cannot be negative (e.g., models may require `PROC NLMIXED` in SAS or `nlme` in R. The `LOCAL` option in `PROC MIXED` provides limited support of these models as does the `form =` option in the `random` or `weights` option in the `lme` R function).

Fixed-Effect Models: Incorporating Slope Heterogeneity

The fixed-effect model example included cluster affiliation predictors as main effects, which is the fixed-effect model analog of a random intercepts model. This assumes that slopes of the within-cluster predictors are constant across schools. However, fixed-effect models can also be specified in a way that permits heterogeneity in the associations of within-cluster predictors and the outcome (i.e., the analog of a random slopes model).

Instead of modeling slope heterogeneity with random effects as in a multilevel model, a fixed-effect model can include interaction terms among cluster affiliation dummies and a predictor. This results in *J* slope coefficients per predictor, one per cluster. Unlike a multilevel model where a subsequent focus is to build a between-cluster submodel for the slope heterogeneity, cluster-specific coefficients in a fixed-effect model already account for all collected and uncollected sources of between-cluster variance. In a fixed-effect model, the cluster-specific coefficients are directly estimated, meaning that there is no distributional assumptions or variance components associated with the predictor but that estimates do not generalize beyond the clusters included in the data.

If the fixed-effect model in Equation 7b were specified to incorporate slope heterogeneity for non-White identity and SES (without the interaction term), the model would be written as

$$\begin{aligned} \text{Math}_i &= \sum_{j=1}^{J} C_j\alpha_j + \sum_{j=1}^{J} \beta_{1j}(\text{SES}_i \times C_j) \\ &\quad + \sum_{j=1}^{J} \beta_{2j}(\text{Non-White}_i \times C_j) + e_i \\ e_i &\sim N(0, \sigma^2) \end{aligned} \tag{9}$$

such that both predictors interact with all cluster affiliation dummies. There are 160 clusters in these data, so the model would attempt to estimate 160 coefficients for SES (one for each cluster) and 160 coefficients for non-White identity (one for each cluster). Note that cluster-specific coefficients are not estimable when there is no within-cluster variability. In these data, 20 schools have no variability in non-White identity (i.e., students in a school exclusively identify as White or exclusively identify as non-White), so there are only 140 cluster-specific coefficients for non-White identity in these data.

As specified in Equation 9, there are no main effects for either SES or non-White identity. This specification will directly estimate the effect in each cluster. If a main effect were included, one of the cluster affiliation dummies would be omitted and serve as the reference cluster. The main effect would correspond to the cluster-specific estimate of this arbitrary reference cluster and the other cluster-specific coefficients would represent the difference between the cluster-specific slope in cluster *j* and the cluster-specific coefficient for the reference cluster. To be clear, the main effect would *not* be the average effect across all clusters. If the average effect across all clusters were sought, this could be calculated by taking an average of the cluster-specific coefficients, weighted by the number of people within the cluster because fixed-effect model coefficients are unpooled.

Blending Methods Together

To this point, approaches have been discussed as if they are mutually exclusive. However, the three methods form a general framework for handling clustered data and benefits from one method can be mimicked or integrated into another method to blend strengths of different approaches into a more coherent overall model that maximizes robustness and minimizes weaknesses of applying each method in isolation. Each permutation of methods is covered in a dedicated subsection.

Clustered Errors and Multilevel Models

Disaggregating Effects With Clustered Errors

Previously, it was noted that there is some ambiguity about conflated coefficients with clustered errors, but the same centering and specification principles from multilevel models can be applied with clustered errors to disaggregate effects. Raudenbush and Bryk (2002) disaggregated effects in single-level models to obtain correct coefficient estimates of within and between effects (p. 141) but noted that the standard errors will be incorrect. McNeish (2019) showed that clustering the errors in a disaggregated single-level model produces accurate estimates of within and between effects with accurate standard errors. Therefore, conflated coefficients are a consequence of centering decisions and model specification, not necessarily the method by which clustering is accommodated. More plainly, disaggregation is not restricted to multilevel models and can be applied with clustered errors (Begg & Parides, 2003; Goetgeluk & Vansteelandt, 2008).

To demonstrate, consider a single-level model for the high school beyond data that is similar to the earlier multilevel model example in Equation 6,

$$\begin{aligned} \text{Math}_{ij} &= \beta_0 + \beta_1\text{SES}_{ij}^{(\text{CMC})} + \beta_2\text{Non-White}_{ij}^{(\text{CMC})} \\ &\quad + \beta_3(\text{SES}_{ij}^{(\text{CMC})} \times \text{Non-White}_{ij}^{(\text{CMC})}) \\ &\quad + \beta_4\overline{\text{SES}}_j + \beta_5\overline{\text{Non-White}}_j^{(\text{GMC})} \\ &\quad + \beta_6(\overline{\text{SES}}_j \times \overline{\text{Non-White}}_j^{(\text{GMC})}) + \beta_7\text{Private}_j \\ &\quad + \beta_8(\text{SES}_{ij}^{(\text{CMC})} \times \text{Private}_j) \\ &\quad + \beta_9(\text{Non-White}_{ij}^{(\text{CMC})} \times \text{Private}_j) + e_{ij} \end{aligned} \tag{10}$$

This model cluster-mean centers all within-school predictors and includes schools means as between-school predictors. The moderation

effect is disaggregated as well. The only difference is that Equation 10 does not model heterogeneity in the coefficients (i.e., the β terms do not have *j* subscripts). As the random slope variability in Table 3 was rather large, GEE with an exchangeable working structure is used rather than the classical clustered errors to improve efficiency of the correction.

Table 5 compares fitting the single-level model in Equation 10 with GEE and fitting the multilevel model from Equation 6. The coefficient estimates and standard errors are nearly identical, corroborating previous claims that it is possible to disaggregate predictors without a multilevel model.

As there is no slope heterogeneity with clustered errors, the two models are not generally equivalent and multilevel models retain advantages if quantifying slope heterogeneity is an interest. Because clustered errors do not partition the variance, they also have more limited variance explained options than multilevel models (see Chapter B in the online supplemental materials for more details on variance explained). Coefficients from clustered errors are completely pooled whereas coefficients from multilevel models are partially pooled, so estimates can diverge if the number of people per cluster varies widely and/or random slopes have large variances because an exchangeable working structure becomes less effective for accommodating complex covariance structures (J. W. Twisk, 2003). As a reference, cluster sizes in these data ranged from 14 to 67, but differences were minimal.

