Psychological Methods
2017, Vol. 22, No. 1, 114–140

© 2016 American Psychological Association
1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000078

On the Unnecessary Ubiquity of Hierarchical Linear Modeling

Daniel McNeish
University of Maryland, College Park and Utrecht University

Laura M. Stapleton and Rebecca D. Silverman
University of Maryland, College Park

In psychology and the behavioral sciences generally, the use of the hierarchical linear model (HLM) and its extensions for discrete outcomes are popular methods for modeling clustered data. HLM and its discrete outcome extensions, however, are certainly not the only methods available to model clustered data. Although other methods exist and are widely implemented in other disciplines, it seems that psychologists have yet to consider these methods in substantive studies. This article compares and contrasts HLM with alternative methods including generalized estimating equations and cluster-robust standard errors. These alternative methods do not model random effects and thus make a smaller number of assumptions and are interpreted identically to single-level methods with the benefit that estimates are adjusted to reflect clustering of observations. Situations where these alternative methods may be advantageous are discussed including research questions where random effects are and are not required, when random effects can change the interpretation of regression coefficients, challenges of modeling with random effects with discrete outcomes, and examples of published psychology articles that use HLM that may have benefitted from using alternative methods. Illustrative examples are provided and discussed to demonstrate the advantages of the alternative methods and also when HLM would be the preferred method.

Keywords: HLM, GEE, cluster robust errors, clustered data, multilevel model

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The hierarchical linear model (HLM) and its discrete outcome extensions are useful and popular methods for analyzing data that have a clustered structure, a common occurrence in psychological research (for brevity, HLM will be used to encompass all outcomes distributions for such models including HLM, hierarchical generalized linear models [HGLM], etc.). As a testament to the widespread use of HLM, in Footnote 1 in Bauer and Sterba (2011), in a search of the psychological literature from 2006 to 2011, 211 studies were located using HLM's many aliases such as "multilevel model," "random coefficient model," "mixed model," or "hierarchical linear model." On the other hand, only 14 studies were found in the same timeframe using keywords related to generalized estimating equations (GEE), another method used to account for clustered data that is more popularly utilized in biology, epidemiology, and medicine. Although this search did not play a large role in their study (as evidenced by a majority of the information appearing in a footnote), it does clearly illustrate the overwhelming preference for HLM in psychology compared with generalized estimating equations by a wide margin of about 15 to 1.

This article was published Online First May 5, 2016.

Daniel McNeish, Department of Human Development and Quantitative Methodology, University of Maryland, College Park, and Department of Methodology and Statistics, Utrecht University; Laura M. Stapleton, Department of Human Development and Quantitative Methodology, University of Maryland, College Park; Rebecca D. Silverman, Department of Counseling, Higher Education, and Special Education, University of Maryland, College Park.

Correspondence concerning this article should be addressed to Daniel McNeish, P.O. Box 80140, 3508 TC Utrecht, the Netherlands. E-mail: d.m.n.mcneish@uu.nl

Goal and Outline for This Article

Although the distinction between HLM and other methods was not the intended focus of Bauer and Sterba (2011), their study did quantify the overwhelming tendency toward HLM with clustered data even though it seems unlikely that cluster-specific inferences obtained by HLM are desired or are of interest in 94% of psychological studies (details on cluster-specific inferences will be discussed in more detail in subsequent sections). This article's primary aim is to raise awareness among psychologists that having clustered data does not necessitate the use of HLM. Although HLM is quite useful in many situations and lines of research, when the clustering of the data is more a nuisance to accommodate rather than a substantive interest, alternative population-averaged methods (PAMs; a class of methods that account for clustering without explicitly splitting the model into multiple levels) may be a viable option to address such scenarios in a simplified manner.

Although the popularity of HLM and the benefit of making inferences at the cluster level or modeling interactions across levels cannot be denied, we provide some evidence through a review of graduate course syllabi that researchers in psychology and related sciences may employ HLM with such frequency not because of the inherent advantages but rather because alternative models for clustered data such as cluster-robust standard errors (CR-*SE*s; a.k.a. sandwich or empirical estimators), variance estimation methods such as Taylor series linearization and replication, or GEE are not fully considered or even known among psychologists while the tradition of using HLM is firmly cemented. Thus, a goal of this article is to raise awareness among instructors of clustered data analysis courses that PAMs could prove useful for many researchers but are largely overlooked in research methods and statistics courses aimed toward psychology students.

Although we intend to provide a fair comparison between HLM and alternative methods, we admittedly focus our presentation on alternatives to HLM because these methods are rarely applied in the psychological literature. There are many extant resources available that effectively introduce readers to these alternative methods (e.g., Ballinger, 2004; Burton, Gurrin, & Sly, 1998; Gardiner, Luo, & Roman, 2009; Ghisletta & Spini, 2004; Hanley, Negassa, Edwardes, & Forrester, 2003; Hubbard et al., 2010; Twisk, 2004; Zorn, 2001), but, as argued throughout this article, they are not targeted toward psychologists, and psychologists may not be aware of their utility to psychological research. Although this article will provide some didactic detail, the primary motivation is to highlight where the use of HLM may unnecessarily overcomplicate analyses and where alternative methods may be better suited to researchers' interests. As such, because there is a wealth of resources in the psychological literature that discuss the merits of HLM, this article primarily focuses on the advantages that can be realized with alternative methods.

To outline the structure of this article, we first introduce the foundational details of HLM and alternative clustered data methods and compare these methods with a particular focus on what these alternative methods can provide for psychological researchers. Assumptions of each method are compared and contrasted as are advantages and disadvantages of the use of each method. We conclude with three illustrative examples that compare the use of HLM and alternative methods for a model with a continuous outcome and a model with a binary outcome. Broader implications for the field are then discussed.

Syllabus Review

As evidence for the claim that psychologists and researchers in related fields tend almost exclusively toward HLM, a convenience sample of 67 syllabi ranging from 2008 to 2015 that listed readings and a tentative outline of topics from graduate school courses in the United States and Canada on modeling clustered, multilevel, and/or longitudinal data were reviewed. Syllabi were found through a simple Google search for "multilevel model," "longitudinal data analysis," "correlated data," "clustered data," "nested data," "complex survey data," "survey data analysis," "cluster robust errors," generalized estimating equations," and "hierarchical linear model" with "syllabus" placed after the course keyword. As a result, only syllabi available publically were included. To be in the sample, courses must have (a) focused primarily on clustered data methods (e.g., economics syllabi were notably absent because clustered data methods are commonly interspersed within broader econometrics course sequences); and (b) have focused primarily on regression-based methods.

Syllabi were found from eight different academic disciplines from both the social sciences and the natural/health sciences.[1] The most widely represented disciplines were statistics (24%), biostatistics (19%), education (15%), and psychology (15%). Table 1 shows the number of syllabi found from each respective discipline that included HLM, PAMs, or both. Sixty-two of the 67 syllabi (93%) included HLM (or one of its aliases such as mixed models, random effects models, etc.) as a topic to be covered either in lecture or through readings. However, only 64% (43/67) included any coverage of at least one PAM.[2] Moreover, large differences were observed between the social sciences (criminology, education, political science, psychology, and sociology) compared with the natural and health sciences (biostatistics, public health, and statistics). Only 24% (7/29) of the social science discipline syllabi included PAMs as a topic to be covered in lecture or readings while 92% (35/38) of natural and health science discipline syllabi did so. Of particular note is that only one psychology syllabi mentioned PAMs (in a course titled "Categorical Data Analysis") while at least 85% of syllabi in every natural or health science discipline included PAMs. Although economics syllabi were not included, the results would be likely be even more stark because economics follows an opposing trend to psychology where HLM is almost *never* used. As support, Petersen (2009) included a review of methods to account for clustering in economics and he found that fewer than 3% of studies included in his survey used HLM (p. 464).

Table 1
Frequency of Syllabi Listing PAMs and HLM by Department Affiliation

Discipline	PAMs Only	HLM Only	HLM and PAMs	Total
Biostatistics	0	1	12	13
Criminology	0	1	0	1
Education	0	8	2	10
Political science	1	2	1	4
Psychology	0	9	1	10
Public health	1	0	8	9
Sociology	1	1	2	4
Statistics	2	2	12	16

Note. PAM = population-averaged method; HLM = hierarchical linear model.

Although this sample was convenient and not random or exhaustive and broad inference is not entirely warranted, it provides some evidence that, assuming what is being taught in graduate courses is indicative of what is applied in practice—in social sciences, students do not appear to be exposed to a variety of clustered data methods. Consequently, the 15 to 1 MLM-to-GEE ratio found by Bauer and Sterba (2011) in psychological studies is not surprising.

Overview of Methods

We will next provide an overview of three methods that are useful when modeling clustered data: HLM, cluster-robust stan-

[1] Some courses were cross-listed in multiple departments but syllabi were ultimately classified within only a single discipline. In instances of cross-listed departments, the department affiliation of the instructor was used to uniquely classify syllabi in such cases. If the department was ambiguous (e.g., educational psychology could be education or psychology), the college was used for classification (e.g., College of Education vs. College of Arts and Sciences). In one instance, the instructor of a cross-listed course held appointments in multiple departments, so the specialization of her Ph.D. was used.

[2] We want to note that in 12 out of the 67 syllabi (18%), the course title indicated that HLM was the specific interest of the course (e.g., courses with titles like "Hierarchical Linear Models" or "Introduction to Multilevel Models"). However, the remaining syllabi had broader titles that did not narrow the scope with titles such as "Analysis of Correlated Data," "Advanced Regression Analysis," or "Longitudinal Data Analysis."

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

dard errors, and generalized estimating equations. To present a more targeted narrative, the following discussion will primarily focus on cross-sectional clustering (e.g., people within families) although readers should keep in mind that the same principles apply to longitudinal clustering (e.g., repeated measures clustered within individuals). The following discussion will not present extensive statistical detail: interested readers can find additional mathematical information in Appendix A.

HLM

Conceptual overview. HLM accounts for the clustered nature of data by directly modeling the clustering with random coefficients (Laird & Ware, 1982; Stiratelli, Laird, & Ware, 1984). Regression coefficients in HLM consist of two possible types of effects: a fixed effect and a random effect. Fixed effects are estimated to represent the relationship between a predictor and the outcome irrespective of which cluster observations belong to (assuming the predictor is not cluster-mean centered), similar to a standard single-level regression model (Raudenbush & Bryk, 2002). For each cluster, a cluster-specific random effect may be estimated (but is not required). Random effects capture how much the relation between the predictor and the outcome differs from the fixed effect estimate.

Mathematically, HLM for continuous outcomes can be expressed as

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

where $\mathbf{Y}_j$ is an $m_j \times 1$ vector of responses for cluster j where m_j is the number of units within cluster j, $\mathbf{X}_j$ is an $m_j \times p$ design matrix for the predictors in cluster j (at either level in this notation) where p is the number of predictors (which includes the intercept), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $\mathbf{Z}_j$ is an $m_j \times q$ design matrix for the random effects of cluster j, $\mathbf{u}_j$ is a $q \times 1$ vector of random effects for cluster j where q is the number of random effects and $p \geq q$, $E(\mathbf{u}_j) = \mathbf{0}$ and $Cov(\mathbf{u}_j) = \mathbf{G}$ where $\mathbf{G}$ is $q \times q$, and $\boldsymbol{\varepsilon}_j$ is an $m_j \times 1$ vector of residuals of the observations in cluster j where $E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$, $Cov(\boldsymbol{\varepsilon}_j) = \mathbf{R}_j$. HLM in frequentist settings is typically estimated with either maximum likelihood or restricted maximum likelihood whose details are beyond the scope of this introduction (see Raudenbush & Bryk, 2002 for more details). Variance of regression coefficients in HLM are calculated by

$$Var^{HLM}(\hat{\boldsymbol{\beta}}) = \left\{\sum_{j=1}^{J}\left(\mathbf{X}_j^{\mathrm{T}}\hat{\mathbf{V}}_j^{-1}\mathbf{X}_j\right)\right\}^{-1} \tag{2}$$

where $\mathbf{V}_j = Var(\mathbf{Y}_j) = \mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\mathrm{T}} + \mathbf{R}_j$ and standard errors are obtained by taking the square root of the diagonal elements of $Var^{HLM}(\hat{\boldsymbol{\beta}})$.

As a conceptual example of HLM, consider a model for depression inventory scores taken from many clinics that contain an overall intercept (a fixed effect) for all clinics. However, the sample may contain some clinics with relatively few symptoms (on average) and also some clinics with more severe symptoms (on average) for which the intercept fixed effect may not be entirely representative. So, a random effect for the intercept may be included to more accurately reflect that depression scores are partially dependent upon the clinic that patients visit. The variance of the outcome is then partitioned into two-parts (or more if the data have more levels to the hierarchy): within-cluster variance (**R**) and between-cluster variance (**G**). The between-cluster variance captures the dispersion of the random effects from cluster to cluster—if the between cluster variance is high, then knowing to which cluster an observation belongs will be more informative for modeling an individual's score. The between-cluster variance, which is not explicitly modeled in single-level models, helps to obtain more appropriate standard error estimates for regression coefficient standard errors to account for the violation of the independence assumption made by single-level models. Within-cluster variance is interpreted similarly to error variance in single-level models and is largely a measure of the accuracy of predictions from the model. However, it is important to note that single-level models do not partition the variance between levels and modeling clustered data with a model that does not partition the variance will result in an error variance term that combines variance from all levels.

Assumptions, properties, and advantages of HLM. When modeling clustered data with HLM, 10 assumptions are made:

1. All relevant predictors are included in the model.
2. All relevant random effects are included in the model.
3. The covariance structure of the within-cluster residuals, **R**, is properly specified (when the outcome is continuous).
4. The covariance structure of the random effects, **G**, is properly specified (for all outcomes scales).
5. The within-cluster residuals and the random effects do not covary [$Cov(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$].
6. The within-cluster residuals follow a multivariate normal distribution (when the outcome is continuous).
7. The random effects follow a multivariate normal distribution (for all outcome scales).
8. The predictor variables do not covary with the residuals/random effects at any other level [$Cov(\mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $Cov(\mathbf{X}, \mathbf{u}) = \mathbf{0}$].
9. Sample size is sufficiently large for asymptotic inference at each level.
10. With or without preprocessing, missing data are assumed to be missing completely at random (MCAR) or missing at random (MAR).

When modeling with HLM, the researcher must explicitly specify which Level-1 slopes and/or the intercept vary randomly in the population and the covariance structure of the random effects. For instance, the researcher must decide whether the random effects for the slope covary with random effects for the intercept or if the random effects for one predictor may covary with the random effects for another predictor. Referring back to the depression example, if "hours of therapy" were a predictor of depression symptoms, with HLM, researchers would need to decide whether the effect of hours of therapy differs for each cluster and, if so,

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

whether the cluster-specific effects for number of hours of therapy covary with the cluster-specific intercepts. Similarly, for continuous outcomes, the structure of the within-cluster residuals must also be specified. That is, covariance matrix of the within-cluster residuals has to be selected by the researcher although it is often assumed to have equal values on the diagonal and zeroes on the off diagonal (i.e., an independence structure).

The inclusion of random effects in HLM makes it distinct from the other methods that will be discussed in this article and, as a result, Assumptions 2, 4, 5, 6, 7, and 8 are unique to HLM. To varying degrees, these assumptions are important because standard error estimates of regression coefficients likely will be biased otherwise. Efficiency could also be decreased (standard errors are larger than they need to be) which adversely affects power (Agresti, Caffo, & Ohman-Strickland, 2000; Ferron, Dailey, & Yi, 2002; LeBeau, 2013).

However, violations of Assumptions 2, 4, and/or 6 do not always have a large impact on model estimates. Verbeke and Lesaffre (1997) showed that assuming normality of the random effects (Assumption 7) even when the distribution is non-normal still yielded consistent regression coefficient estimates so long as all variables have fourth moments. Standard error estimates were problematic when random effects were non-normal with small or moderate samples (fewer than 120). Jacqmin-Gadda, Sibillot, Proust, Molina, and Thiébaut (2006), Litière, Alonso, and Molenberghs (2007), and Agresti, Caffo, and Ohman-Strickland (2004) have shown that misspecifying the structure of either the random effects covariance or the distribution of the errors has a large effect on standard error estimates throughout the model and, consequently, on Type-I error rates and power. Assumptions 6 and 7 can also be addressed directly in the model by specifying a distribution other than normal (see, e.g., Lin & Lee, 2008; Liu & Yu, 2008; Muthén & Asparouhov, 2008) although implementation of these methods often requires a fairly high level of programmatic skill.

Misspecifying the number of random effects and/or their covariance structure can also lead to biased point estimates when the outcome variable is discrete (Litière et al., 2007). Thus, when choosing to model clustered data with HLM, researchers with continuous outcomes and large sample sizes can be fairly confident that their results are robust to a misspecified covariance matrix or the exclusion of a random effect. However, with continuous outcomes with small or moderate number of clusters or with discrete outcomes, a violation of either assumption can adversely affect inference from model estimates. As such, with the type of data frequently seen in applied psychological research, the common strategy of solely including a random intercept into the model without considering additional random effects may not adequately account for clustering despite common perception to the contrary.

Estimation with discrete outcomes is one noted difficulty in the use of HLM due to the inclusion of the random effects. In models for continuous outcomes, the random effects can be integrated out of the likelihood meaning that the likelihood function is averaging over the random effects distribution. The result is that estimation of the model is not much more complex than other models estimated with likelihood techniques. However, with discrete outcomes, there is no closed form solution for integrating the random effects out of the likelihood function (Fitzmaurice, Laird, & Ware, 2012; McCalloch & Searle, 2001). Therefore, the likelihood must be approximated with numerical integration as with Adaptive Gaussian Quadrature (AGQ) or a Laplace Approximation or linearized as with penalized quasi-likelihood (PQL; Breslow & Clayton, 1993). These estimation methods for HGLM have more limitations including vast computational overhead with multiple random effects or biased estimates even when the number of clusters exceeds 100 (e.g., Diaz, 2007; Pinheiro & Bates, 1995).

Very broadly, AGQ breaks up the likelihood into several small components, evaluates each component, and then takes a weighted sum of all components to approximate the integral. The number of partitions is determined by $Q + 1$ where Q is the number of quadrature points (which are user selected). As more quadrature points are selected, the approximation becomes more accurate because the likelihood surface is partitioned into smaller and smaller pieces. The trade-off associated with many quadrature points, however, is increased computational burden, so selecting the appropriate number of points often requires balancing accuracy with computational demand (Fitzmaurice, Laird, & Ware, 2012; Givens & Hoeting, 2005). The number of computations per iteration of the maximum likelihood algorithm is equal to JQ^q where J is the number of clusters and q is the number of random effects: the computation grows exponentially as the number of random effects increases and model can take several hours to converge in such cases (Kim, Choi, & Emery, 2013). The Laplace Approximation, attempts to approximate AGQ with Taylor series expansions, which reduces much of the computation burden (equivalent to using a single quadrature point). Previous studies have noted that its performance is hampered by a small number of clusters and/or small cluster sizes (Clarkson & Zhan, 2002; Diaz, 2007; Joe, 2008; Kim et al., 2013; Raudenbush, Yang, & Yosef, 2000). As an alternative, PQL approximates the model rather than the likelihood function. That is, HLM with discrete outcomes require a nonlinear link function to relate the mean of the outcome distribution to a linear predictor—PQL attempts to linearly approximate the model and then apply estimation methods that are suitable for linear models (where the random effects are able to be integrated out of the likelihood function). Although appealing for its time-efficient estimation, the approximation tends to be worse than AGQ and model estimates have been found to have substantial bias under nonideal conditions (Diaz, 2007; Pinheiro & Bates, 1995; Zhou, Perkins, & Hui, 1999).

Methods for Cluster Sampled Data

The main focus of methods for single-level analyses of cluster sampled data (e.g., cluster-robust standard errors, Taylor series linearization, balanced repeated replication, jackknife replication) is to estimate the standard errors of the regression coefficients in a way that accurately reflects the process by which data were collected. This is done by considering features of complex sampling designs such as clusters, strata, and/or sampling weights. Because of the relative scarcity of explicitly cluster sampled primary data in psychology compared to naturally clustered data (which do not have strata or weighting information), we will not discuss the complex survey elements or methods to accommodate them in more detail. For an accessible treatment of these methods, interested readers are referred to Heeringa, West, and Berglund (2010).

Regardless of whether observations are naturally clustered or explicitly cluster sampled, methods for cluster sampled data can be

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

applied to obtain standard errors that reflect the nature of the clustering. Many clustered data methods can be implemented in general software packages such as SPSS (complex samples), Stata (svyset and svy commands with cluster option), SAS (Proc SurveyReg and Proc SurveyLogistic), M*plus* (TYPE = COMPLEX), or a variety of R packages (sandwich, plm, and clusterSEs) while additional software programs are dedicated specifically to clustered data methods such as SUDAAN, WesVar, and IVEware (Heeringa, West, & Berglund, 2010).

Cluster robust-standard error conceptual overview. Cluster-robust standard errors (CR-*SE*s; also referred to as empirical standard errors or the sandwich estimator) are methods for estimating standard errors of fixed effects for a variety of models, including HLM (i.e., CR-*SE* and HLM are not mutually exclusive; Kovacevic & Rai, 2003; Pfefferman, Skinner, Holmes, Goldstein, & Rasbash, 1998; Rabe-Hesketh & Skrondal, 2006; Rao, Verret, & Hidiroglou, 2013). In fact, CR-*SE* are a germane step in the GEE algorithm discussed in the next section. In this article, we will refer to CR-*SE* in the context of estimating a single-level general or generalized linear model whose standard errors are then estimated CR-*SE*s but readers should note that CR-*SE* are much more widely applicable.

Although the specific derivational details are provided in Appendix A, as a basic overview for the notation for CR-*SE*s, consider a standard single-level regression formulated by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

for $\mathbf{Y}$ an $n \times 1$ vector of outcomes, $\mathbf{X}$ an $n \times p$ design matrix, $\boldsymbol{\beta}$ a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ an $n \times 1$ vector of residuals assumed to be distributed $N^{i.i.d.}(0, \sigma^2)$ for n the total sample size, p the number of predictors, and σ^2 the estimate of the residual variance. Under ordinary least squares (OLS), the regression coefficients have a closed from solution such that

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{4}$$

with the variance of the regression coefficients calculated by

$$Var^{OLS}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \tag{5}$$

if the assumptions are upheld. CR-*SE*s alter the calculation of the regression coefficient variance to accommodate assumption violations (i.e., homogeneity of variance and independence) such that

$$Var^{\mathrm{CR}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sum_{j=1}^{J}(\mathbf{X}_j^{\mathrm{T}}\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^{\mathrm{T}}\mathbf{X}_j)(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \tag{6}$$

where standard errors are obtained by taking the square root of the diagonal elements of $Var^{\mathrm{CR}}(\hat{\boldsymbol{\beta}})$.

Conceptually, with CR-*SE*s, first a standard single-level model (e.g., with OLS or maximum likelihood as shown in Equation 3) is used to estimate the regression coefficients (e.g., $\hat{\boldsymbol{\beta}}_{OLS}$). In the presence of clustering, bias in the standard error estimates are the primary concern, so the regression coefficients with CR-*SE*s will be identical to what would be obtained if the clustering was completely ignored. More plainly, compared with a single-level model, only the standard errors (and any quantities that require them such as *t* statistics) will be different (notice the more complex specification of Equation 6 compared with Equation 5). The standard errors from Equation 5 will be underestimated to the extent that the clustering is informative. CR-*SE* address this with a statistical correction based on the residuals (using Taylor series linearization) to yield standard error estimates that more accurately reflect the variability in the regression coefficient estimates given that clustering is present in the data. In essence, this results in residuals being summed by clusters (where clusters are assumed to be independent) rather than by individual (which are known to be dependent on cluster, noted by the middle parenthetical term in Equation 6 taking a *j* subscript). In most cases, this process will inflate the standard error estimates although it may possibly deflate estimates if the ICC is negative.[3] Note that no random effects are included in the model to explicitly model variability across clusters—only fixed effect regression coefficient estimates are obtained from the model with standard error estimates that account for clustering.

Assumptions, properties, and advantages of cluster-robust standard errors. When estimating regression coefficient standard errors with CR-*SE*s, three assumptions must be upheld:

1. All relevant predictors are included in the model.
2. Observations between clusters are not related (there is not a higher level of the hierarchy).
3. The sample size is sufficiently large for asymptotic inferences at the cluster level.

Because CR-*SE*s can be implemented with a variety of models, other assumptions unrelated to clustering must be upheld for valid inference (e.g., the independence assumption of the general linear model would not be required and heteroskedasticity due to clustering would be permissible as CR-*SE* would address these issues).

Similar to HLM, the CR-*SE* estimates fully address the clustered nature of the data. Unlike HLM, the model does not provide random effect estimates and model output appears and is interpreted *identically* to a single-level model. Although this may not intuitively seem highly advantageous, it does afford CR-*SE*s a few distinct advantages over HLM and also GEEs (which are discussed in the next section). Because there are no modeled random effects and the variance is not partitioned between different levels, for continuous outcomes, the expected mean square from a model with CR-*SE* is identical to a single-level model. That is, the statistical adjustment for clustering only affects the standard error estimates of the regression coefficients, leaving the regression coefficient estimates unaffected compared to a single-level model. This means that CR-*SE*s can output model R^2 and effect size measures that are identical to what would be obtained through OLS because quantities used in these calculations (sum of squares, expected mean squares) are unaffected by the statistical correction to the standard error estimates and the computational formulas are equivalent to a single-level model (Hayes & Cai, 2007).[4] This makes CR-*SE*s an

[3] A negative ICC may be present, for instance, if a school district intentionally tries to balance the demographic makeup of different classrooms in which case students are less like others in their own classroom than those in different classrooms. We thank one of the anonymous reviewers for providing the authors with this example.

[4] There is some debate whether R^2 values are equally interpretable when the residuals are no longer independently and identically distributed and some software programs suppress R^2 values for such models. However, Wooldridge (2003) states that "R^2 (is a) consistent estimator of the population R-squared whether or not the homoskedasticity assumption holds" (p. 265).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

attractive option for the multiple moderated regression strategy popular in psychology when the data happen to be clustered because the familiar regression metrics and the change in R^2 between blocks can still be calculated. Additionally, intended single-level models that have unforeseen clustering or partial clustering can also be easily accommodated.

Again returning to the depression example, imagine a researcher was interested in whether sex moderates the effect of hours of therapy on depression symptoms. In a standard multiple moderated regression, one would record R^2 values first for a model with just the hours of therapy variable (and also with possible relevant covariates), then both hours of therapy and sex (along with relevant covariates), and finally a model with hours of therapy, sex, and their interaction (along with relevant covariates). ΔR^2 is often the primary interest of this common type of analysis in psychology; however, if data are clustered such that patients are nested within clinics, HLM does not yield an analogous R^2 value. CR-*SE* could be implemented without any change to the standard procedure (with the added benefit that the standard errors would account for clustering).

HLM analogues of R^2 and effect sizes exist but it is much less transparent how to treat the variance estimates at each level (i.e., Should only the within-cluster residual variance be used or should it be combined with the between cluster variance? See Recchia, 2010 or Snijders & Bosker, 1994 for further discussion). It should be noted that the Adjusted R^2 statistic between CR-*SE*s and OLS will not be identical because the clustering complicates the degrees of freedom calculation (i.e., degrees of freedom are a function of the number of clusters rather than the number of people with CR-*SE*s).

Generalized Estimating Equations

Conceptual overview. Similar to CR-*SE*s, instead of accounting for the clustering by directly modeling random effects (and their associated covariance structure) as in HLM, the clustered structure of the data is treated more as a nuisance with GEE. Also similar to CR-*SE*, GEE use regression coefficient estimates from a single-level general(ized) linear model. Contrary to CR-*SE*s, GEE uses the residuals to iteratively estimate a working correlation matrix for observations within a cluster and these correlations are then used to obtain updated estimates of the regression coefficients that take clustering into account (Liang & Zeger, 1986; Zeger & Liang, 1986; Zeger, Liang, & Albert, 1988). CR-*SE*s are then applied at the end of the process to account for possible misspecifications of the working correlation matrix.

Somewhat similar to HLM, with GEE, researchers provide the initial working correlation structure that captures the general relation between observations within a cluster. However, unlike HLM, the working correlation structure does not have to be correct (and it is in fact not assumed that the working structure is correct) and only has to be in the very general vicinity of the population structure (Zeger et al., 1988). This is in opposition to HLM which assumes a specific form for the variance, namely $\mathbf{V}_j = Var(\mathbf{Y}_j) = \mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\mathrm{T}} + \mathbf{R}_j$ as described in Equation 2.

Using the initial working correlation matrix as a starting point, an algorithmic process estimates and updates the working correlation matrix based on the residuals to more accurately reflect the strength with which individuals within a cluster are related to one another. Unlike CR-*SE*s, both standard error and regression coefficient estimates are then updated to reflect the correlation between observations. That is, the correlation between the observations is used when estimating the regression coefficients, unlike CR-*SE* which assume working independence when estimating the regression coefficients. That is, generalized estimating equations will yield potentially different regression coefficient estimates than if clustering were ignored. This is in opposition to CR-*SE* which only corrects the standard errors and does not affect regression coefficient estimates.