Nonetheless, if the research question concerns disaggregated effects but not necessarily heterogeneity, clustered errors can be serviceable to differentiate among within, between, and contextual effects.

Clustering Errors in Multilevel Models

Multilevel models allow for the broadest possibilities of the three methods covered, but this flexibility comes with additional assumptions. For instance, random coefficients require a distribution to be specified. Normality is typically assumed but may be violated in empirical data (e.g., Alonso et al., 2010). When normality of the random coefficients is not upheld, fixed effect standard errors and variance component estimates may be too large (Litière et al., 2007; Schielzeth et al., 2020) and cluster-specific coefficients may be inaccurate (McCulloch & Neuhaus, 2011a).

These problems tend to be mild and dissipate as the number of clusters increases when the outcome is continuous (McCulloch & Neuhaus, 2011b). Having 100 or more clusters tends to be sufficient for robustness to mild or moderation normality violations (Jacqmin-Gadda et al., 2007; Verbeke & Lesaffre, 1997), but the assumption should be assessed to identify gross violations of normality (Schielzeth et al., 2020). The assumption that the random effect distribution is correct also extends to selecting the correct covariance structure, even if normality is reasonable (Wolfinger, 1993). This means that covariances between random effects are properly modeled and that the correct number of random effects have been included, which can increase the risk for nonpositive definite estimated covariance matrices or other convergence-related issues increases if there are several random effects (Bates et al., 2015, p. 18; Eager & Roy, 2017; McNeish & Bauer, 2022).

To protect against possible violations of the random effect distributional assumptions (e.g., normality, a misspecified covariance structure, or omitted random effects), it is possible to use clustered errors within a multilevel model (Maas & Hox, 2004). Similar to making standard errors robust to independence violations (or, robust to covariance structure misspecification as in GEE or FGLS) in single-level models, clustering the errors in a multilevel model can make standard errors robust to random effect distributional assumptions (Gurka et al., 2011; Verbeke & Lesaffre, 1997).

Multilevel models imply a covariance matrix for observations within the same cluster—which is informed by model assumptions—and is used to quantify the sampling variability of the fixed effects. The estimated sampling variability based on the model-implied covariance matrix will be correct to the extent that assumptions are correct. However, as with single-level models, multilevel models can replace the model-implied covariance matrix with an empirical covariance matrix. Just as clustered errors can provide inference that is robust to single-level assumption violations like homoskedasticity and independence, clustered errors in multilevel models can provide inference that is robust to mild-to-moderate multilevel assumption violations like omitted random slopes or nonnormality of random effects. Chapter C in the online supplemental materials describes clustering errors in multilevel models.

Table 5

Comparison of Estimates and Standard Errors for a Linear Regression Model Using Cluster-Mean Centering With Clustered Errors Using an Exchangeable Working Correlation Matrix and a Multilevel Model With Cluster-Mean Centering

	Estimate		*SE*	
Effect	CE	MLM	CE	MLM
Student characteristics				
Intercept	11.85	11.85	0.19	0.20
SES	2.38	2.39	0.16	0.16
Non-White identity	−3.96	−3.98	0.36	0.36
SES × Non-White Identity	−0.59	−0.67	0.33	0.31
School characteristics				
SES school mean	4.40	4.41	0.41	0.42
Non-White school mean	−2.61	−2.61	0.60	0.57
SES School Mean × Non-White School Mean	−0.99	−1.02	1.00	1.03
Private	1.64	1.64	0.29	0.30
Private × SES School Mean	−1.03	−1.03	0.23	0.23
Private × Non-White School Mean	2.09	2.12	0.49	0.50

Note. CE = clustered errors; MLM = multilevel model; SES = socioeconomic status.

Multilevel Models and Fixed-Effect Models

Relaxing Exogeneity Assumptions in Multilevel Models

A key assumption in multilevel models is *exogeneity* (Antonakis et al., 2010), also called the *zero conditional mean* assumption (Grilli & Rampichini, 2011). In single-level models, a version of this assumption is made, which states that the errors are not systematically related to the predictors (i.e., $E(\mathbf{e}|\mathbf{X}) = \mathbf{0}$). Multilevel models include multiple error terms, so this assumption extends not just to the within-cluster error (r_{ij}) but also to the random effects (u_j), meaning that predictors at either level cannot be related to error terms at either level (Antonakis et al., 2021). Practically, this assumption is violated when the functional form of predictors is incorrect (e.g., nonlinear effects modeled as linear; Bauer & Cai, 2009) or when relevant predictors are omitted (Kim & Frees, 2006, 2007; Kim & Swoboda,

2010; Tofighi & Kelley, 2016), particularly in the between-cluster submodel. Essentially, multilevel models offer more possibilities, but require that all submodels are properly specified.

A primary benefit of fixed-effect models is that the between-cluster submodel is immune from misspecifications. The cluster affiliation dummies necessarily explain all the between-cluster variance, so there is no mechanism for between-cluster submodel misspecifications to affect the within-cluster submodel. That is, fixed-effect models limit the scope of exogeneity. However, the tradeoff with a fixed-effect model is that effects of specific predictors in the between-cluster submodel are inestimable. As noted by A. Bell et al. (2019), this can limit the utility of fixed-effect models because they "reveal almost nothing about the level-2 entities in the model ... and only ever present a partial picture of the substantive phenomenon represented by the model" (p. 1058).

However, this property of fixed-effect models can be recreated in a multilevel model without sacrificing the ability to estimate effects of specific predictors in the between-cluster submodel (A. Bell & Jones, 2015; Dieleman & Templin, 2014; Hazlett & Wainstein, 2022; McNeish & Kelley, 2019). In a multilevel model, this can be accomplished by (a) cluster-mean centering all predictors in the within-cluster submodel and (b) including the cluster means of all within-cluster predictors as predictors of the intercept in the between-cluster submodel. This specification has been referred to as the *within-between specification* (A. Bell & Jones, 2015; Dieleman & Templin, 2014; McNeish & Kelley, 2019) or the *bias-corrected multilevel model* (Hazlett & Wainstein, 2022) and is related to the Mundlak (1978) and Chamberlain (1982) devices from econometrics, where it is better known as the *correlated random effects* approach (Schunck, 2013; Wooldridge, 2010).

Specifying a multilevel model this way artificially forces the two submodels to be orthogonal (Hamaker & Muthén, 2020; Hazlett & Wainstein, 2022), mimicking the process in fixed-effect models but by a different mechanism. A fixed-effect model explains all between-cluster variance with cluster affiliation dummies, making between-cluster information orthogonal to within-cluster information (i.e., if unexplained between-cluster variance is zero, any covariance is also necessarily zero). A within-between multilevel model does not explain all between-cluster variance, but centering and including cluster means as predictors similarly imposes zero covariance between submodels.

In doing so, omitted variables or model misspecification in the between-cluster submodel cannot permeate to the within-cluster submodel. This narrows the scope of exogeneity because within-cluster estimates remain accurate so long as the within-cluster submodel is properly specified, which is identical to assumptions of a fixed-effect model. However, because the between-cluster variance is not explained by this specification, it is still possible to estimate effects of specific between-cluster predictors.

As an example, consider estimating the within-cluster main effects of non-White and SES math achievement gaps in the High School Beyond data (the interaction term from previous models is omitted to facilitate comparisons). Imagine that the research question is aligned with a fixed-effect model such that the goal is to isolate the within-cluster effect while completely removing effects of any collected or uncollected between-cluster predictors. Table 6 compares estimates from a fixed-effect model and a multilevel model with a within-between specification as fit in SAS `PROC GLM` and SAS `PROC MIXED`, respectively. The coefficient estimates, standard errors, and test statistics for the within-cluster SES achievement gap and non-White achievement gap are identical. In this multilevel specification, school means of SES and non-White identity serve the same purpose as the cluster affiliation dummies in a fixed-effect model, absorbing potential misspecifications in the between-cluster submodel such that the within-cluster submodel is independent and isolated from possible between-level misspecification.[6]

Table 6
Comparison of Estimates and SEs for a Multilevel Model With Cluster-Mean Centering and Cluster Means Included as Predictors and a Fixed-Effect Model

	Estimate		*SE*	
Effect	MLM	FEM	MLM	FEM
Student characteristics				
Intercept	13.11	—	0.21	—
SES	1.95	1.95	0.11	0.11
Non-White identity	−2.90	−2.90	0.22	0.22
Residual variance	36.14	36.12		
School characteristics				
SES school mean	5.33	—	0.40	—
Non-White school mean	−1.54	—	0.55	—
Intercept variance	2.56	—		

Note. MLM = multilevel model; FEM = fixed-effect model; SES = socioeconomic status.

Importantly, the multilevel model permits specific between-cluster predictors and allows quantifying slope heterogeneity. That is, the mechanism used to make submodels orthogonal in a within-between multilevel model reduces the scope of exogeneity assumptions (similar to fixed-effect models) without conceding the ability to estimate effects in the between-cluster submodel (unlike fixed-effect models). Note that between-cluster effect estimates are sensitive to omitted or uncollected between-cluster variables (A. Bell et al., 2019, p. 1059).

Of course, multilevel models assume that clusters are randomly sampled from the population, so the fixed-effect model retains advantages for situations where the target of inference is a specific group of clusters or if clusters are not representative of the population. Nonetheless, protecting effects in the within-cluster submodel from potential misspecifications in the between-cluster submodel is not exclusive to fixed-effect models and can be recreated in a multilevel model.

Hybrid Models for Incidental Clustering

Certain levels of clustering may not be relevant to the research questions. For instance, the interest may be quantifying heterogeneity in the SES achievement gap and identifying possible school characteristics that explain this heterogeneity. However, to recruit a sufficient number of schools, schools are clustered in multiple districts such that there is a three-level hierarchy (students within schools within districts). The research questions may only concern

[6] There are special cases where grand-mean centering or no centering can produce the same within effect if cluster means are included as predictors (e.g., Hamaker & Grasman, 2015, Section 2.1; Snijders & Bosker, 2012, Ch 5). However, such a specification does not ensure that the within-cluster submodel is orthogonal to the between-cluster submodel (Hamaker & Muthén, 2020, p. 377), especially with random slopes (Snijders & Bosker, 2012, Section 5.3).

the school level such that the district level is incidental, and the research questions could be answered without the district level.

Existing studies find that incidental levels cannot be ignored without adverse statistical ramifications unless the DEFT is small (Moerbeek, 2004; Pornprasertmanit et al., 2014; Tranmer & Steel, 2001; van den Noortgate et al., 2005). Unlike a two-level hierarchy where ignoring the second level inflates Type-I error rates, ignoring the third level of a three-level hierarchy tends to decrease power (Moerbeek, 2004), so it is in researchers' interest to model incidental levels.

A three-level multilevel model could be employed such that the variance is partitioned in student, school, and district components. Although possible, there may be practical limitations for three-level models with incidental levels. For instance, incidental levels often have few clusters and few collected variables given that no research questions exist at the level, which can exogeneity difficult to satisfy.

Alternatively, different levels of the hierarchy could be handled differently, depending on the relevance of each level to the research question (McNeish & Kelley, 2019). For instance, it may be simplest to completely control for the district level to ensure that district-level influences do not distort the student-level or school-level estimates. In this case, a "hybrid" model could be built such that district affiliation dummies are included in the model to absorb all sources of district-level variance, but the student-level and school-level variance are explicitly modeled with a multilevel model (McNeish & Wentzel, 2017). The overall model is not strictly a multilevel or fixed-effect model, but instead incorporates different approaches to accommodate each level in a manner that is consistent with its relevance to the research questions. If the district level is wholly uninteresting, including district affiliation dummies limits exposure to multiple issues (sample size, random sampling of clusters, exogeneity, normality) given that fixed-effect models require fewer assumptions.

Fixed-Effect Models and Clustered Errors

This combination may not be intuitive initially—cluster affiliation dummies in a fixed-effect model make the errors conditionally independent, so there would be no remaining within-cluster correlation for clustered errors to correct. This is accurate in the purest form of clustered data where there is only one source of within-cluster correlation. However, real data are not always pristine, so it is possible that the errors may not be entirely independent despite including cluster affiliation dummies if there is another source of clustering other than the level whose dummies have been included in the model (e.g., Bertrand et al., 2004). In this case, clustering the errors in a fixed-effect model can correct for any unforeseen sources of clustering beyond the level whose cluster affiliation dummies were included.

For instance, if school affiliation dummies are included but there are neighborhood influences that do not overlap with school affiliation, the errors may still be dependent from the unmodeled neighborhood level. In this case, clustering the errors could clean up unintended dependence that exists beyond the primary source of clustering (Pustejovsky & Tipton, 2018).

Alternatively, school affiliation dummies may be used but perhaps students are further clustered within classrooms. Clustering the errors on top of adding school affiliation dummies would result in accurate standard errors if there were dependence due to classrooms (Lee & Pustejovsky, 2023). As another example, errors from a fixed-effect model may be heteroskedastic, so clustered errors could be applied to ensure proper inference if regression assumptions beside independence are not upheld. Combining fixed effects and clustered errors is also more common in longitudinal data to address serial dependence (i.e., high correlations between nearby timepoints; Arellano, 1987; Moody & Marvell, 2020).