The first step in the GEE algorithm fits the model assuming the data were independent (i.e., not clustered) as is similarly done with CR-*SE* estimation. Then, using information from the residuals of the independence model estimates, the initial values for the working correlation matrix are estimated in accordance with the structure provided by the researcher. The working structure is a blend of the **G** and **R** matrices in HLM (the model is not split into multiple levels, so the relation of observations within a cluster is a function of both **G** and **R**). Then, using the working correlation matrix, the covariance matrix of the outcome for individuals within a cluster is then estimated which is the same **V** matrix estimated in HLM. This matrix is then used to update the regression coefficient and standard error estimates to reflect the dependent relation between observations. The residuals from this updated model are then calculated and the process iterates between updating the working correlation matrix, the covariance matrix for the outcome, and the model estimates until the regression coefficients no longer change between iterations whereby the model is said to have converged to a solution. After this convergence, the cluster-robust estimator (the same one from CR-*SE*s) is applied to account for any potential misspecifications in the covariance structure and the final regression coefficient and standard error estimates are output, with the clustering taken into account. The specifics of the algorithm are rather involved and are presented in full detail in Appendix A for interested readers. A conceptual flowchart of the algorithm is shown in Figure 1 to help simply the process.

Assumptions, properties, and advantages of GEE. As with CR-*SE*s, GEE yield estimates that take the clustering of observations into account without specifying any random effects in the model. The GEE algorithm will attempt to accommodate the covariance that exists between observations due to clustering and GEE will yield regression coefficient estimates with the standard error estimates being corrected for the clustered nature of the data. That is, coefficient estimates with GEE and CR-*SE* will be almost, if not, identical if GEE uses an independent working structure but estimates can be quite different if GEE uses a more complex working structure, especially with discrete outcomes.

Modeling with GEE requires a smaller quantity of assumptions compared to HLM because GEEs do not require assumptions about random effects—five assumptions are made:

1. All relevant predictors are included in the model.
2. Observations between clusters are not related (there is not a higher level of the hierarchy).
3. Sample size is sufficiently large for asymptotic inferences at the cluster level.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

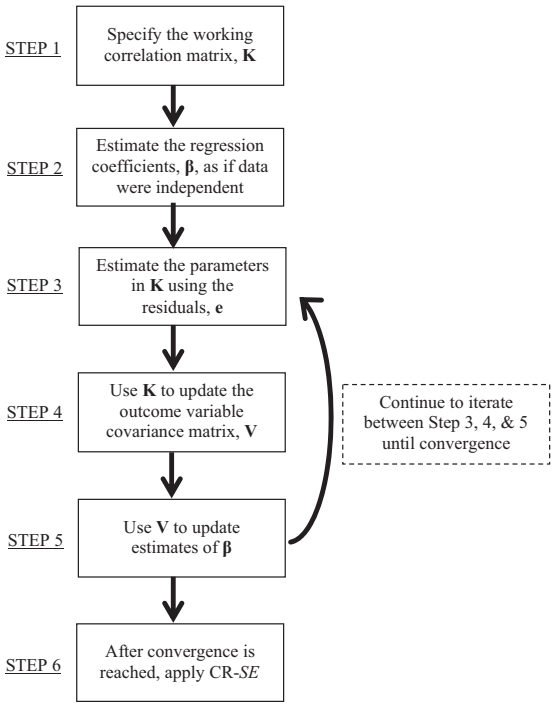

Figure 1. Conceptual flowchart of GEE algorithm.

4. The working correlation matrix is "reasonably close" to the population structure.

5. Without preprocessing, missing data are assumed to be MCAR (this issue will be discussed in more detail when PAMs and HLM are directly compared in a subsequent section).

With GEE, there is some researcher input required when selecting the working correlation structure; however, it does not have to be properly specified and estimates are robust to fairly large misspecifications (Ballinger, 2004; Zeger et al., 1988). To explicate more on "grossly misspecified," Zeger, Liang, and Albert (1988) found that for an ICC of 0.30 or less, using an independent working correlation structure (the most basic structure) resulted in similar estimates to an exchangeable structure,[5] so selection of the working correlation matrix for cross-sectional clustering should not present too much of an issue for data common in psychology. For researchers in fields of psychology where higher ICCs are common, choosing the appropriate working structure is still rather free from deliberation because an exchangeable working matrix is usually the most appropriate for cross-sectionally clustered data regardless of ICC. With higher ICCs, an independent structure may not be a viable option and thus GEE with an exchangeable working structure may be preferred to CR-*SE*s because the estimation will be more efficient meaning that power will be augmented due to increased precision of the estimates (Hanley et al., 2003).

The main difference between GEE and other PAMs is that GEE does not use traditional likelihood methods to arrive upon model estimates. As disadvantages, model comparison procedures based on the likelihood such as likelihood ratio tests or traditional information criteria like AIC or BIC are not available with the traditional GEE approach nor are model fit measures such as R^2 that can be computed simply with CR-*SE*s with continuous outcomes. As an alternative, Pan (2001) developed the Quasi-Likelihood Information Criteria (QIC) that extends the idea of AIC to GEE where lower values indicate better fit. Additional criteria, particularly for adjudicating the appropriate working structure, include Rotnitzky and Jewell's criterion (Rotnitzky & Jewell, 1990), correlation information criterion (Hin & Wang, 2009), and Gosho's criterion (Gosho, Hamada, & Yoshimura, 2011).

As an advantage of using an alternative estimation scheme, GEE are far faster and simpler to estimate with discrete outcomes compared with HLM and the estimation scheme used by GEE does not depend on whether the outcome is continuous or discrete.[6] Thus, GEE are commonly advantageous compared to HLM for discrete outcomes or for CR-*SE*s with longitudinal data because ICC values associated with repeated measures data are often rather large.[7]

Comparison of PAMs and HLM

To ensure that estimates are consistent and appropriately incorporate the clustered nature of the data, either PAMs or HLM can be used provided that assumptions of each method are met and that sample sizes are adequately large. The range of potential research questions that can be answered using HLM is greater because the estimation of random effects allows for more nuanced analyses that can more fully exploit information arising from the multilevel structure. However, as a result of the inclusion of random effects, HLM requires a greater quantity of assumptions than PAMs, although (as noted earlier) HLM is robust to violations of some of these assumptions under certain conditions. A comparison of what information is provided by each method is presented in Table 2 and the relative advantages and disadvantages of each method are listed in Table 3.

Sample Size

Both HLM and PAMs require large sample sizes to produce consistent estimates that effectively account for clustering. With PAMs, researchers must be concerned with the sample size at the cluster level whereas the sample size at both the cluster level and

[5] An independent working structure is the most basic structure and constrains all nondiagonal entries to be 0. An exchangeable structure is also fairly simple and constrains all nondiagonal entries of the working structure to be a single value, thus only requiring two total parameters (one for the diagonal entries and one for the off-diagonal entries). Reverting to the depression score example, the exchangeable structure would be reasonable if one considered that people are equally correlated with all other people in the clinic and there was no underlying reason why certain pairs of people should be more correlated than other pairs.

[6] GEE are typically estimated with quasi-likelihood methods which only require a mean and variance function. Therefore, the full likelihood function is not utilized and the lack of a closed form solution is not problematic for estimation.

[7] Although we are not extensively discussing longitudinal data, we would like to note that GEE can incorporate many working structures for longitudinal data that are preferable to working independence assumed with CR-*SE*s such as autoregressive structures.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Comparison of Information Available for Each Method

	Ignoring clustering	CR-*SE*	GEE	HLM
SE based on	Information/closed form (OLS)	Sandwich	Sandwich	Information
Within-cluster correlations	None	Exchangeable	User-specified	Fully modeled
Inference for fixed effects	Yes, if no clustering	Yes	Yes	Yes (assuming proper specification)
Inference for variance parameters	N/A	N/A	N/A	Yes, (depending on software*)
Supports cross-classification	N/A	Yes	No	Yes
Discrete outcomes				
Assumption for point estimates	Independence	Independence	Working correlation	Fully modeled
Interpretation of fixed effects	Textbook	Population-averaged	Population-averaged	Cluster-Specific
Estimation			Based on quasi-likelihood, only first two moments are needed, so numeric integration not necessary	The likelihood does not have closed form, requires numeric integration or approximation

Note. We thank an anonymous reviewer for recommending this table and for suggesting the format and a wealth of the information.
* The popular lme4 package for HLM in R purposefully does not provide standard error estimates for variance components on theoretical grounds because the sampling distribution is likely to be strongly asymmetric and standard errors are an inadequate measure of uncertainty (Bates, 2009). GEE = generalized estimating equations; HLM = hierarchical linear model; CR-*SE* = cluster-robust standard errors.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

within-cluster level must be considered with HLM. Recommendations for how large is "large enough" vary across and within each method and common suggestions for HLM are 30 clusters of size 30 (Kreft, 1996), at least 20 clusters (Snijders & Bosker, 2011), or 50 clusters of size 20 for cross-level interactions or 100 clusters of size 10 for interest in variance components (Hox, 1998, 2010 p. 235). For GEE and CR-*SE*, "large enough" is typically considered to be a minimum of about 50 clusters (e.g., Angrist & Pischke, 2008; Cameron, Gelbach, & Miller, 2011; Lu et al., 2007; Mancl & DeRouen, 2001; Morel, Bokossa, & Neerchal, 2003). Because HLM relies on reasonable samples at the within-cluster level to estimate the random effects, a simulation by McNeish (2014) suggested that PAMs are more advantageous with cluster sizes less than five, particularly for discrete outcomes.

There are also some differences between methods relating to how individuals are allocated among clusters (i.e., balanced vs. unbalanced clusters). CR-*SE*s can be sensitive to data where some clusters have many more observations than other clusters (Nichols and Schafer, 2007). Both HLM and GEE are fairly robust to unbalanced clusters although the GEE algorithm may encounter convergence issues if there is extreme imbalance (Verbeke, Fieuws, Molenberghs, & Davidian, 2014). If researchers have large samples and the clusters widely vary with regard to the number of observations within each, then HLM may be the best method to handle these data. This can be common, for instance, when people are clustered within geographical areas (countries, states, etc.) because certain areas are larger or more populous than others. Our third real data example will demonstrate this point.

For each method, many small sample procedures have been developed. A comprehensive discussion of small sample corrections is outside the scope of this basic commentary. Readers interested in small sample problems with clustered data are referred to McNeish and Stapleton (2014) for a nontechnical review of small samples with HLM, to Lu et al. (2007) or Westgate (2013) for simulations comparing select small sample corrections for GEE and CR-*SE*s, or McNeish and Stapleton (in press) for a comparative simulation of several corrections used in both HLM and GEE.

Specifying Random Effects

Although HLM is robust to misspecification of random effect-related assumptions under certain conditions, it is still important to attempt to model this portion of the model correctly because the random effects are substantively meaningful (otherwise a method that does not incorporate random effects could more parsimoniously be employed). The general concern is that the variances of the random effects are used to assess whether the random effects should be retained in the model. If the random effects have a large amount of variance, then the effect of the particular predictor on the outcome is quite different across clusters and the random effect should be retained in the model to capture this variability.

Although conceptually straightforward, the difficulty in model selection emerges from the fact that variances are typically constrained so that they are bounded below by zero (although, see Savalei & Kolenikov, 2008 for a discussion of when this is appropriate and how it affects inference). An inferential test that the variance of the random effects is equal to zero in the population tests a value at the boundary of the parameter space. That is, the test assesses the parameter at the lower bound of the possible values it can take (if the estimate is constrained to be non-negative as is the default in most software). The resulting distribution of the test statistic (typically a Z or χ^2 statistic) for this hypothesis may not follow the appropriate test distribution, meaning that inferences made from such tests may be untrustworthy (Molenberghs & Verbeke, 2004, 2007; Stram & Lee, 1994). Therefore, traditional likelihood ratio tests for the variance of the random effects may not be χ^2 distributed and traditional Z tests are not appropriate since the variance of the random effects is not symmetric.

Table 3
*Advantages and Disadvantages of HLM, CR-*SE*, and GEE*

Method	Advantages	Disadvantages
HLM	1. Can directly incorporate substantive multilevel theory into the model	1. Requires many explicit assumptions and is not always robust to violations
	2. Provides information about specific predictors having cluster-level variance, allows for cluster-specific inferences to be made	2. Cluster-specific interpretations and estimation difficult with discrete outcomes; Likelihood does not have a closed form solution with discrete outcomes which requires approximation or linearization
	3. Can more easily partition the variance into more than two levels and allows for full decomposition of cluster-level and within-level effects	3. Difficult to determine if the covariance is modeled correctly
	4. Accommodates either longitudinal or cross-sectionally clustered data well	4. Lacks an overall R^2 for continuous outcomes
CR-*SE*	1. Can output OLS-equivalent R^2 and effect sizes while accounting for clustering	1. Assumes working independence, coefficient estimates may be affected when the ICC is greater than .30
	2. Allow for multiple moderated or blockwise regression with clustered data	2. Less efficient than GEE for longitudinal analyses where the ICC is typically high
	3. Along with GEE, less affected by small cluster sizes (Level-1 sample size)	3. Compared with HLM, more affected by small number of clusters
GEE	1. Straightforward estimation with discrete outcomes; full likelihood not needed	1. Limited ability to compare models or gauge fit
	2. Estimates are robust to misspecifications to the covariance structure of the outcome	2. Compared with CR-*SE*, fewer advantages for cross-sectionally clustered data with continuous outcomes such as no R^2
	3. Along with CR-*SE*, less affected by small cluster sizes (Level-1 sample size)	3. Compared to HLM, more affected by small number of clusters and highly unbalanced clusters
	4. No distributional assumptions concomitant with random effects	4. Cannot fully decompose effects into between-level and within-level components.

Note. GEE = generalized estimating equations; HLM = hierarchical linear model; CR-*SE* = cluster-robust standard errors.

One potential remedy has been to compare the likelihood ratio test statistic with a mixture χ^2 distribution, which has been shown to produce better results in correctly identifying which variance components to retain in the model (Morrell, 1998). When selecting structures for the covariance of the random effects and residuals, methods such as AIC and BIC also have not performed well and in simulation studies they have been shown to recover the correct structure less than half of the time (Keselman, Algina, Kowalchuk, & Wolfinger, 1998; Wolfinger, 1993). It has been suggested that BIC performs poorly because sample size is included in the calculation; it is not clear if one should use the number of clusters, the total number of observations, or effective sample size that is a combination of the two which may not necessarily be the same for all model parameters (e.g., Newton & McCoach, 2015). This topic has been recently researched (e.g., Gurka, 2006; Lukočienė, Varriale, & Vermunt, 2010; Whittaker & Furlow, 2009) although no consensus has been reached.

Coefficient Interpretation

In single-level models such as OLS regression, most researchers are familiar with the general, "textbook" interpretation of the unstandardized regression coefficients of the general form: for a one-unit change in the predictor variable *X,* the outcome variable *Y* is expected to change by the value of the regression coefficient β, holding all other predictors in the model constant. This interpretation is referred to as the population-averaged or marginal interpretation because it applies to each observation in the dataset and the effect of predictor variable *X* on the outcome variable *Y* does not differ across observations. More formally, this relation can be written as $E(\mathbf{Y}|\mathbf{X}) = \mathbf{X\beta}$ - the expected value of the outcome conditional on the values of the predictors is equal to the values of the predictors times the regression coefficients.

With PAMs, the regression coefficients also take a population-averaged interpretation where $E(\mathbf{Y}|\mathbf{X}) = \mathbf{X\beta}$ meaning that the regression coefficients have an identical interpretation as a comparable single-level model. Conceptually, this follows from the nature of the correction applied to account for the clustering: the corrections make use of the residuals to adjust the standard error estimates so that the dependence of observations within clusters is adequately captured. No random effects are estimated for individual clusters and the model retains a fully fixed-effect specification, resulting in comparable interpretations to a traditional single-level model. The same cannot be said from HLM model estimates.

As a result of including the random effects in the model, the interpretation of the regression coefficients refer to cluster-specific estimates rather than the population-averaged estimates produced by single-level models or PAMs. Rather than regression coefficients having the population-averaged interpretation of a single-level model, HLM regression coefficients are interpreted as: for a one-unit change in the predictor variable *X,* the outcome variable *Y* is expected to change by the value of the regression coefficient β, holding all other predictors in the model constant *and given equal values for the random effects*. Notationally, $E(\mathbf{Y}|\mathbf{X}, \mathbf{u}) = \mathbf{X\beta} + \mathbf{Zu}$. This difference is minimal when the outcome is continuous (Ballinger, 2004), but, when the outcome is discrete, the PAM estimates can be quite different and will tend to be

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

slightly smaller in magnitude as compared with the HLM estimates because the two methods are estimating regression coefficients that are representative of truly different quantities (Ghisletta & Spini, 2004). Furthermore, Carlin, Wolfe, Brown, and Gelman (2001) have investigated the cluster-specific interpretation of HLM with discrete outcomes and questioned whether estimates actually reflect this interpretation and to which quantity the estimates were actually referring. Coefficients for predictors at the cluster level would not be affected by the population-averaged or cluster-specific interpretation and their magnitude would be expected to be the same between methods under ideal circumstances.

If the specific clusters or estimates of the variability of clusters are relevant to the underlying research question, then an HLM approach is warranted; however, the differential interpretation of regression coefficients is one such peril of resorting to HLM solely to produce estimates that account for clustering. If a researcher essentially desires a single-level model with estimates that properly account for clustering, this is *not* what it is reflected by HLM estimates and researchers may risk interpreting the coefficients incorrectly or estimating quantities which do not align with their research interest. This contrast will be demonstrated in detail in the second and third real data examples.

Contextual Effects

Kreft, de Leeuw, and Aiken (1995) demonstrated that researchers can decompose the effect of a single predictor at each level to determine whether the effect is the same within-clusters and between-clusters. This is done by centering the within-cluster predictor around either the cluster mean or the grand mean and then including the same predictor at the between-cluster level (aggregated over the observations within a cluster; Enders & Tofighi, 2007). If the effect is not the same at different levels of the hierarchy then there is said to be a *contextual* or *compositional* effect.

Among researchers who advocate the use of HLM, the ability of HLM to fully decompose effects into between-cluster effects and within-cluster effects is typically cited as this is not permissible with OLS regression (Hoffmann & Gavin, 1998). However, if researchers are not interested in explicitly understanding the variability of the random effects, then random effects are not required to estimate the population-averaged contextual effect (Begg & Parides, 2003; Berkhof & Kampen, 2004; Snijders & Bosker, 2011, p. 106). That is, if the investigation of a contextual effect is concerned only with inferentially testing if there is an effect in the population but not necessarily inspecting the effect for specific cluster, then this same information can be obtained from PAMs and does not necessitate HLM. PAMs make use of the clustering variable to correct standard errors and therefore, within-cluster predictors can be centered just as in HLM and the cluster mean added as an additional variable into the model and effects can be similarly decomposed, at least for population-averaged inference. As noted in previous sections, PAMs will not be able to partition the variance into within and between components. PAMs are suitable for simply estimating the population averaged regression coefficient, which this is routinely done in epidemiological and economic studies (for empirical examples, see Agerbo, Sterne, & Gunnell, 2007; Huynh, Parker, Harper, Pamuk, & Schoendorf, 2005; Kontos, Burchinal, Howes, Wisseh, & Galinksy, 2002; Marschall, 2004; Petronis & Anthony, 2003).

Missing Data

A common concern in psychological research is missing values in the data. As a result of the process used to fit the model, GEE are only implicitly consistent when data are MCAR based on the classification in Rubin (1976). Standard GEE is not a likelihood method and therefore likelihood-based corrections cannot be applied to data that are MAR (Ghisletta & Spini, 2004). While GEE's assumption that missing data are MCAR may cause researchers concern, especially those with longitudinal data because of the high prevalence of missing data in such designs, Fitzmaurice, Laird, and Rotnitzky (1993) found that bias of GEE with data that are MAR was small—relative bias was found to be less than 5% unless the amount of missing data was quite large (50%) and the model was misspecified. Furthermore, researchers are not bound to the MCAR assumption with GEE and methods such as weighted GEE (Chen, Yi, & Cook, 2010; Lipsitz, Ibrahim, & Zhao, 1999; Robins, Rotnitzky, & Zhao, 1995) or preprocessing the data with multiple imputation (Rubin, 1987) are valid ways to accommodate MAR missingness with GEE, provided that certain assumption are met (e.g., specifying a proper imputation model; for discussion and comparisons of these methods, readers are referred to Beunckens, Sotto, and Molenberghs, 2008; Carpenter, Kenward, & Vansteelandt, 2006; Clayton, Spiegelhalter, Dunn, & Pickles, 1998; Scharfstein, Rotnitzky, & Robins, 1999). Alternatively, likelihood GEE-type models can be estimated as well, which can implicitly handle MAR missingness (see Example 38.12 in the SAS 9.2 Manual, p. 2381). Although we will not go into detail regarding weighted GEE, we do note that weighted GEE is a preprogrammed option in the new Proc GEE procedure in SAS 9.4.

On the other hand, HLM is typically estimated with likelihood based methods and, consequently, estimates are consistent when the outcome is MAR or MCAR provided that the observed Fisher information matrix is used (Kenward & Molenberghs, 1998).[8] If the expected Fisher information matrix is used, then HLM estimates are consistent only under the MCAR missingness (Verbeke & Molenberghs, 2007). For more detail on the distinction between observed and expected Fisher information, interested readers are referred to Enders (2010, pp. 100–102), Kenward and Molenberghs (1998), or Savalei (2010). These studies address missingness only on the outcome variable and missing covariates can present more of an analytical challenge (Horton & Laird, 1999, 2001). M*plus* is capable of handling missing data with likelihood methods regardless of which variables in the model have missing values (Allison, 2012; Enders, 2010).

As a cautionary note, we remind readers that when variables have missing values, the condition of MAR is not to be assumed

[8] As a brief software note, some software programs (e.g., SAS Proc Mixed) only use likelihood methods to accommodate missing values on the outcome variable and listwise delete observations that are missing values on the predictor variables. Researchers with missing data will want to confirm that their software is handling missing values as intended (see, Allison, 2012; Enders, 2010).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

by default—if addressed with likelihood methods, then the model must contain all variables that are related to missingness for this assumption to be upheld. As such, the requirement that data must be preprocessed with traditional GEE to accommodate MAR missingness, while possibly inconvenient at times, is not necessarily a strict disadvantage. As noted in Enders (2010) "Given that the two procedures (likelihood methods and multiple imputation) frequently produce very similar results, the choice of technique is often personal preference" (p. 336).

Illustrative Real Data Examples

Three illustrative example datasets with clustered observations are used to demonstrate the near equivalence of the regression coefficients from HLM, CR-*SE*, and GEE in applied research with continuous outcomes and their divergence with discrete outcomes.

The first dataset comes from an Institute of Educational Sciences (IES) grant that investigated the efficacy of a reading intervention to assess whether word knowledge and comprehension at posttest were greater for students receiving a treatment applied at the classroom level compared to students in the control group (there were six classrooms in each group). Two examples from this study will be shown—one for modeling *word knowledge* and one for modeling *receptive vocabulary*. The data used for the example models include 203 kindergarten students clustered within 12 classrooms in a semiurban, Mid-Atlantic, school district.[9] Word knowledge was measured by the Peabody Picture Vocabulary Test Growth Score Value (PPVT-GSV) and was predicted by treatment group status, English language learner (ELL) status, and PPVT-GSV pretest score. Receptive vocabulary was measured by a researcher-constructed scale and the model featured the same predictors except that receptive vocabulary pretest scores were used instead of PPVT-GSV pretest scores. Our goals for these first two examples are (a) to show the similarity of estimates between methods when the outcome is continuous and interest is on inference regarding the regression coefficients, and (b) to show how violations of the additional assumptions when using HLM to account for clustering can adversely affect inference. Assumption checking with each method is explicitly shown in the word knowledge model to concretize the number of assumptions that must be tested between methods.

The third example utilizes a second dataset that appeared in Snijders and Bosker (2011) and originated from Ruiter and Van Turbergen (2009) which models the probability that 135,508 people clustered within 59 countries attend religious services at least once per week as predicted by sex, age, education level, income, unemployment, marital status, urbanization, and the Gini Index, which measures the degree of wealth distribution in a country. With this example our goals are (a) to show how estimation with HLM can be trying with discrete outcomes, (b) to demonstrate how the coefficient estimates will be different between HLM and PAMs because the distinction between cluster-specific and population-averaged interpretations is relevant, and (c) to show how HLM encounters fewer problems with very large disparities in cluster sizes compared to GEE.

Word Knowledge Model

The unconditional ICC for PPVT-GSV posttest scores was estimated to be 0.21 and the square root of the design effect (DEFT) was 2.21, indicating that the clustering of students within classrooms would have a non-negligible impact on standard error estimates if clustering were ignored. The statistical models for HLM (in Raudenbush & Bryk, 2002 notation) and for GEE/CR-*SE* are provided below. To facilitate implementing these models in applied studies, Appendix B provides software code for HLM, GEE, and CR-*SE* in SAS, Stata, and M*plus* for the word knowledge model.

HLM:

$$\begin{aligned} Post\text{-}Test_{ij} &= \beta_{0j} + \beta_{1j} \times ELL_{ij} + \beta_{2j} \times Pre\text{-}Test_{ij} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} \times Treatment_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} \times Treatment_j \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} \times Treatment_j \end{aligned}$$

GEE/CR − *SE*:

$$\begin{aligned} Post\text{-}Test_{ij} = \beta_{0j} &+ \beta_1 \times ELL_{ij} + \beta_2 \\ &\times Pre\text{-}Test_{ij} + \beta_3 \times Treatment_j + \beta_4 \\ &\times (Pre\text{-}Test_{ij} \times Treatment_j) + \beta_4 \\ &\times (ELL_{ij} \times Treatment_j) + e_{ij} \end{aligned}$$

GEE/CR-*SE* assumptions. The assumptions of GEE and CR-*SE* are fairly similar so the assumption tests will be discussed together. Assumption 1 of both methods is that all relevant predictors are included; the word knowledge model was determined theoretically and model fitting is not of interest, so it will be assumed that the appropriate predictors are in the model. Assumption 2 requires that observations between clusters (e.g., students in different classrooms) are not related to one another. The ICC for a third level (school) was calculated but it was quite small (less than 0.03, DEFT = 1.19) so this assumption appears to be reasonably upheld. Assumption 3 concerns the sample size at the cluster-level—although there are only 12 classrooms, the Kauermann-Carroll correction to the sandwich estimator has been documented to perform well with as few as 10 clusters (Lu et al., 2007) and will be used here. Assumption 4 for GEE requires that the working covariance matrix be "reasonably close" and, for cross-sectional data, the most logical choice for the working correlation matrix is either an exchangeable or independent structure. QIC favored the independent structure to the exchangeable structure (QIC_{IND} = 14,007.33 *vs.* QIC_{EXCH} = 14,087.54). However, Hin, Carey, and Wang (2007) noted that the Rotnizky-Jewell criterion (RJC) is more suitable for distinguishing between independent and exchangeable working structures. For the word knowledge model, RJC_{IND} = 12141.20, RJC_{EXCH} = 7542.74 as calculated by the CriteriaWorkCorr SAS macro (Gosho, 2014). Thus, GEE was used with an exchangeable working structure. There were no missing values in these data.

HLM assumptions. Similar to GEE/CR-*SE*, Assumption 1 requires that all relevant predictors be included in the model and is assumed to hold for the same reason as above. Assumption 2 requires that all relevant random effects be included in the model; as discussed earlier, there is no straightforward method to test this assumption. Although not without its disadvantages, a likelihood ratio test compared to a mixture χ^2 distribution has been recom-

[9] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A110142 to the University of Maryland. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

mended and will be implemented here through the Covtest statement in SAS Proc Glimmix. A model with random effects for all within-cluster predictors fits significantly better than a model with zero random effects ($\chi^2 = 77.41$, $p < .001$). A model with just random intercepts did not fit appreciably worse than a model with three random effects ($\chi^2 = 0.11$, $p = .613$), a model with random effects for the intercept and pretest ($\chi^2 = 0.10$, $p = .751$), or a model with random effects for the intercepts and ELL ($\chi^2 = 0.01$, $p = .956$) and therefore only the random intercept was included in the model.

Assumptions 3 and 4 address the covariance structures of the within-cluster residuals and random effects, respectively. Because the model has only one random effect, Assumption 4 addressing the covariance structure of the random effects is not a concern. With regard to the within-cluster residual covariance structure, the cross-sectional nature of the data makes a diagonal structure the most reasonable which was confirmed as other structures (e.g., compound symmetric) did not improve the fit of the model.

Assumptions 6 and 7 address the normality of the within-cluster residuals and the random effects, respectively. Figure 2 shows the histogram of the within-cluster residuals on the left panel and the histogram of the random effects on the right panel. The left panel shows a slight negative skew and inferential normality tests such as Cramer-Von Mises ($W^2 = 0.11$, $p = 0.08$), and Anderson-Darling test ($A^2 = 0.68$, $p = 0.08$) were not significant at the 0.05 level. The right panel is more difficult to interpret because of the small number of clusters, a pervasive problem in behavioral sciences (Dedrick et al., 2009; McNeish & Stapleton, 2014). With 12 clusters, it is difficult to discern whether the distribution is normal and inferential tests are not trustworthy because they are highly underpowered at small sample sizes. There is no drastic violation of Assumption 7, but it is difficult to be confident that it is upheld.

Assumption 8 requires that the random effects are not correlated with predictors in the model. For discrete predictors with few categories (treatment status and ELL in this model), the variance of the random effects can be calculated separately by group. Table 4 shows the intercept variance estimate when estimated separately by group for ELL and treatment status. In Table 4, there appears to be little difference in the variance components by ELL but the difference between treatment status may be worrisome. This raises another difficulty in assumption testing with HLM in that it is difficult to compare the variance components by group. The scale of the variance components is not always intuitive and the variance components are often constrained to be non-negative, so placing a traditional confidence interval around the estimate using its standard error is not very informative. Furthermore, for the modest sample sizes often seen in psychology, power will be quite low and, the standard error estimates of variance components are highly sensitive to the deviations from normality (Maas & Hox, 2004). For the continuous predictor (pretest score), this assumption can be tested by plotting the predictor values against the random effect estimates. The linear correlation of about .17 resulted in a *p*–value slightly below 0.05. Figure 3 shows that there is a slight quadratic trend and that this assumption may be questionably upheld—given uncertainty associated with the smaller sample size, we proceed presuming that this assumption was reasonably upheld. To address Assumption 9, the small number of clusters was addressed with a Kenward-Roger correction which has been demonstrated in the literature to perform well with as few as 10 clusters (e.g., McNeish & Harring, 2015). The number of students within each cluster was also in the double digits which is adequately large (e.g., McNeish, 2014).