Considerations for Longitudinal Data

So far, the focus has been on data with people clustered within organizational units. However, longitudinal data are a special case of clustered data where repeated measures are clustered in people. Many of the principles discussed previously apply to longitudinal clustering; however, there are some special considerations resulting from design and data structure differences between organizationally and longitudinally clustered data. This section overviews main differences for each method.

Clustered Errors

Classical clustered errors are less useful in longitudinal data because ICCs are often higher than in organizationally clustered data and heterogeneity in growth trajectories is common. As a result, GEE or FGLS tend to be more useful with longitudinal data (Hin & Wang, 2009; Pan & Connett, 2002). In longitudinal data, the main utility of GEE or FGLS is to estimate the marginal or population-averaged growth trajectory, which captures the overall trend across people. This is opposed to multilevel models where person-specific growth trajectories are estimated. Advantages and disadvantages of marginal versus person-specific growth models have been widely debated in biostatistics (e.g., Heagerty & Zeger, 2000; Hu et al., 1998; Hubbard et al., 2010; Zeger et al., 1988), but the main takeaway is the GEE and FGLS are ideally suited for research questions focusing on what variables influence baseline values or the rate of change of the *average* trajectory.

A complication with GEE and FGLS in longitudinal data is that there are more sensible options for the working structure. If students are clustered within schools, there is not typically a strong reason to suspect that Student 1 and Student 2 will be more related than Student 1 and Student 3. People are roughly interchangeable with respect to contextual influences, so an exchangeable working structure is sufficient for many organizationally clustered data. However, observations in longitudinal data are timepoints, so it is often more plausible that Time 1 and Time 2 are more strongly related than Time 1 and Time 3 because they are closer together in time (i.e., time may not be exchangeable).

An exchangeable working structure can be used with longitudinal data, which would assume that correlations between timepoints are constant regardless of the distance between timepoints. However, other structures are worth considering such autoregressive/Markov (correlations between timepoints decay systematically as timepoints are further apart), Toeplitz (correlations between timepoints decay as timepoints are further apart but not necessarily systematically) or unstructured (correlations between timepoints do not follow a discernible pattern).

GEE does not use the full likelihood during estimation, so GEE is only robust to data that are missing completely at random and the popular full-information maximum likelihood method cannot be applied. However, weighting methods (DeSouza et al., 2009; Fitzmaurice et al., 1995; Preisser et al, 2002) or multiple

imputation (Lipsitz, et al., 2004; Paik, 1997; J. Twisk & de Vente, 2002) have been developed to accommodate missing at random data that are commonly associated with dropout or attrition in longitudinal studies. Note that likelihood-based versions of GEE can be implemented (Molenberghs & Kenward, 2007) and have been referred to as *covariance pattern* models (Jennrich & Schluchter, 1986). Covariance pattern models are essentially a maximum likelihood version of FGLS and can similarly be paired with clustered errors.

Multilevel Models

With longitudinal data, the between-person submodel corresponds to individual differences in growth trajectories because the cluster variable is a person. Random coefficients therefore capture heterogeneity in baseline values or growth rates. Correspondingly, multilevel models are ideally suited for research questions interested in quantifying and explaining individual differences in growth trajectories over time given that multilevel models directly quantify heterogeneity in growth parameters (Singer, 1998; Zeger et al., 1988). With longitudinal data, ideas behind multilevel models are sometimes expressed in a structural equation framework where they are called latent growth models (Bauer, 2003; Bollen & Curran, 2006; Curran, 2003; Grimm et al., 2016), although minor differences between multilevel models and latent growth models do exist (McNeish & Matta, 2018).

In the longitudinal context, person-mean centering is common to disaggregate within-person and between-person effects and distinguish between momentary and habitual changes (Curran & Bauer, 2011; Hamaker & Grasman, 2015). There is also a greater focus on the within-person residual covariance structure. With organizationally clustered data, the within-cluster residual variance is typically modeled as being constant across all people in the same cluster (σ^2). However, the within-person residual variance in longitudinal data refers to time, so there are more potentially sensible options (Hoffman, 2015).

A main distinction in longitudinal data is whether the within-person residual variance is constant or varies over time (Grimm & Widaman, 2010; Kwok et al., 2007). Time tends to be the main focal predictor in longitudinal data, so there are often more interactions with predictors and time or more complicated functional forms for the predictors or for time to better capture how the outcome changes (e.g., polynomials, nonlinear models; Cudeck & Harring, 2007).

Missing longitudinal data are more likely to be missing not at random (MNAR) due to dropout and attrition related to the outcome. For example, participants in a depression study may drop out because their depression is too high to continue participating (i.e., dropout is related to the value that would have been reported if the participant remained in the study). Specialized MNAR models exist such as Diggle–Kenward selection models (Diggle & Kenward, 1994) or pattern mixture models (Hedeker & Gibbons, 1997). Enders (2011) provides an accessible overview of MNAR models in psychology.

Fixed-Effect Models

The prevailing utility of fixed-effect models remains the ability to completely control for collected or uncollected sources of between-cluster variance. In longitudinal data, this means that fixed-effect models control for time-invariant sources of variance and isolate time-varying sources (Allison, 2009; A. Bell & Jones, 2015). This may be particularly helpful in observational studies where an intervention cannot be randomly assigned or situations where potential confounders were not or could not be collected (Gunasekara et al., 2014; Kaufman, 2008; Neuhaus & Kalbfleisch, 1998). Fixed-effect models allow causal claims under weaker assumptions with longitudinal data because they rule out any possible time-invariant sources of confounding (Brüderl & Ludwig, 2015).

One caveat for fixed-effect models in longitudinal data is that effects of specific time-invariant effects are not estimable. This may be problematic if intervention effects are time-invariant (i.e., intervention group assignments are constant over time). Although, a main interest is often whether *growth* is different between intervention conditions rather than whether groups are different at baseline. Intervention effects on growth trajectories *can* be directly estimated with a Treatment × Time interaction, which is not collinear with the person-affiliation dummies because the interaction has time-varying variance through inclusion of the Time variable.