Table 4
Intercept Variance Estimates by Group

Group	Intercept variance	Standard error
ELL status		
Not ELL	5.78	7.24
ELL	4.34	5.84
Treatment status		
Control	2.70	4.68
Treatment	9.34	8.74

Note. ELL = English language learner.

Comparison of results. Table 5 compares the estimates between HLM, CR-*SE*, and GEE as estimated in SAS 9.3. From Table 5, it can be seen that the coefficient estimates, their

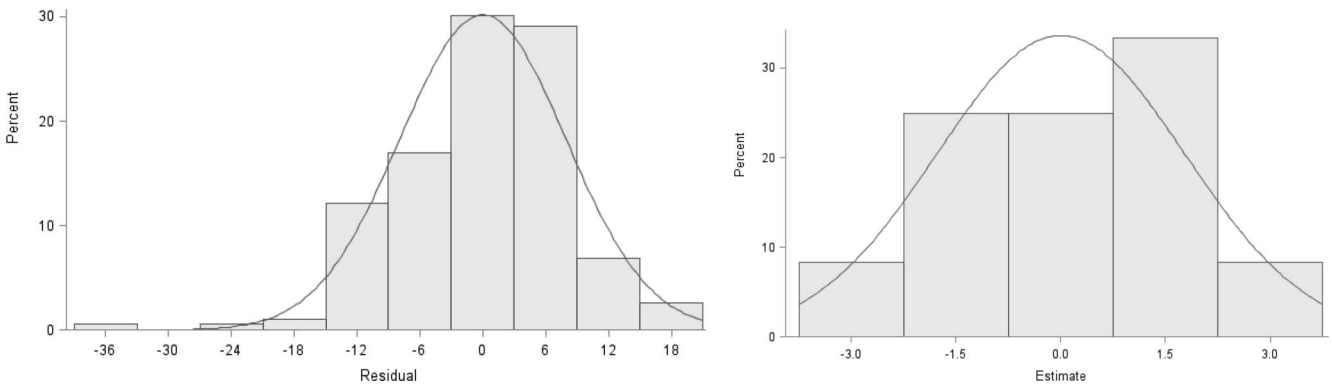

Figure 2. Normality plots for within-cluster residuals (left) and intercept random effects (right). The within-cluster residuals look approximately normal; the intercept random effects are difficult to interpret because of the rather small number of clusters.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

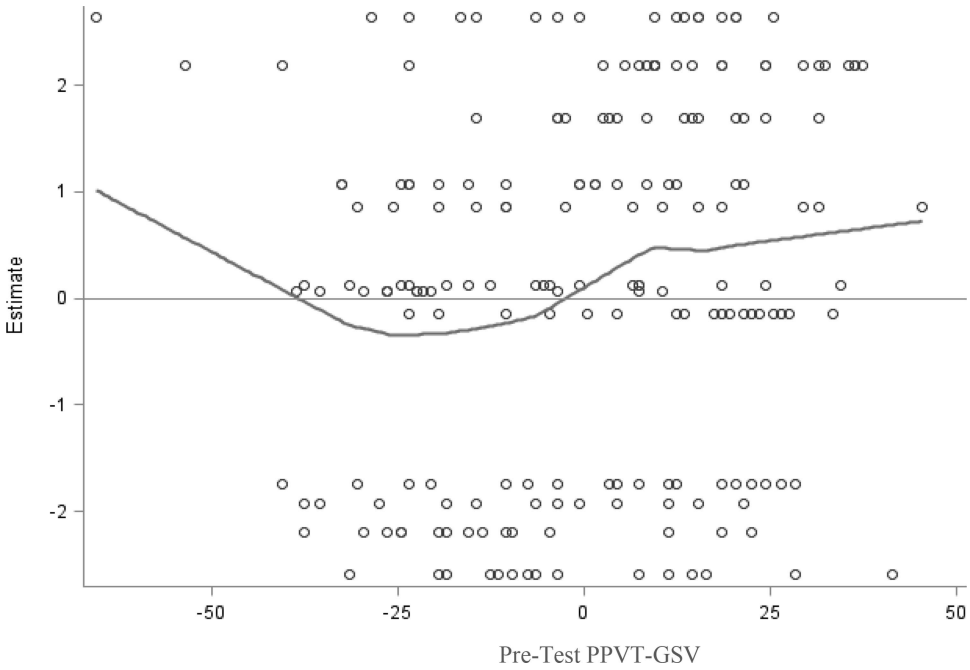

Figure 3. Random intercepts versus pretest PPVT-GSV scores with imposed linear trend. The linear Pearson correlation was .17 which significant at the 0.05 level but not the 0.01 level. The plot of the within-cluster residual and random intercepts is not shown, but the linear correlation was −.01 and showed no pattern of violation.

standard errors, and their *p* values are fairly close across methods. As also mentioned previously, the CR-*SE* model can also estimate an R^2 that is analogous to OLS which cannot be similarly done in HLM (although comparable level-specific R^2 values can be computed from HLM estimates for models without random slopes). Additionally, although both GEE and CR-*SE* use the sandwich estimator to estimate standard errors, the effect of incorporating the working covariance matrix into the coefficient estimates in GEE can be seen readily in the estimates of the ELL and ELL × Treatment estimates: the inference does not differ at the .05 level (although they would differ at the 0.10 level and perhaps with a larger sample size) but the estimates are noticeably different between the two methods.

The primary interest was determining if the treatment was effective for increasing word knowledge and, by virtue of the population of interest, students were nested within classrooms which had to be accounted for in the model. The clustered structure was not an inherent research interest, so accounting for clustering without modeling any cluster-specific random effects could more simply account for the clustering while avoiding the assumptions required when modeling with random effects. The interpretation of the regression coefficients with GEE or CR-*SE* are also as clear, if not clearer, than an HLM

Table 5
*Comparison of Estimates for Word Knowledge Data From HLM, GEE, and CR-*SE *With* SE*s in Parenthesis*

	HLM		GEE		CR-*SE*	
Effect	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
Intercept	123.51 (4.09)	—	123.51 (4.29)	—	129.78 (3.15)	—
Treatment	10.62 (5.74)	.09	10.62 (6.14)	.11	9.88 (6.34)	.15
ELL	2.89 (2.48)	.25	2.90 (4.04)	.47	−8.21 (5.25)	.26
ELL × Treatment	−6.39 (3.35)	.06	−6.38 (3.56)	.08	−4.70 (2.84)	.58
Pretest	.87 (.06)	<.01	.86 (.07)	<.01	.73 (.08)	<.01
Pretest × Treatment	−.20 (.08)	<.01	−.20 (.09)	.02	−.19 (.08)	.03
Intercept variances	85.56	—	—	—	—	—
Residual variances	64.07	—	—	—	115.77	—
Exchangeable correlation	—	—	.51	—	—	—
R^2	—		—		.81	

Note. Pretest was cluster-mean centered. GEE = generalized estimating equations; HLM = hierarchical linear model; CR-*SE* = cluster-robust standard errors; ELL = English language learner.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

(including the ability to calculate an R^2 more straightforwardly with CR-*SE*). More plainly, the clustering mechanism served no substantive interest, so the clustering was treated as a nuisance with PAMs which simplified the modeling process. HLM can still be used to account for clustering, but, if the interest is primarily on the regression coefficient estimates, there is a reliance on proper modeling of the covariance structures and an assumption that the random effects are not correlated with the predictors; these assumptions are required to obtain estimates of quantities that the researcher may not care about (i.e., random effect variance). PAMs obviate the need to properly model quantities that are not a direct research interest and adhering to all HLM assumptions will result in the same estimates provided by PAMs. The next example will demonstrate the ramifications when assumptions are not upheld and one employs HLM merely to account for clustering.

Receptive Vocabulary Model

With the receptive vocabulary model, two aspects exacerbated possible violations of HLM Assumption 5 regarding covariance between random effects and residuals and Assumption 8 regarding relations between random effects/residuals and predictor variables. First, pretest scores were a stronger predictor of posttest scores of receptive vocabulary than word knowledge. Second, possibly due to the smaller sample size, the randomization process was not completely successful and the treatment group was about 0.35 of a standard deviation higher than the control group at baseline on the pretest measure. Thus, when fitting the HLM model from Equation 1, the intercept random effects and the within-cluster residuals were clearly correlated (left panel of Figure 4) as were the pretest scores and the intercept random effects (right panel of Figure 4), violating Assumptions 5 and 8. HLM is not robust to violations of this assumption (Bates, Castellano, Rabe-Hesketh, & Skrondal, 2014; Kim & Frees, 2007). As a result, although the coefficient estimates for HLM (with Kenward-Roger adjustment) and CR-*SE* (with Kauermann-Carroll adjustment) are quite close in Table 6, the standard errors are quite different which changes the inference decision (using a .05 level of significance) for the ELL predictor (GEE estimates were fairly close to CR-*SE* and are not reported for brevity). In an attempt to allay concerns of these violations due to baseline imbalance, the data were weighted by inverse propensity scores[10] to obtain more equivalent covariate values at baseline between the treatment group and the control group (which initially differed by 0.35 of a standard deviation). The data were then remodeled with the weights incorporated. As can be seen in Table 6, the weighted HLM and weighted CR-*SE* coefficient estimates are again very similar but the weighted HLM standard error estimates are much smaller than the unweighted HLM standard errors and fairly closely mirror both the weighted and unweighted CR-*SE* standard error estimates (for which the assumptions were not violated). Therefore, it can be assumed that the reweighting for baseline nonequivalence successfully addressed assumption violation in the HLM estimation.

Importantly, this example shows a potential pitfall of using HLM to account for clustering when cluster-specific inference or variance partitioning are not of interest—HLM can certainly accomplish the task of estimating standard error estimates that account for clustering but does so in a more complex manner and entails a larger number of assumptions. If cluster-specific inference or variance partitioning are not of interest, GEE or CR-*SE* can streamline the process of accounting for clustering by making as few assumptions as possible.

Religious Attendance Model

The religion data will be used to demonstrate how the choice of method, as well as the estimation method, can affect results when the outcome is discrete. In equation form, the models can be written as,

HLM:

$$\ln\left(\frac{p(Attend_{ij})}{1-p(Attend_{ij})}\right) = \beta_{0j} + \beta_{1j} \times Female_{ij} + \beta_{2j} \times Income_{ij} + \beta_{3j} \times Education_{ij} + \beta_{4j} \times Single_{ij} + \beta_{5j} \times Divorced_{ij} + \beta_{6j} \times Widowed_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \times Gini_j + \gamma_{02} \times Urbanization_j + \gamma_{03} \times College\ Enrollment_j + \gamma_{04} \times Years\ of\ Education_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \times Gini_j + \gamma_{12} \times Years\ of\ Education_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\mathbf{u} \sim MVN(0,\ \boldsymbol{\tau}), \boldsymbol{\tau} = \begin{pmatrix} \tau_{00} & 0 & 0 \\ 0 & \tau_{11} & 0 \\ 0 & 0 & \tau_{22} \end{pmatrix}$$

GEE/CR–*SE*:

$$\ln\left(\frac{p(Attend_{ij})}{1-p(Attend_{ij})}\right) = \beta_0 + \beta_1 \times Female_{ij} + \beta_2 \times Income_{ij} + \beta_3 \times Education_{ij} + \beta_4 \times Single_{ij} + \beta_5 \times Divorced_{ij} + \beta_6 \times Widowed_{ij} + \beta_7 \times Gini_j + \beta_8 \times Urbanization_j + \beta_9 \times College\ Enrollment_j + \beta_{10} \times Years\ of\ Education_j + \beta_{11} \times (Female_{ij} \times Years\ of\ Education_j) + \beta_{11} \times (Female_{ij} \times Gini_j)$$

The clustering is cross-sectional and the outcome variable is binary meaning that the residual covariance matrix for HLM models does not have to be explicitly modeled because the model does not have an explicit error term (i.e., HLM Assumptions 3 and 6 do not apply with discrete outcomes). The covariance matrix of the random effects does still need to be explicitly modeled with HLM with discrete outcomes. Based upon nested mixture χ^2 tests available in the Covtest statement in Proc Glimmix (see SAS

[10] The propensity score model was built using logistic regression with treatment status as the outcome and the battery of pretest scores given to students as well as various demographic information and interactions thereof. Propensity weights were then stabilized using the method outlined in Harder, Stuart, and Anthony (2010).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

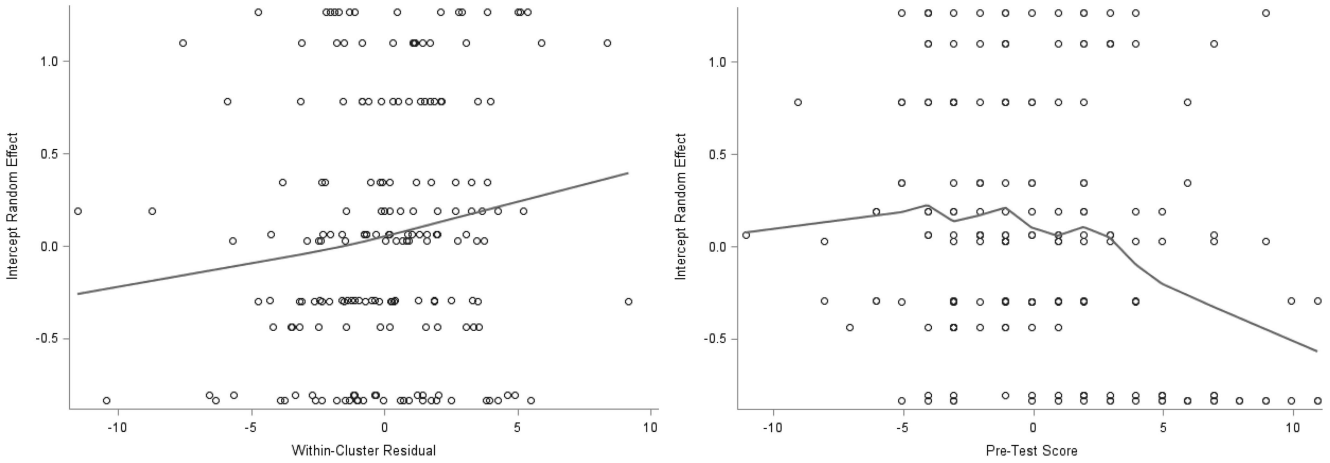

Figure 4. Random intercepts versus within-cluster residuals (left) and random intercepts versus receptive vocabulary pretest scores (right), each with an imposed loess line. The linear Pearson correlation between random intercepts and the within-cluster residual was .14; the Pearson correlation between pretest scores and random intercepts was −.24. Both were statistically significant. If assumptions were met, the loess lines should be approximately horizontal through zero.