Fixed-effects models can be applied in conjunction with multilevel models to control incidental levels of clustering. For instance, if the focus is on students' growth but students happen to be clustered within schools (a three-level hierarchy), school affiliation dummies can be included to control for collected and uncollected school characteristics to ensure that the student-level estimates are not affected by school-level differences. As noted earlier, longitudinal data provide more opportunities to blend fixed-effect models with clustered errors because there is greater concern that the within-person covariance matrix has been properly modeled. That is, errors of different timepoints within the same person often have a more complicated correlation structure than errors between people clustered within the same organization, so there is greater utility in estimating standard errors in a way that provides some protection against within-person covariance misspecifications.

Considerations With Three-Level or Cross-Classified Hierarchies

The focus has been on the simplest and most common two-level case where individuals are nested within one organizational unit, but individuals can be simultaneously clustered within more than one organizational unit (a.k.a. *multiway* clustering, especially in econometrics; Cameron et al., 2011). When there are two organizational level units, clustering can be *nested* or *cross-classified*.

A nested hierarchy with two organizational units occurs when membership in one organizational unit implies membership in another higher organizational unit. For instance, if students are clustered in schools which are clustered within districts (e.g., each school has multiple students, each district has multiple schools), knowing a student's school also reveals their district because schools are nested within districts such that all students belonging to a particular school also belong to the same district.

Contrast this with a cross-classified hierarchy like students being clustered within schools and neighborhoods (Rasbash & Goldstein, 1994). Due to different school types (public, private, magnet, etc.), school attendance is not strictly based on geography. So, knowing that two students attend school together does not imply that they live in the same neighborhood and knowing that two students live

in the same neighborhood does not imply that they attend school together.

If trying to determine which type of hierarchy is present, nested hierarchies easily allow rank ordering of levels. That is, it is clear that Level-3 is composed of Level-2 units and Level-2 is composed of Level-1 units. In cross-classified hierarchies, the two organizational units are harder to rank order, so it makes more sense to label the different units as Level-2a and Level-2b rather than Level-2 and Level-3. For example, neighborhoods are not necessarily composed of schools and schools are not composed of neighborhoods; both levels are composed of students but the structure of how students are arranged into schools is not a function of neighborhood (and vice versa).

Hierarchies with more than one organizational unit are straightforward to accommodate with multilevel models because the variance is simply partitioned into additional sources (Goldstein, 1994; Grady & Beretvas, 2010). Models with additional levels do require more assumptions (e.g., exogeneity across more levels, distributional assumptions of more random effects) and performance can deteriorate when assumptions are not upheld (Lee & Pustejovsky, 2023).

Clustered errors and fixed-effects models can accommodate more than one organizational unit, often with fewer assumptions. An overview of these approaches with multiple organizational units is discussed next.

Nested Three-Level Hierarchies

Clustered Errors

With clustered errors (or GEE or FGLS), three-level nested data are straightforward to accommodate because the cluster variable is just the highest level of the hierarchy (Cameron et al., 2011; Pepper, 2002). If students are clustered in schools which are clustered in districts, clustering the errors by district (Level-3) appropriately quantifies the sampling variability (i.e., Level-2 and Level-3 are corrected simultaneously). This approach is effective because the empirical covariance used in the clustered error correction is unstructured within a cluster such that off-diagonal terms are not constrained to follow a particular pattern. Therefore, off-diagonal elements in block j of $\mathbf{\Omega}$ can be larger for observations that share the same Level-2 unit than for observations from different Level-2 units. Essentially, clustering at Level-3 naturally builds in clustering at Level-2 as a byproduct.

As one caution, samples sizes tend to be smaller at Level-3 because this level tends to be entities like school districts or a governing entity like towns, counties, or states. Effectiveness of clustered errors is based on having many clusters, so small sample corrections may be especially pertinent in three-level data (corrections can be found in the `EMPIRCAL` option in SAS `PROC GLIMMIX` or the `clubSandwich` R package; Pustejovsky, 2020). Corrections are applicable with classical clustered errors, GEE, or FGLS (although software may differ for different approaches). An exchangeable working structure is typically reasonable for GEE or FGLS with three-level organizationally clustered data. Sample code is provided in the online supplemental materials.

Fixed-Effect Models

An intuitive approach to fixed-effect models with three levels may be to create dummies for both the Level-2 and Level-3 units and include both sets of dummies as predictors. However, this would exhaust degrees of freedom with Type-III sums of squares (A. Bell et al., 2019; McNeish & Kelley, 2019) because the Level-3 dummies would be collinear with the Level-2 dummies.

A simpler approach is to include only the Level-2 dummies. The reasoning is similar to the mechanism for clustered errors, only in the reverse direction. The Level-2 dummies fully explain the variance at Level-2, so there is no Level-2 remaining variance to partition among Level-3 units. Dummies for the lowest between-cluster level absorb variance of higher between-cluster levels. If there is a direct interest in the Level-3 dummy estimates in each cluster, dummies for both levels can be included with Type-I sum of squares, although this will produce multiple reference clusters (one per Level-3 unit), so some care may be required for proper interpretation (e.g., McCaffrey et al., 2012).

Cross-Classified Hierarchy

Clustered Errors

Clustered errors are not as straightforward to apply to cross-classified data as they are with three-level nested data because Level-2a and Level-2b both need to be considered when the organizational units are not nested. Full detail will not be provided in this article (see Cameron et al., 2011 or Lee & Pustejovsky, 2023 for specific details about the computational mechanism), but the main idea is that the empirical covariance for cross-classified clustered errors is the sum of three separate calculations—clustering the errors based on Level-2a, clustering the errors based on Level-2b, and clustering the errors on the intersection of Levels-2a and 2b.

Cross-classified clustered errors can be implemented in the `lfe` (Gaure, 2013), or `fixest` (Bergé, 2018) R packages. The online supplemental materials provide example code for cross-classified data. To the author's knowledge, no SAS procedure currently allows for cross-classified clustered errors.[7] GEE has origins in longitudinal clustering where cross-classification in less common, so there tends to be less support for GEE with a cross-classified structure. As of this writing in August 2023, the `lfe` and `fixest` documentation does not mention support for cross-classified clustered errors with FGLS.

Fixed-Effect Models

With cross-classified data, a set of cluster affiliation dummies for Level-2a and a set of cluster affiliation variables for Level-2b are both added to the model. Doing so will account for all variance attributable to either Level-2a or Level-2b such that only the within-cluster variance needs to be modeled. Relatedly, no effects of between-cluster predictors—either at Level-2a or Level-2b—can be directly estimated because they will be collinear with the cluster affiliation dummies.

Computation can sometimes be a challenge and it is not always straightforward to calculate cluster-specific intercepts because the reference cluster corresponds to a membership in specific Level-2a

[7] `PROC SURVEYREG` allows for multiple clusters to be declared but does not appear to treat the different levels as cross-classified. Instead, it creates a new cluster ID based on the permutations of the two cross-classified clusters (i.e., the intersection of Level-2a and Level-2b). The errors are then clustered by the newly created ID variable rather than the separately clustering by the two original cluster variables.

and Level-2b units. The `lfe` (Gaure, 2013) and `fixest` (Bergé, 2018) R packages can be helpful for implementing these models as can SAS `PROC GLM` if clustered errors are not desired in conjunction with the fixed effects (note that `ABSORB` cannot be used for cross-classified data in SAS).

Concluding Remarks

Clustered data can present some researchers with additional opportunities to clarify mechanisms of behavior, but to other researchers clustering can be an annoyance preventing appropriate inferences. Clustered data does not impact all researchers or analyses uniformly, so there is no one-size-fits-all statistical approach that ideally handles clustering in all cases. Multilevel models are a common option that allow researchers to leverage information in clustered data to disentangle individual and contextual contributions, although these opportunities come with assumptions and additional model building steps. Clustered errors and fixed-effect models both see clustering as a nuance to accommodate, although they handle clustering in different ways. Clustered errors focus on statistical issues whereas fixed-effect models are motivated by substantive issues. Importantly, these three approaches are not mutually exclusive. Ideas from different methods can be blended to tailor models to the unique needs and questions of an analysis or to combine the flexibility of multilevel models with the robustness of clustered errors or fixed-effects models. The take-home message is to make statements like "I need to use a multilevel model because my data are clustered" antiquated because, principally, the research question should inform the modeling approach. Clustered data do not require one specific statistical approach, instead, the statistical approach should be selected to best serve the needs of the research questions.

References

Allison, P. D. (2009). *Fixed effects regression models*. Sage.

Alonso, A., Litière, S., & Laenen, A. (2010). A note on the indeterminacy of the random-effects distribution in hierarchical models. *The American Statistician*, *64*(4), 318–324. https://doi.org/10.1198/tast.2010.09244

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods*, *24*(2), 443–483. https://doi.org/10.1177/1094428119877457

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086–1120. https://doi.org/10.1016/j.leaqua.2010.10.010

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434. https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x

Assari, S., Mardani, A., Maleki, M., Boyce, S., & Bazargan, M. (2021). Black-white achievement gap: Role of race, school urbanity, and parental education. *Pediatric Health, Medicine and Therapeutics*, *12*, 1–11. https://doi.org/10.2147/PHMT.S238877

Bailey, D. H., Duncan, G. J., Murnane, R. J., & Au Yeung, N. (2021). Achievement gaps in the wake of COVID-19. *Educational Researcher*, *50*(5), 266–275. https://doi.org/10.3102/0013189X211011237

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*(2), 151–164. https://doi.org/10.1037/a0030642

Balli, H. O., & Sørensen, B. E. (2013). Interaction effects in econometrics. *Empirical Economics*, *45*(1), 583–603. https://doi.org/10.1007/s00181-012-0604-2

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods*, *7*(2), 127–150. https://doi.org/10.1177/1094428104263672

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). *Parsimonious mixed models*. arXiv Preprint, arXiv: 1506.04967. http://arxiv.org/abs/1506.04967

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135–167. https://doi.org/10.3102/10769986028002135

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*(1), 97–114. https://doi.org/10.3102/1076998607310504

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*(4), 373–390. https://doi.org/10.1037/a0025813

Begg, M. D., & Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, *22*(16), 2591–2602. https://doi.org/10.1002/sim.1524

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*(2), 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

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*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–182.

Bergé, L. (2018). *Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm*. Department of Economics at the University of Luxembourg.

Berkowitz, R. (2021). School climate and the socioeconomic literacy achievement gap: Multilevel analysis of compensation, mediation, and moderation models. *Children and Youth Services Review*, *130*, Article 106238. https://doi.org/10.1016/j.childyouth.2021.106238

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275. https://doi.org/10.1162/003355304772839588

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

Brüderl, J., & Ludwig, V. (2015). Fixed-effects panel regression. In H. Best & C. Wolf (Eds.), *The Sage handbook of regression analysis and causal inference* (pp. 327–357). Sage.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, *8*(1), 158–233. https://doi.org/10.3102/0091732X008001158

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, *90*(3), 414–427. https://doi.org/10.1162/rest.90.3.414

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249. https://doi.org/10.1198/jbes.2010.07136

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Canning, E. A., Muenks, K., Green, D. J., & Murphy, M. C. (2019). STEM Faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Science Advances*, *5*(2), Article eaau4734. https://doi.org/10.1126/sciadv.aau4734

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, *18*(1), 5–46. https://doi.org/10.1016/0304-4076(82)90094-X

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, *3*(2), 399–408. https://doi.org/10.1017/psrm.2014.32

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, *58*(1), 615–637. https://doi.org/10.1146/annurev.psych.58.110405.085520

Cui, J., & Qian, G. (2007). Selection of working correlation structure and best model in GEE analyses of longitudinal data. *Communications in Statistics—Simulation and Computation*, *36*(5), 987–996. https://doi.org/10.1080/03610910701539617

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529–569. https://doi.org/10.1207/s15327906mbr3804_5

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*(1), 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

DeSouza, C. M., Legedza, A. T., & Sankoh, A. J. (2009). An overview of practical approaches for handling missing data in clinical trials. *Journal of Biopharmaceutical Statistics*, *19*(6), 1055–1073. https://doi.org/10.1080/10543400903242795

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*(10), Article e110257. https://doi.org/10.1371/journal.pone.0110257

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *43*(1), 49–73. https://doi.org/10.2307/2986113

Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press.

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *Review of Economics and Statistics*, *89*(2), 221–233. https://doi.org/10.1162/rest.89.2.221

Eager, C., & Roy, J. (2017). *Mixed effects models are sometimes terrible*. arXiv Preprint, arXiv:1701.04858.