Usage Note 40724 for more detail), tests suggested to include random effects for the intercept, female status, and income variables and the matrix was specified to have a heterogeneous diagonal structure. Covariances among the random effects were considered but their inclusion did not appreciably improve the fit of the model, so covariances were constrained to zero for the sake of parsimony. Using an exchangeable working structure with GEE led to model convergence issues unless the convergence criterion was heavily relaxed (likely due to the extreme imbalance of the cluster sizes which ranged from 95 to 7,745; Verbeke et al., 2014); therefore, an independent working structure was used with GEE.

Comparing estimation times, HLM with PQL took approximately 8 min to converge, HLM with AGQ took about 18 hr to converge, and GEE took about 3 s to converge. Computational times can be reduced by relaxing convergence criterion. For instance, M*plus* uses a default convergence criterion of 1E-3 for this model (compared with 1E-8 in Proc Glimmix) and reached convergence in 4.25 hr with roughly the same parameter estimates. If the same convergence criteria are used, then computational times between programs are comparable. Readers should note that computational times are high with this example because of the vast sample size; more reasonably sized data sets that are typical in psychology will not feature such computational overhead.

Table 7 compares estimates from an HLM estimated with PQL, HLM estimated with AGQ and 10 integration points (as recommended by Pinheiro & Bates, 1995; Pinheiro & Chao, 2006), and GEE with an independent working correlation matrix. HLM with a Laplace Approximation yielded estimates and *p* values that were equal to the second decimal point with HLM estimated by AGQ. Similarly, GEE with an independent working matrix yielded estimates and *p* values that were equal to the second decimal point with CR-*SE*s. We therefore only presented one of each related method for brevity. To facilitate implementing these models in applied studies, Appendix C provides software code for all three HLM methods, GEE, and CR-*SE* in SAS, Stata, and M*plus* for estimation methods that are available in each respective program.

Unlike the first example, the outcome in these models is discrete meaning that the estimates between HLM and GEE are expected to

Table 6
*Comparison of HLM and CR-*SE *Estimates for Receptive Vocabulary Where HLM Assumptions are Seriously Violated*

	HLM		CR-*SE*		IPW HLM		IPW CR-*SE*	
Effect	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
Intercept	14.35 (.82)	—	14.34 (.52)	—	14.31 (.52)	—	14.30 (.50)	—
Treatment	5.63 (1.15)	<.01	5.62 (.93)	<.01	4.95 (.79)	<.01	4.97 (.84)	<.01
ELL	−2.10 (1.03)	.07	−2.21 (.73)	<.01	−1.99 (.80)	.01	−1.98 (.81)	.01
ELL × Treatment	−.95 (1.43)	.52	−.88 (1.11)	.43	−.10 (1.09)	.93	−.10 (1.09)	.93
Pretest	.50 (.16)	<.01	.48 (.13)	<.01	.59 (.08)	<.01	.59 (.09)	<.01
Pretest × Treatment	.33 (.16)	.04	.32 (.13)	.01	.27 (.13)	.04	.25 (.12)	.04
Intercept variance	.90	—	—	—	.34	—	—	—
Residual variance	12.63	—	12.35	—	12.51	—	12.96	—
R^2	—	—	.67	—	—	—	.61	—

Note. GEE = generalized estimating equations; HLM = hierarchical linear model; CR-*SE* = cluster-robust standard errors; ELL = English language learner; IPW = inverse propensity weighted; pretest was cluster-mean centered.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 7
Comparison of Estimates for Religion Data With HLM With Three Different Estimation Methods

Effect	HLM-PQL		HLM-AGQ-10		GEE	
	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value	Estimate (*SE*)	*p*-value
Intercept	−1.74 (.171)	<.001	−1.81 (.163)	<.001	−1.46 (.148)	<.001
Female	.49 (.015)	<.001	.46 (.044)	<.001	.39 (.040)	<.001
Gini (mean centered)	.04 (.020)	.046	.04 (.018)	.048	.04 (.016)	.026
Female × Gini (mean centered)	<.01 (.002)	.035	.01 (.005)	.169	−.00 (.004)	.745
College enrollment (mean centered)	−.02 (.011)	.124	−.02 (.011)	.075	−.01 (.012)	.372
Urbanization (mean centered)	−.01 (.011)	.440	−.01 (.011)	.478	−.02 (.009)	.031
Years of education (mean centered)	−.01 (.004)	.009	−.02 (.005)	<.001	−.05 (.022)	.023
Female × Yrs. of Education (mean centered)	−.05 (.006)	<.001	−.04 (.006)	<.001	−.06 (.015)	<.001
Income	−.08 (.008)	<.001	−.09 (.018)	<.001	−.07 (.025)	.004
Married	Reference		Reference		Reference	
Single	−.28 (.019)	<.001	−.28 (.019)	<.001	−.08 (.075)	.270
Divorced	−.52 (.036)	<.001	−.53 (.036)	<.001	−.63 (.080)	<.001
Widowed	.48 (.026)	<.001	.46 (.027)	<.001	.29 (.081)	.001
Intercept variance	1.687	—	1.493	—	—	—
Female variance	.092	—	.090	—	—	—
Income variance	.014	—	.014	—	—	—

Note. HLM = hierarchical linear model; PQL = penalized quasi-likelihood; AGQ = Adaptive Gaussian Quadrature; GEE = generalized estimating equations.

be incongruent because they are representative of different quantities. That is, the GEE coefficient estimates are interpreted as the change in the log-odds of attending religious services holding all other predictors constant, yielding a population-averaged interpretation. On the other hand, the estimates for the HLM models are interpreted as the change in the log-odds of attending religious services holding all other predictors *and the random effects* constant. This results in a cluster-specific interpretation. With the exception of the urbanization and single predictors, inferences based on *p* values aligned between HLM and GEE even though the values of the coefficients were much different due to the different interpretations of the coefficients. Additionally, it can be seen in Table 7 that the estimates between the different HLM estimation methods do not produce congruent results. The Female × Gini cross-level interaction is marginally significant with PQL ($p = .035$) but is not significant with AGQ ($p = .169$). Even under this near ideal case of an extremely large sample size, PQL and AGQ are noticeably different. The Laplace Approximation was comparable with AGQ for these data; however, this is not always the expectation. As noted previously, the Laplace Approximation tends to be less accurate with smaller sample sizes at either level, neither of which were a concern with these data.

Despite the difference in the interpretation, there are heuristic approximations that can be used to convert cluster-specific coefficients to population averaged coefficients. Molenberghs and Verbeke (2004) provided one for a random intercepts logistic model and Mroz and Zayats (2008) discussed how this can be done for logistic model with multiple random effects (pp. 409–410). From Mroz and Zayats (2008), to obtain population-averaged estimates from cluster-specific estimates, multiply the cluster-specific inferences by $\sqrt{\frac{\pi^2/3}{\pi^2/3 + \sum_{q=0}^{Q} g_{qq}}}$ − the variance of the logistic function divided by the logistic variance plus the sum of the random effects variances. In the religion data, this quantity is equal to 0.82 and multiplying the AGQ HLM coefficients by 0.82 yields the GEE coefficients in almost every case (except for the marital status predictors, which might suggest a possible misspecification; the model did not include which religion individuals practice, e.g., which would have a clear effect on whether people attend religious services).

Discussion

Although HLM has historically been closely associated to clustered data in psychology, population-averaged methods (PAMs) can also be used to analyze clustered data without the additional step of explicitly modeling the random effects or covariance structures which also allows researchers to bypass the assumptions inherent with random effects that are required in HLM but not in PAMs. Researchers are encouraged to determine if their research question and interests truly call for cluster-specific inferences or a substantive reason to partition the residual variance between levels. If the clustering within the data is seen as a facet that merely needs to be accommodated to yield appropriate estimates or is a byproduct of the data collection and one simply wants to ensure that the estimates adequately account for the clustered structure, PAMs offer researchers an alternative to HLM that similarly performs inferential tests on regression coefficients without having to partake in HLM-specific model building steps such as fitting the random effects and the covariance structures for the random effects under the assumption that these structures have been properly specified. More pragmatically, if researchers are fitting a random intercepts model to "account for clustering," this can more simply (and sometimes more appropriately) be done with PAMs.

The use of CR-*SE*, in particular, offers researchers far simpler computations of familiar and often desirable metrics like R^2 or effect sizes with continuous outcomes. That is, the statistical correction is made to the standard errors of regression coefficients but quantities involved in R^2 and effect size calculations such as the sums of squares are identical to a single-level model because the model contains only fixed effects. Objectively speaking, HLM can still model clustered data even if the clustering is more of a

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

nuisance rather than a substantive interest, particularly with continuous outcomes. However, researchers subscribing to this philosophy are making many more assumptions of their data and trusting that they are modeling all covariance structures correctly, both of which can be obviated with PAMs. Furthermore, the limited amount of space for research articles often means that discussion of and inspection of the wealth of HLM modeling assumptions is not included in published studies (Dedrick et al., 2009). This can make it difficult to assess research from a methodological perspective because model estimates may not be trustworthy in the event that one of the many assumptions is potentially violated.

Based on the syllabus review conducted, the strong preference for HLM in psychology (and behavioral sciences in general) appears to be at least partially driven by tradition as reflected by what is taught to graduate students in the classroom. Instructors of multilevel, longitudinal, and correlated data analysis courses in the behavioral sciences are especially encouraged to integrate some discussion of PAMs into their courses. This article is not suggesting that PAMs are superior to HLM, that HLM is an inferior method for modeling clustered data, or that behavioral science should strive to emulate biomedical or econometric applications where PAMs are much more common. Rather, we are trying to argue that nonmethodological researchers often strive for the simplest analysis that can adequately handle their data structure and the methodological literature has even gone so far to note an aversion to more advanced methodology in empirical studies (Sharpe, 2013). Although HLM is undoubtedly useful in many contexts and affords many advantages that can address a broader range of research questions, researchers who instinctively model clustered data with HLM are making fairly rigid assumptions of their data in order to obtain estimates that are not always of interest to their research question. As an exception to the methodological norm discussed in Sharpe (2013), researchers who resort to HLM as an almost knee-jerk reaction to clustered data are often using a *more* complicated and strict method that makes *more* assumptions than may be necessitated by their data and cluster-specific estimates may be more difficult to interpret or perhaps even misaligned with researchers' expectations if the outcome is discrete.

In fact, it is commonplace for psychology articles to report using HLM without reporting the variance component estimates or other vital pieces of information that provide a rationale for using HLM over PAMs (Dedrick et al., 2009). To exemplify this point, using only the results on the first page of a Google Scholar search for "HLM," since 2010, in a single arbitrarily selected flagship psychology journal (*Psychological Science*), five of the 10 studies on the first page of results at the time of this writing used straightforward, conditionally univariate HLM models and did not report nor discuss the random effects, partitioning of the variance between levels, or covariance structure(s) that were used for the random effects or the residuals (Gebauer, Sedikides, & Neberich, 2012; Job, Dweck, & Walton, 2010; Sherman, Haidt, & Clore, 2012; Silberzahn & Uhlmann, 2013; Ziegler et al., 2010) and one study discussed these model components in the statistical analysis section but did not report them in the results (Lopez, Hofmann, Wagner, Kelley, & Heatherton, 2014). Furthermore, two of these studies had discrete outcomes (Job et al., 2010; Lopez et al., 2014) meaning that a cluster-specific model was fit without any mention of the modeling components that make the model cluster-specific, a meaningful decision that affects the magnitude and interpretation of the coefficients.

We are not claiming that these studies were done incorrectly, that their conclusions are faulty, or that there are any methodological deficiencies in *Psychological Science* nor are we trying to criticize the reporting practices contained with these studies as we can certainly appreciate the strict length restrictions encountered in substantive journals. Rather, we suspect that this information was not reported because it was not a concern to the research question and we are trying to emphasize that random effects and cluster-specific inference are at the core of HLM, not simply estimating standard errors that account for clustering. When the latter is the interest, PAMs allow for simplified model interpretations, avoid random effect distributional assumptions, are more straightforward to estimate with discrete outcomes, and are more straightforward to implement in the context of these papers which seemed to be solely interested in estimating and interpreting the regression coefficients while accounting for clustering.[11]

Broader Implications for the Field

Although the ubiquity of HLM may seem fairly harmless, the implications for current and future research are not innocuous. From a pedagogical perspective, teaching HLM as the default method to handle clustered observations can make statistics more unapproachable to substantive researchers than it is already perceived to be and can unnecessarily complicate otherwise straightforward analyses. Although HLM is appropriate in certain circumstances, it is one of the more complex methods to handle clustering and models are far less intuitive, more difficult to build, and are more difficult to interpret than PAMs. Although the statistical theory behind PAMs is likely not any more approachable than HLM, PAMs provide a more intuitive conceptual transition from independent data to clustered data and offer substantive researchers tools to address a wealth of research questions that arise in psychological research that may align more closely with their research interests. More plainly, material in clustered data courses in psychology, anecdotally, tends to focus more so on what HLM is rather than contexts when it should be used.

We hope that this article has provided some context for why PAMs should be more relevant to psychologists and to researchers in the behavioral sciences broadly. As we hope to have shown in this article, PAMs can play an integral role in analyzing clustered data in and future research on their properties and performance in behavioral science-specific contexts could be quite helpful.

[11] As one anonymous reviewer pointed out, it would be fairer to also provide examples of similar transgressions with PAMs. However, this elicits one of our primary motivations for this article—one cannot easily obtain examples of PAMs used in psychological research and a similar random review would require us to inspect biomedical or economics journals, which are outside the scope of *Psychological Methods.*

References

Agerbo, E., Sterne, J. A., & Gunnell, D. J. (2007). Combining individual and ecological data to determine compositional and contextual socioeconomic risk factors for suicide. *Social Science & Medicine, 64,* 451–461. http://dx.doi.org/10.1016/j.socscimed.2006.08.043

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis, 47,* 639–653. http://dx.doi.org/10.1016/j.csda.2003.12.009

Allison, P. D. (2012, April). Handling missing data by maximum likelihood. Paper presented at the SAS Global Forum, Orlando, FL. Retrieved from http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

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

Bates, D. (2009). *Assessing the precision of estimates of variance components*. Seewiesen, Austria: Presentation at the Max Planck Institute for Ornithology.

Bates, M. D., Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Handling correlations between covariates and random slopes in multilevel models. *Journal of Educational and Behavioral Statistics, 39,* 524–549.

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

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,* 2591–2602. http://dx.doi.org/10.1002/sim.1524

Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics, 29,* 201–218. http://dx.doi.org/10.3102/10769986029002201

Beunckens, C., Sotto, C., & Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics & Data Analysis, 52,* 1533–1548. http://dx.doi.org/10.1016/j.csda.2007.04.020

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88,* 9–25.