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

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, *57*(4), 1198–1206. https://doi.org/10.1111/j.0006-341X.2001.01198.x

Fitzmaurice, G. M., Molenberghs, G., & Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society: Series B*, *57*(4), 691–704. https://doi.org/10.1111/j.2517-6161.1995.tb02056.x

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, *28*(2), 221–239. https://doi.org/10.1002/sim.3478

Gaure, S. (2013). Lfe: Linear group fixed effects. *The R Journal*, *5*(2), 104–117. https://doi.org/10.32614/RJ-2013-031

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. https://doi.org/10.1198/004017005000000661

Giesselmann, M., & Schmidt-Catran, A. W. (2022). Interactions in fixed effects regression models. *Sociological Methods & Research*, *51*(3), 1100–1127. https://doi.org/10.1177/0049124120914934

Goetgeluk, S., & Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, *64*(3), 772–780. https://doi.org/10.1111/j.1467-9868.2008.00673.x

Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, *60*(310), 539–547. https://doi.org/10.1080/01621459.1965.10480811

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, *22*(3), 364–375. https://doi.org/10.1177/0049124194022003005

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*(3), 393–419. https://doi.org/10.1080/00273171.2010.483390

Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology*, *7*(4), 121–133. https://doi.org/10.1027/1614-2241/a000030

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford.

Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent growth curve modeling. *Structural Equation Modeling*, *17*(3), 424–442. https://doi.org/10.1080/10705511.2010.489006

Gunasekara, F. I., Richardson, K., Carter, K., & Blakely, T. (2014). Fixed effects analysis of repeated measures data. *International Journal of Epidemiology*, *43*(1), 264–269. https://doi.org/10.1093/ije/dyt221

Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, *30*(22), 2696–2707. https://doi.org/10.1002/sim.4293

Hamaker, E. L., & Grasman, R. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, Article 1492. https://doi.org/10.3389/fpsyg.2014.01492

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics*, *140*(2), 670–694. https://doi.org/10.1016/j.jeconom.2006.07.011

Hanushek, E. A., Light, J. D., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2022). Long-run trends in the US SES—Achievement gap. *Education Finance and Policy*, *17*(4), 608–640. https://doi.org/10.1162/edfp_a_00383

Hashim, S. A., Kane, T. J., Kelley-Kemple, T., Laski, M. E., & Staiger, D. O. (2020). *Have income-based achievement gaps widened or narrowed? (No. w27714)*. National Bureau of Economic Research.

Hazlett, C., & Wainstein, L. (2022). Understanding, choosing, and unifying multilevel and fixed effect approaches. *Political Analysis*, *30*(1), 46–65. https://doi.org/10.1017/pan.2020.41

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, *15*(1), 1–26. https://doi.org/10.1214/ss/1009212671

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

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. https://doi.org/10.1002/sim.5338

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Henry, D. A., Betancur Cortés, L., & Votruba-Drzal, E. (2020). Black–White achievement gaps differ by family socioeconomic status from early childhood through early adolescence. *Journal of Educational Psychology*, *112*(8), 1471–1489. https://doi.org/10.1037/edu0000439

Hin, L. Y., & Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, *28*(4), 642–658. https://doi.org/10.1002/sim.3489

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, *42*(4), 609–629. https://doi.org/10.1080/00273170701710072

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Hoffman, L., & Walters, R. W. (2022). Catching up on multilevel modeling. *Annual Review of Psychology*, *73*, 659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Howard, T. C. (2019). *Why race and culture matter in schools: Closing the achievement gap in America's classrooms*. Teachers College Press.

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer.

Hox, J. J., & McNeish, D. (2020). Small samples in multilevel modeling. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 215–225). Routledge.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, *147*(7), 694–703. https://doi.org/10.1093/oxfordjournals.aje.a009511

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. *The Journal of Experimental Education*, *84*(1), 175–196. https://doi.org/10.1080/00220973.2014.952397

Huang, F. L. (2018). Multilevel modeling and ordinary least squares regression: How comparable are they? *The Journal of Experimental Education*, *86*(2), 265–281. https://doi.org/10.1080/00220973.2016.1277339

Huang, F. L. (2022). Analyzing cross-sectionally clustered data using generalized estimating equations. *Journal of Educational and Behavioral Statistics*, *47*(1), 101–125. https://doi.org/10.3102/10769986211017480

Huang, F. L., & Li, X. (2022). Using cluster-robust standard errors when analyzing group-randomized trials with few clusters. *Behavior Research Methods*, *54*, 1181–1199. https://doi.org/10.3758/s13428-021-01627-0

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*(4), 467–474. https://doi.org/10.1097/EDE.0b013e3181caeb90

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, *51*(10), 5142–5154. https://doi.org/10.1016/j.csda.2006.05.021

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*(4), 805–820. https://doi.org/10.2307/2530695

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396. https://doi.org/10.1198/016214501753382309

Kaufman, J. S. (2008). Commentary: Why are we biased against bias? *International Journal of Epidemiology*, *37*(3), 624–626. https://doi.org/10.1093/ije/dyn035

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583–2595. https://doi.org/10.1016/j.csda.2008.12.013

Kim, J. S., & Frees, E. W. (2006). Omitted variables in multilevel models. *Psychometrika*, *71*(4), 659–690. https://doi.org/10.1007/s11336-005-1283-0

Kim, J. S., & Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika*, *72*(4), 505–533. https://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). Routledge.

Kish, L. (1965). *Survey sampling*. Wiley.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, *11*(1), 55–77.

Kvålseth, T. O. (1985). Cautionary note about R2. *The American Statistician*, *39*(4), 279–285. https://doi.org/10.1080/00031305.1985.10479448

Kwok, O. M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, *42*(3), 557–592. https://doi.org/10.1080/00273170701540537

Lai, M. H., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *The Journal of Experimental Education*, *83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Lang, J. W., Bliese, P. D., & Adler, A. B. (2019). Opening the black box: A multilevel framework for studying group processes. *Advances in Methods and Practices in Psychological Science*, *2*(3), 271–287. https://doi.org/10.1177/2515245918823722

Lang, J. W., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence in groups using longitudinal multilevel methods. *Personnel Psychology*, *71*(2), 255–281. https://doi.org/10.1111/peps.12260

Lee, Y. R., & Pustejovsky, J. E. (2023). Comparing random effects models, ordinary least squares, or fixed effects with cluster robust standard errors for cross-classified data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000538

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M., & Ibrahim, J. G. (2004). Protective estimator for linear regression with nonignorably missing Gaussian outcomes. *Statistical Modelling*, *4*(1), 3–17. https://doi.org/10.1191/1471082X04st066oa

Litière, S., Alonso, A., & Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, *63*(4), 1038–1044. https://doi.org/10.1111/j.1541-0420.2007.00782.x

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models*. SAS Institute.

Maas, C. J., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*(3), 427–440. https://doi.org/10.1016/j.csda.2003.08.006

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, *57*(1), 126–134. https://doi.org/10.1111/j.0006-341X.2001.00126.x

McCaffrey, D. F., Lockwood, J. R., Mihaly, K., & Sass, T. R. (2012). A review of Stata commands for fixed-effects estimation in normal linear models. *The Stata Journal*, *12*(3), 406–432. https://doi.org/10.1177/1536867X1201200305

McCulloch, C. E., & Neuhaus, J. M. (2011a). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, *67*(1), 270–279. https://doi.org/10.1111/j.1541-0420.2010.01435.x

McCulloch, C. E., & Neuhaus, J. M. (2011b). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, *26*(3), 388–402. https://doi.org/10.1214/11-STS361

McNeish, D. (2019). Effect partitioning in cross-sectionally clustered data without multilevel models. *Multivariate Behavioral Research*, *54*(6), 906–925. https://doi.org/10.1080/00273171.2019.1602504

McNeish, D. (2021). Specifying location-scale models for heterogeneous variances as multilevel SEMs. *Organizational Research Methods*, *24*(3), 630–653. https://doi.org/10.1177/1094428120913083

McNeish, D., & Bauer, D. J. (2022). Reducing incidence of nonpositive definite covariance matrices in mixed-effect models. *Multivariate Behavioral Research*, *57*(2–3), 318–340. https://doi.org/10.1080/00273171.2020.1830019

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*(1), 20–35. https://doi.org/10.1037/met0000182

McNeish, D., & Matta, T. (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods*, *50*(4), 1398–1414. https://doi.org/10.3758/s13428-017-0976-5

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. https://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*(2), 200–215. https://doi.org/10.1080/00273171.2016.1262236

Merolla, D. M., & Jackson, O. (2019). Structural racism as the fundamental cause of the academic achievement gap. *Sociology Compass*, *13*(6), Article e12696. https://doi.org/10.1111/soc4.12696

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*(1), 129–149. https://doi.org/10.1207/s15327906mbr3901_5

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

Moody, C. E., & Marvell, T. B. (2020). Clustering and standard error bias in fixed effects panel data regressions. *Journal of Quantitative Criminology*, *36*(2), 347–369. https://doi.org/10.1007/s10940-018-9383-z

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69–85. https://doi.org/10.2307/1913646

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. https://doi.org/10.2307/271070

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, *54*(2), 638–645. https://doi.org/10.2307/3109770

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, *92*(440), 1320–1329. https://doi.org/10.1080/01621459.1997.10473653

Pan, W., & Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, *12*(2), 475–490.

Pepper, J. V. (2002). Robust inferences from random clustered samples: An application using data from the panel study of income dynamics. *Economics Letters*, *75*(3), 341–345. https://doi.org/10.1016/S0165-1765(02)00010-1

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, *49*(6), 518–543. https://doi.org/10.1080/00273171.2014.933762

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*(2), 189–205. https://doi.org/10.1037/met0000052

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

Preisser, J. S., Lohman, K. K., & Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, *21*(20), 3035–3054. https://doi.org/10.1002/sim.1241

Primo, D. M., Jacobsmeier, M. L., & Milyo, J. (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly*, *7*(4), 446–459. https://doi.org/10.1177/153244000700700405

Pustejovsky, J. E. (2020). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (Version 0.4.2) [Computer software]. https://cran.r-project.org/web/packages/clubSandwich/clubSandwich.pdf

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, *36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, *19*(4), 337–350. https://doi.org/10.3102/10769986019004337

Rast, P., & Ferrer, E. (2018). A mixed-effects location scale model for dyadic interactions. *Multivariate Behavioral Research*, *53*(5), 756–775. https://doi.org/10.1080/00273171.2018.1477577

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

Rights, J. D., & Sterba, S. K. (2020). New recommendations on the use of R-squared differences in multilevel model comparisons. *Multivariate Behavioral Research*, *55*(4), 568–599. https://doi.org/10.1080/00273171.2019.1660605

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Sanders, E. A., & Konold, T. R. (2023). X matters too: How the blended slope problem manifests differently in unilevel vs. multilevel models. *Methodology*, *19*(1), 1–23. https://doi.org/10.5964/meth.9925

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal: Promoting Communications on Statistics and Stata*, *13*(1), 65–76. https://doi.org/10.1177/1536867X1301300105

Shah, B. V., Holt, M. M., & Folsom, R. E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, *47*(3), 43–57.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*(4), 323–355. https://doi.org/10.3102/10769986023004323

Snijders, T. A., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*(3), 748–761. https://doi.org/10.1111/ajps.12001

Theobald, E. j., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., II, Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., … Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, *117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

Tofighi, D., & Kelley, K. (2016). Assessing omitted confounder bias in multilevel mediation models. *Multivariate Behavioral Research*, *51*(1), 86–105. https://doi.org/10.1080/00273171.2015.1105736

Tranmer, M., & Steel, D. G. (2001). Ignoring a level in a multilevel model: Evidence from UK census data. *Environment and Planning A: Economy and Space*, *33*(5), 941–948. https://doi.org/10.1068/a3317

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology*, *55*(4), 329–337. https://doi.org/10.1016/S0895-4356(01)00476-0

Twisk, J. W. (2003). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology*, *19*(8), 769–776. https://doi.org/10.1023/B:EJEP.0000036572.00663.f2

Van den Noortgate , W., Opdenakker, W., & & Onghena, M. C. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*(3), 281–303. https://doi.org/10.1080/09243450500114850

Van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*(1), Article 25216. https://doi.org/10.3402/ejpt.v6.25216

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, *23*(4), 541–556. https://doi.org/10.1016/S0167-9473(96)00047-3

Viechtbauer, W., & López-López, J. A. (2022). Location-scale models for meta-analysis. *Research Synthesis Methods*, *13*(6), 697–715. https://doi.org/10.1002/jrsm.1562

Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*(1), 63–83. https://doi.org/10.1037/met0000030

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2021). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*, *26*(1), 74–89. https://doi.org/10.1037/met0000270

Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods*, *51*(5), 1968–1986. https://doi.org/10.3758/s13428-019-01255-9

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics—Simulation and Computation*, *22*(4), 1079–1106. https://doi.org/10.1080/03610919308813143

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121–130. https://doi.org/10.2307/2531248

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*(4), 1049–1060. https://doi.org/10.2307/2531734

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, *57*(3), 795–802. https://doi.org/10.1111/j.0006-341X.2001.00795.x

Ziegler, A. & Vens, M. (2014). Generalized estimating equations. In W. Ahrens & I. Pigeot (Eds.), *Handbook of epidemiology* (pp. 1337–1376). Springer-Verlag.

Received November 18, 2022
Revision received August 30, 2023
Accepted September 28, 2023 ■