Burton, P., Gurrin, L., & Sly, P. (1998). Tutorial in biostatistics. Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine, 17,* 1261–1291. http://dx.doi.org/10.1002/(SICI)1097-0258(19980615)17:11<1261::AID-SIM846>3.0.CO;2-Z

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics, 29,* 238–249. http://dx.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,* 317–372.

Carlin, J. B., Wolfe, R., Brown, C. H., & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics, 2,* 397–416. http://dx.doi.org/10.1093/biostatistics/2.4.397

Carpenter, J. R., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society Series A, 169,* 571–584. http://dx.doi.org/10.1111/j.1467-985X.2006.00407.x

Chen, B., Yi, G., & Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association, 105,* 336–353. http://dx.doi.org/10.1198/jasa.2010.tm08551

Clarkson, D. B., & Zhan, Y. (2002). Using spherical-radial quadrature to fit generalized linear mixed effects models. *Journal of Computational and Graphical Statistics, 11,* 639–659. http://dx.doi.org/10.1198/106186002439

Clayton, D., Spiegelhalter, D., Dunn, G., & Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society Series B. Methodological, 60,* 71–87. http://dx.doi.org/10.1111/1467-9868.00109

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

Diaz, R. E. (2007). Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomized trials. *Computational Statistics & Data Analysis, 51,* 2871–2888. http://dx.doi.org/10.1016/j.csda.2006.10.005

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

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

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37,* 379–403. http://dx.doi.org/10.1207/S15327906MBR3703_4

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics, 51,* 309–317. http://dx.doi.org/10.2307/2533336

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science, 8,* 284–299. http://dx.doi.org/10.1214/ss/1177010899

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

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

Gebauer, J. E., Sedikides, C., & Neberich, W. (2012). Religiosity, social self-esteem, and psychological adjustment: On the cross-cultural specificity of the psychological benefits of religiosity. *Psychological Science, 23,* 158–160. http://dx.doi.org/10.1177/0956797611427045

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics, 29,* 421–437. http://dx.doi.org/10.3102/10769986029004421

Givens, G. H., & Hoeting, J. A. (2005). *Computational statistics*. New York: John Wiley & Sons.

Gosho, M. (2014). Criteria to select a working correlation structure for the generalized estimating equations method in SAS. *Journal of Statistical Software, 57,* 1–10. http://dx.doi.org/10.18637/jss.v057.c01

Gosho, M., Hamada, C., & Yoshimura, I. (2011). Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. *Communications in Statistics Theory and Methods, 40,* 3839–3856. http://dx.doi.org/10.1080/03610926.2010.501938

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician, 60,* 19–26. http://dx.doi.org/10.1198/000313006X90396

Hanley, J. A., Negassa, A., Edwardes, M. D., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology, 157,* 364–375. http://dx.doi.org/10.1093/aje/kwf215

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15,* 234–249.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39,* 709–722. http://dx.doi.org/10.3758/BF03192961

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman and Hall/CRC. http://dx.doi.org/10.1201/9781420080674

Hin, L. Y., Carey, V. J., & Wang, Y. G. (2007). Criteria for working–correlation structure selection in GEE. *The American Statistician, 61,* 360–364. http://dx.doi.org/10.1198/000313007X245122

Hin, L. Y., & Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine, 28,* 642–658.

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

Horton, N. J., & Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research, 8,* 37–50. http://dx.doi.org/10.1191/096228099673120862

Horton, N. J., & Laird, N. M. (2001). Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics, 57,* 34–42. http://dx.doi.org/10.1111/j.0006-341X.2001.00034.x

Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician, 53,* 160–169.

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). Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-3-642-72087-1_17

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., . . . 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,* 467–474. http://dx.doi.org/10.1097/EDE.0b013e3181caeb90

Huber, P. J. (1967, June). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1,* 221–233.

Huynh, M., Parker, J. D., Harper, S., Pamuk, E., & Schoendorf, K. C. (2005). Contextual effect of income inequality on birth outcomes. *International Journal of Epidemiology, 34,* 888–895. http://dx.doi.org/10.1093/ije/dyi092

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,* 5142–5154. http://dx.doi.org/10.1016/j.csda.2006.05.021

Job, V., Dweck, C. S., & Walton, G. M. (2010). Ego depletion—Is it all in your head? Implicit theories about willpower affect self-regulation. *Psychological Science, 21,* 1686–1693. http://dx.doi.org/10.1177/0956797610384745

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis, 52,* 5066–5074. http://dx.doi.org/10.1016/j.csda.2008.05.002

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13,* 236–247. http://dx.doi.org/10.1214/ss/1028905886

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics Simulation and Computation, 27,* 591–604. http://dx.doi.org/10.1080/03610919808813497

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

Kim, Y., Choi, Y. K., & Emery, S. (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician, 67,* 171–182. http://dx.doi.org/10.1080/00031305.2013.817357

Kontos, S., Burchinal, M., Howes, C., Wisseh, S., & Galinsky, E. (2002). An eco-behavioral approach to examining the contextual effects of early childhood classrooms. *Early Childhood Research Quarterly, 17,* 239–258. http://dx.doi.org/10.1016/S0885-2006(02)00147-3

Kovacevic, M. S., & Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics Theory and Methods, 32,* 103–121. http://dx.doi.org/10.1081/STA-120017802

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript, California State University, Los Angeles, CA.

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

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

LeBeau, B. (2013). *Misspecification of the covariance matrix in the linear mixed model: A Monte Carlo simulation* (Doctoral dissertation). University of Minnesota, St. Paul, MN.

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

Lin, T. I., & Lee, J. C. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medicine, 27,* 1490–1507.

Lipsitz, S. R., Ibrahim, J. G., & Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association, 94,* 1147–1160. http://dx.doi.org/10.1080/01621459.1999.10473870

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

Liu, L., & Yu, Z. (2008). A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine, 27,* 3105–3124.

Lopez, R. B., Hofmann, W., Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2014). Neural predictors of giving in to temptation in daily life. *Psychological Science, 25,* 1337–1344.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I., & Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics, 63,* 935–941.

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower-and higher-level classes in multilevel latent class analysis. *Sociological Methodology, 40,* 247–283. http://dx.doi.org/10.1111/j.1467-9531.2010.01231.x

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,* 427–440. http://dx.doi.org/10.1016/j.csda.2003.08.006

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

Marschall, M. J. (2004). Citizen participation and the neighborhood context: A new look at the coproduction of local public goods. *Political Research Quarterly, 57,* 231–244. http://dx.doi.org/10.1177/106591290405700205

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London, UK: Chapman and Hall. http://dx.doi.org/10.1007/978-1-4899-3242-6

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York, NY: Wiley.

McNeish, D. M. (2014). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological Methods, 19,* 552–563. http://dx.doi.org/10.1037/met0000024

McNeish, D. M., & Harring, J. R. (2015). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics. Simulation and Computation*. Advance online publication. http://dx.doi.org/10.1080/03610918.2014.983648

McNeish, D. M., & Stapleton, L. M. (2014). The effect of small sample size on two level model estimates: A review and illustration. *Educational Psychology Review*. Advance online publication. http://dx.doi.org/10.1007/s10648-014-9287-x

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

Molenberghs, G., & Verbeke, G. (2004). Meaningful statistical model formulations for repeated measures. *Statistica Sinica, 14,* 989–1020.

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician, 61,* 22–27. http://dx.doi.org/10.1198/000313007X171322

Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal Biometrische Zeitschrift, 45,* 395–409. http://dx.doi.org/10.1002/bimj.200390021

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics, 54,* 1560–1568. http://dx.doi.org/10.2307/2533680

Mroz, T. A., & Zayats, Y. V. (2008). Arbitrarily normalized coefficients, information sets, and false reports of "biases" in binary outcome models. *The Review of Economics and Statistics, 90,* 406–413. http://dx.doi.org/10.1162/rest.90.3.406

Muthen, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

Newton, S. D., & McCoach, D. B. (2015, April). *Growth modeling: Akaike and Bayesian Information Criteria and sample size considerations*. Paper presented at the annual meeting of the American Educational Research Association (AERA), SIG: Multilevel Modeling, Chicago, IL.

Nichols, A., and Schafer, M. (2007, July). Clustered standard errors in Stata, United Kingdom Stata Users' Group Meetings.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics, 57,* 120–125. http://dx.doi.org/10.1111/j.0006-341X.2001.00120.x

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

Petronis, K. R., & Anthony, J. C. (2003). A different kind of contextual effect: Geographical clustering of cocaine incidence in the USA. *Journal of Epidemiology and Community Health, 57,* 893–900. http://dx.doi.org/10.1136/jech.57.11.893

Pfefferman, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel model. *Journal of Royal Statistical Society, 60,* 23–40.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4,* 12–35.

Pinheiro, J. C., & Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics, 15,* 58–81. http://dx.doi.org/10.1198/106186006X96962

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, 169,* 805–827. http://dx.doi.org/10.1111/j.1467-985X.2006.00426.x

Rao, J. N. K., Verret, F., & Hidiroglou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology, 39,* 263–282.

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

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9,* 141–157.

Recchia, A. (2010). R-squared measures for two-level hierarchical linear models using SAS. *Journal of Statistical Software, 32,* 1–9. http://dx.doi.org/10.18637/jss.v032.c02

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90,* 106–121. http://dx.doi.org/10.1080/01621459.1995.10476493

Rotnitzky, A., & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika, 77,* 485–497. http://dx.doi.org/10.1093/biomet/77.3.485

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. http://dx.doi.org/10.1002/9780470316696

Ruiter, S., & Van Tubergen, F. (2009). Religious attendance in cross-national perspective: A multilevel analysis of 60 countries. *American Journal of Sociology, 115,* 863–895. http://dx.doi.org/10.1086/603536

Savalei, V. (2010). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods, 15,* 352–367. http://dx.doi.org/10.1037/a0020143

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13,* 150–170. http://dx.doi.org/10.1037/1082-989X.13.2.150

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association, 94,* 1096–1120. http://dx.doi.org/10.1080/01621459.1999.10473862

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods, 18,* 572–582. http://dx.doi.org/10.1037/a0034177

Sherman, G. D., Haidt, J., & Clore, G. L. (2012). The faintest speck of dirt: Disgust enhances the detection of impurity. *Psychological Science, 23,* 1506–1514. http://dx.doi.org/10.1177/0956797612445318

Shults, J., Ratcliffe, S. J., & Leonard, M. (2007). Improved generalized estimating equation analysis via xtqls for quasi-least squares in Stata. *The Stata Journal, 7,* 147–166.

Silberzahn, R., & Uhlmann, E. L. (2013). It pays to be Herr Kaiser: Germans with noble sounding surnames more often work as managers than as employees. *Psychological Science*. Advance online publication.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22,* 342–363. http://dx.doi.org/10.1177/0049124194022003004

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London, UK: Sage. http://dx.doi.org/10.1007/978-3-642-04898-2_387

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

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Twisk, J. W. (2004). Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology, 19,* 769–776. http://dx.doi.org/10.1023/B:EJEP.0000036572.00663.f2

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research, 23,* 42–59. http://dx.doi.org/10.1177/0962280212445834

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,* 541–556. http://dx.doi.org/10.1016/S0167-9473(96)00047-3

Verbeke, G., & Molenberghs, G. (2007). What can go wrong with the score test? *The American Statistician, 61,* 289–290. http://dx.doi.org/10.1198/000313007X243089

Westgate, P. M. (2013). A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. *Statistics in Medicine, 32,* 2850–2858. http://dx.doi.org/10.1002/sim.5709

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

White, H. (1984). *Asymptotic theory for econometricians*. San Diego, CA: Academic Press.

Whittaker, T. A., & Furlow, C. F. (2009). The comparison of model selection criteria when selecting among competing hierarchical linear models. *Journal of Modern Applied Statistical Methods, 8,* 173–193.

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

Wooldridge, J. (2003). *Introductory econometrics: A modern approach*. Boston, MA: Cengage.

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials, 33,* 869–880. http://dx.doi.org/10.1016/j.cct.2012.05.004

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42,* 121–130. http://dx.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,* 1049–1060. http://dx.doi.org/10.2307/2531734

Zhou, X.-H., Perkins, A. J., & Hui, S. L. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician, 53,* 282–290.

Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., . . . Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science, 21,* 551–559. http://dx.doi.org/10.1177/0956797610363406

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science, 45,* 470–490. http://dx.doi.org/10.2307/2669353

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix A

Technical Details for Each Method

Cluster-Robust Standard Errors

We will outline how CR-*SE*s account for clustering for continuous outcomes using OLS because CR-*SE*s are most advantageous for continuous outcomes but readers should note that CR-*SE*s can similarly be applied to models estimated with maximum likelihood with only minor changes in the computation.

As noted in the main text, $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ for $\mathbf{Y}$ where $\boldsymbol{\varepsilon} \sim N^{i.i.d.}(0, \sigma^2)$. The standard errors of the regression coefficients are taken from the square root of the diagonal elements of variance of $Var(\hat{\boldsymbol{\beta}})$ which is most generally calculated by $Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$ (e.g., Wooldridge, 2003). Assuming independently and identically distributed residuals, $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ can be summarized by the average squared residuals $\sigma^2 = (n - p)^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ which results in a diagonal matrix $\sigma^2\mathbf{I}$ where σ^2 is a single estimate of the residual variance. Through the assumption of independently and identically distributed residuals, the estimate of $Var(\hat{\boldsymbol{\beta}})$ then simplifies to $(\sigma^2)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$, and the first multiplication simplifies because $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}$, leaving $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ if the assumptions are upheld. However, when the assumption of independently and identically distributed residuals is violated, then summarizing $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ with $(n - p)^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ is not appropriate and results in estimates of $Var(\hat{\boldsymbol{\beta}})$ being too small because covariances between observations are constrained to zero, meaning that the standard errors are underestimated because terms from the most general formula should not cancel, and, consequently, Type I errors are inflated (Cameron & Miller, 2015).

When data are dependent through clustering, the residuals of observations within clusters are likely related meaning that the assumption of independently and identically distributed residuals is unlikely to be upheld (e.g., Raudenbush & Bryk, 2002). Robust standard errors (perhaps more appropriately called heteroskedasdicity corrected covariance or empirical estimators) address this problem by replacing the average of squared residuals $(n - p)^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ with the squared residual $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ which does not require diagonal elements to be identical. After this substitution, $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ can no longer be summarized by $\sigma^2\mathbf{I}$ and the variance of the regression coefficients no longer simplifies and thus reverts to its original formulation as $Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$(Huber, 1967; White, 1980).

This substitution only addresses violations to the residuals being identically distributed; to address violations of the residuals being independently distributed, $\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}$ must be calculated for each cluster (rather than each individual) and then summed across all clusters, $\sum_{j=1}^{J}\mathbf{X}_j^T\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^T\mathbf{X}_j$. This quantity is then pre and post multiplied by $(\mathbf{X}^T\mathbf{X})^{-1}$ to obtain the standard errors that account for clustering (i.e., cluster-robust standard errors) such that $Var^{CR}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sum_{j=1}^{J}(\mathbf{X}_j^T\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^T\mathbf{X}_j)(\mathbf{X}^T\mathbf{X})^{-1}$ (White, 1984). This calculation is robust to dependence within clusters but still requires that observations between clusters be independent.

With discrete outcomes, the process is similar except that the calculation of the residuals differs. With continuous outcomes, the residuals are straightforwardly calculated by the difference between the observed and the predicted values, $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. With discrete outcomes, the relation differs because the conditional mean of the outcome variable distribution must be nonlinearly related to predictor variables through a nonlinear link function, $E(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X\beta})$ where $g(\cdot)$ is the nonlinear link function (e.g., a logistic function for binary outcomes; more detail is provided in the GEE section that follows). Due to the nonlinear link function, deviance residuals are often used instead of the traditional raw residuals. Deviance residuals measure the contribution of each unit to the model deviance ($-2 \times$ log-likelihood) and are calculated by $e_i^D = \begin{cases} \sqrt{2\{\log(1 + e^{\mathbf{X}_i\hat{\boldsymbol{\beta}}}) - \mathbf{X}_i\hat{\boldsymbol{\beta}}\}}, & \text{if } y_i = 1 \\ -\sqrt{2\{\log(1 + e^{\mathbf{X}_i\hat{\boldsymbol{\beta}}})\}}, & \text{if } y_i = 0 \end{cases}$ for logistic models. $\boldsymbol{\varepsilon}^D$ would then replace $\boldsymbol{\varepsilon}$ in the formulas presented earlier in this section.

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Generalized Estimating Equations (GEE)

GEE is an algorithmic method to estimate generalized linear models which, as briefly alluded to in the previous section, relate the conditional mean of an outcome variable distribution $E(\mathbf{Y}_j|\mathbf{X}_j) = \boldsymbol{\mu}_j$ to a linear predictor $\mathbf{X}\boldsymbol{\beta}$ through a link function $g(\cdot)$ (McCullagh & Nelder, 1989; McCulloch & Searle, 2001). In behavioral sciences, common link functions are the identity function for normally distributed outcomes, $g(\boldsymbol{\mu}_j) = \boldsymbol{\mu}_j$, the logit link for binary outcomes, $g(\boldsymbol{\mu}_j) = \log(\boldsymbol{\mu}_j/(1 - \boldsymbol{\mu}_j))$, or the log link for count outcomes, $g(\boldsymbol{\mu}_j) = \log(\boldsymbol{\mu}_j)$. The variance of $\mathbf{Y}_j$ is then specified as $Var(\mathbf{Y}_j) = \nu(\boldsymbol{\mu}_j)\phi$ where ϕ is a possibly unknown scale parameter ($\phi = 1$ for binary and Poisson responses) and $\nu(\boldsymbol{\mu}_j)$ is a known variance function ($\nu(\boldsymbol{\mu}_j) = 1$ for normally distributed outcomes, $\boldsymbol{\mu}_j(1 - \boldsymbol{\mu}_j)$ for binary outcomes, and $\boldsymbol{\mu}_j$ for Poisson distributed outcomes).

Broadly speaking, estimating equations specify how parameters in a model are estimated with salient examples including ordinary least squares and maximum likelihood. When data are independent (i.e., clustering is not meaningful), the maximum likelihood estimate of the vector of regression coefficients $\boldsymbol{\beta}$ in a generalized linear model can be obtained using independence estimating equations. Where $\hat{\boldsymbol{\beta}}$ is estimated with score equations such that $\sum_{j=1}^{J}(\mathbf{X}_j^{\mathrm{T}}\mathbf{A}_j\mathbf{S}_j) = \mathbf{0}$ where $\mathbf{X}_j$ is an $m_j \times p$ design matrix for the jth cluster, $\mathbf{A}_j = diag[(\mathrm{Var}(\mu_{j1}), \ldots, \mathrm{Var}(\mu_{jm_j})]$ for m_j the number of within-cluster units in cluster j, and $\mathbf{S}_j = \mathbf{Y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta})$ for $\mathbf{Y}_j$ is an $m_j \times 1$ vector of outcomes for the jth cluster and $\boldsymbol{\mu}_j(\boldsymbol{\beta})$ the conditional mean of the outcome which is based up the regression coefficients (see, e.g., Fitzmaurice, 1995; Liang & Zeger, 1986). As seen by the diagonal structure of $\mathbf{A}_j$, this assumes that covariance is directly calculable from the model and observations within clusters are not related, which introduces bias into the standard errors estimates of the regression coefficients if data are meaningfully clustered. As in HLM, this issue can be addressed by directly modeling the source of clustering. However, Liang and Zeger (1986) generalized independence estimating equation (hence the name "generalized estimating equations") to handle situations in which specifying a modeling for the correlation of observations is not desired. Rather, the covariance matrix is updated as a function of unknown parameters.

Liang and Zeger (1986) define generalized estimating equations for regression coefficients $\hat{\boldsymbol{\beta}}$ such that $\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{S}_j = \mathbf{0}$ where $\mathbf{D}_j = \mathbf{X}_j^{\mathrm{T}}\mathbf{A}_j = \frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\beta}}$ and $\mathrm{V}_j = \hat{\boldsymbol{\phi}}\mathbf{A}_j^{1/2}\mathbf{K}_j(\alpha)\mathbf{A}_j^{1/2}$ for $\hat{\phi}$ a scale parameter estimated by $\hat{\phi} = \frac{1}{N-p}\sum_{j=1}^{J}\sum_{i=1}^{m_j} e_{ij}^2$, and $\mathbf{K}_j$ is an $m_j \times m_j$ working correlation matrix comprised of unknown parameters $\boldsymbol{\alpha}$ that estimate the correlation of observations within clusters rather than it being explicitly modeled. The structure of Kj is specified by the researcher a priori but its elements are updated algorithmically. For cross-sectionally clustered data, an exchangeable structure is typically suitable[12] where $Corr(\mathrm{Y}_{ij},\mathrm{Y}_{kj}) = \begin{cases} 1 & i=k \\ \alpha & i \neq k \end{cases}$ meaning that an arbitrary within-cluster observation has equal correlation with all other observations within the same cluster. The value of α is conceptually similar to the traditional ICC as calculated with HLM in an unconditional model (Wu, Crespi, & Wong, 2012).

As mentioned previously, GEE iteratively updates the parameters in the working structure, α. First, $\hat{\boldsymbol{\beta}}$ is estimated assuming independence. Then, $\mathbf{K}_j(\boldsymbol{\alpha})$ is estimated from the errors from the model that assumes independence. The estimation of $\mathbf{K}_j(\boldsymbol{\alpha})$ depends of the working structure specified by the researcher. For an exchangeable structure that is typical with cross-sectional clustering (Horton & Lipsitz, 1999), $\hat{\alpha} = \frac{1}{\hat{\phi}(N^* - p)}\sum_{j=1}^{J}\sum_{i<k}\mathrm{e}_{ij}\mathrm{e}_{ik}$ where $N^* = 0.5\sum_{j=1}^{J}\mathrm{m}_j(\mathrm{m}_j - 1)$. Because GEE does not require the full likelihood, with discrete outcomes, deviance residuals cannot be used so Pearson residuals (e^p) are used instead where $e_{ij}^P = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{Var(\hat{\mu}_{ij})}}$. Once a value(s) for $\hat{\alpha}$ is obtained, then $\mathbf{V}_j$ can be calculated by $\mathbf{V}_j = \hat{\phi}\mathbf{A}_j^{1/2}\mathbf{K}_j(\boldsymbol{\alpha})\mathbf{A}_j^{1/2}$. $\hat{\boldsymbol{\beta}}$ is then updated by $\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + (\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{D}_j)^{-1}(\sum_{j=1}^{J}\mathbf{D}_j^{\mathrm{T}}\mathbf{V}_j^{-1}\mathbf{S}_j)$ where r is the index for the iteration. When $r = 1$, $\hat{\boldsymbol{\beta}}_{\mathbf{j}}$ houses the coefficient estimates under the independence assumption. For readers within a greater interest in longitudinal data, an AR(1) autoregressive structure is a common choice for the working correlation structure such that $Corr(\mathrm{Y}_{ij}, \mathrm{Y}_{i+t,j}) = \alpha^t$ for $t = 0, 1, 2, \ldots, m_j - i$ for the I the total number of repeated measures. With an AR(1) working structure, $\hat{\alpha} = \frac{1}{\hat{\phi}(J_1 - p)}\sum_{j=1}^{J}\sum_{i \leq m_j}\mathrm{e}_{ij}\mathrm{e}_{i+1,j}$ where $J_1 = \sum_{j=1}^{J}(m_j - 1)$.

[12] Ballinger (2004) states that "[when] there is no logical ordering for observations within a cluster (such as when data are clustered within subject or within an organizational unit but not necessarily collected over time), an exchangeable correlation structure should be used" (p. 133).

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Once the iterative process has successfully converged (assuming convergence can be reached; convergence may be more difficult to obtain if the clusters are very unbalanced or if the working structure is grossly incorrect such that the resulting estimating form a nonpositive definite matrix; Shults, Ratcliffe, & Leonard, 2007), *Var* ($\hat{\boldsymbol{\beta}}$) is calculated using a cluster-robust method similar to the one outlined above in the CR-*SE* section. The naïve estimator of *Var* ($\hat{\boldsymbol{\beta}}$) that ignores clustering is calculated by $(\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j)^{-1}$(e.g., McCullagh & Nelder, 1989). Similar to CR-*SE*s, the naïve estimator "sandwiches" a quantity that takes the clustering into the account. In GEE, the middle term is formulated by $\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j$ making the GEE estimate of $Var(\hat{\boldsymbol{\beta}})$ equal to $Var^{GEE}(\hat{\boldsymbol{\beta}}) = (\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j)^{-1} (\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{S}_j \mathbf{S}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j) (\sum_{j=1}^{J} \mathbf{D}_j^{\mathrm{T}} \mathbf{V}_j^{-1} \mathbf{D}_j)^{-1}$.

Appendix B

Software Code for Word Knowledge Example

HLM

SAS Proc Mixed

```
proc mixed data=Reading;
model post=Pre|Trt ELL|Trt/solution;
random int/ sub=Teacher; run;
```

Stata xtmixed

```
xtmixed Post Pre Trt Trt_Pre Trt_ELL ||Teacher:,
reml cov(unstruct)
```

M*plus*

```
VARIABLE:
cluster = Teacher;
within = Pre Trt_Pre Trt_ELL; between = Trt;

ANALYSIS:
Estimator=ML; type=Twolevel Random;

MODEL:
%within%
Post ON Pre Trt_Pre Trt_ELL;

%between%
Post ON Trt; Post;
```

GEE

SAS Proc Genmod

```
proc genmod data=Reading;
class teacher;
model Post=Pre|Trt ELL|Trt;
repeated subject=Teacher/ type=exch;run;
```

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Stata xtgee

```
xtgee Post Pre Trt Trt_Pre Trt_ELL,
fam(gaus)
link(iden)
i(Teacher)
corr(exc)
```

M*plus*- not available (http://www.statmodel.com/discussion/messages/11/635.html?1114217498)

CR-*SE*

SAS Proc Glimmix

```
proc glimmix data=Reading empirical;
class teacher;
model Post= Pre|Trt ELL|Trt/ solution;
random _residual_ /sub=teacher; run;
```

Stata regress

```
regress Post Pre Trt Trt_Pre Trt_ELL, cluster(Teacher)
```

M*plus*

```
VARIABLE:
cluster = teacher;

ANALYSIS:
estimator=MLR; type= complex;

MODEL:
Post ON Post Pre Trt Trt_Pre Trt_ELL;
```

Appendix C

Software Code for Religion Example

HLM—Adaptive Gaussian Quadrature, 10 Points

SAS Proc Glimmix

```
proc glimmix data=Religion method=quad(qpoints=10);
class country;
model ReligiousAttendance= Female|GINI College Urban Educ|Female Income
Single Divorced Widowed/solution link=logit dist=b;
random int female income/ subject=country;run;
```

Stata xtmixed

```
xtmelogit ReligiousAttendance Female GINI Female*GINI College Urban Educ Educ*Female
Income Single Divorced Widowed || Country: Female Income, intpoints(10)
```

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

M*plus*

```
VARIABLE:
categorical = ReligiousAttendance; cluster = Country;
within = Female Income Single Divorce Widowed Educ Fe_FINI F_Educ;
between = GINI College Urban;

ANALYSIS:
Estimator=ML;
type=Twolevel Random; Algorithm=Integration; Integration=Standard (10);

MODEL:
%within%
ReligiousAttendance ON Single Divorce Widowed Educ Fe_FINI F_Educ;
b1 | ReligiousAttendance ON Female; b2 | ReligiousAttendance ON Income;

%between%
ReligiousAttendance ON GINI College Urban;
b1;b2; ReligiousAttendance;
```

HLM—Laplace

SAS Proc Glimmix

```
proc glimmix data=Religion method=laplace;
class country;
model ReligiousAttendance= Female|GINI College Urban Educ|Female Income
Single Divorced Widowed/solution link=logit dist=b;
random int female income/subject=country;run;
```

Stata xtmixed

```
xtmelogit ReligiousAttendance Female GINI Female*GINI College Urban Educ Educ*Female
Income Single Divorced Widowed || Country: Female Income, laplace
```

M*plus* — not available (Bauer & Sterba, 2011)

HLM—PQL

SAS Proc Glimmix

```
proc glimmix data=Religion;
class country;
model ReligiousAttendance= Female|GINI College Urban Educ|Female Income
Single Divorced Widowed/solution link=logit dist=b;
random int female income/ subject=country; run;
```

Stata — not available (Kim, Choi, & Emery, 2013 p. 174).
M*plus* — not available (Bauer & Sterba, 2011)

(*Appendices continue*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

GEE

SAS Proc Genmod

```
proc genmod data=Religion descending;
class country;
model ReligiousAttendance= Female|GINI College Urban Educ|Female Income
Single Divorced Widowed/link=logit dist=b;
repeated subject=country/type=ind;run;
```

Stata xtgee

```
xtgee ReligiousAttendance Female GINI Female*GINI College Urban Educ Educ*Female Income
Single Divorced Widowed,
fam(bi)
link(logit)
i(Country)
corr(ind)
```

M*plus* - not available (http://www.statmodel.com/discussion/messages/11/635.html?1114217498)

CR-*SE*

SAS Proc Glimmix

```
proc glimmix data=Religion empirical;
class country;
model ReligiousAttendance= Female|GINI College Urban Educ|Female Income
Single Divorced Widowed /Solution link=logit dist=b;
random _residual_/subject=country; run;
```

Stata logit

```
logit ReligiousAttendance Female GINI Female*GINI College Urban Educ Educ*Female Income
Single Divorced Widowed, cluster(Country)
```

M*plus*

```
VARIABLE:
categorical = ReligiousAttendance; cluster = Country;

ANALYSIS:
estimator=MLR; type= complex;

MODEL:
ReligiousAttendance ON
Female Income GINI Urban College Single Divorce Widowed Educ Fe_FINI F_Educ;
```

Received January 30, 2015
Revision received January 10, 2016
Accepted January 26, 2016 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.