American Psychological Association
Division of School Psychology

School Psychology Quarterly

1045-3830/18/$12.00

2018, Vol. 33, No. 3, 492–499
http://dx.doi.org/10.1037/spq0000272

Multilevel Modeling Myths

Francis L. Huang
University of Missouri

The use of multilevel modeling (MLM) to analyze nested data has grown in popularity over the years in the study of school psychology. However, with the increase in use, several statistical misconceptions about the technique have also proliferated. We discuss some commonly cited myths and golden rules related to the use of MLM, explain their origin, and suggest approaches to dealing with certain issues. Misunderstandings related to the use of the intraclass correlation, design effects, minimum sample size, multilevel factor structures, model R^2, and the misestimation of standard errors are reviewed. Many of the cited myths have much truth in them—though at times, researchers may not be aware of the exceptions to the rules that prevent their overall generalization. Although nesting should be accounted for, researchers should realize that MLM, which is a powerful and flexible technique, is not the only method that can be used to account for the clustering effect.

Impact and Implications

School psychology is inherently a multilevel field that often makes use of multilevel modeling (MLM) for the analysis of clustered data. Given the widespread use of various rules of thumb and based on the findings of more recent studies, we provide guidance for applied researchers who are exploring the use of MLM in their own studies. Many of the myths have kernels of truth though researchers should be aware of the exceptions that make broad generalizations of the rules difficult.

Keywords: multilevel modeling, hierarchical linear modeling, statistical misconceptions

Supplemental materials: http://dx.doi.org/10.1037/spq0000272.supp

> The greatest enemy of knowledge is not ignorance; it is the illusion of knowledge.
>
> —Stephen Hawking

The use of multilevel modeling (MLM, also known as hierarchical linear modeling or HLM) has become increasingly popular when analyzing nested data. As indicated by Graves and Frohwerk (2009), "the discipline of school psychology is inherently a multilevel field" (p. 84) with students nested within schools. Observations within one group or cluster tend to be more alike with each other compared with observations within other groups violating a well-known regression assumption of observation independence (Cohen, Cohen, West, & Aiken, 2003). Further, group membership may influence individual behavior and outcomes (Bliese & Hanges, 2004).

A large number of books and articles have been written on how to analyze clustered data (e.g., Luke, 2004; Raudenbush & Bryk, 2002; Singer, 1998). The popularity of MLM in school psychology is suggested in that the most cited article from 2010 to 2015 in the *Journal of School Psychology* (Elsevier, 2015) was not one that focused on a particular substantive area of school psychology, but a primer on MLM (Peugh, 2010). A search on the number of peer-reviewed articles using the PsycNET database of the American Psychological Association (APA) with keywords[1] related to MLM indicated that in 2017, there were 179 articles published related to MLM, more than three times the number (i.e., 50 articles) published in 2007.

General and specialized MLM software, both free (e.g., R) and commercial (e.g., HLM and SAS), are readily available. However, together with the growth of MLM as an analytic technique, several myths regarding the method abound and are found in many well-respected journals suggesting that both authors and reviewers may not be fully aware of more recent developments in the field related to the analysis of clustered data. We highlight some of these myths and golden rules which deserve some attention as newer studies, which we focus on in the View Today section of each myth, may have clarified some prior ambiguous modeling related issues.

The goal of the current article is to provide suggestions and guidance to applied researchers who are considering MLM techniques in their own research. We refrain from citing studies that

This article was published Online First August 2, 2018.

Correspondence concerning this article should be addressed to Francis L. Huang, Department of Educational, School, and Counseling Psychology, College of Education, University of Missouri, 16 Hill Hall, Columbia, Missouri 65211. E-mail: huangf@missouri.edu

[1] APA PsycNET search at http://psycnet.apa.org. Keywords were related to the procedure or the software used in the analysis: "MLM," "HLM," "multilevel," "HGLM," "xtmixed," "glimmix," "mlwin," "PROC MIXED," "nlme," "lmer." We do not use the term "hierarchical" as at times, hierarchical regression is used which is not HLM.

may have followed these myths (though are available in the online supplemental materials Appendix) so as not to cast concerns about these studies. However, the reliance on these rules of thumb illustrates the complexity of the issues that exist in the field with regard to making decisions related to the use of MLM. As with various myths, there are also kernels of truth embedded within them. Often, the myths may be true, but are conditional on certain factors of which researchers should be aware.

Myth 1: When the Intraclass Correlation (ICC) Is Low, MLM Is Not Needed

The ICC (ρ) is a well-known statistic routinely used when conducting multilevel analysis. The ICC represents the amount of variance attributable to the group level and is commonly estimated using a null model (i.e., a model with no predictors) or equivalently, a one-way random effects analysis of variance. The ICC is computed as $\rho = \tau_{00}/\tau_{00} + \sigma^2$ where τ_{00} and σ^2 are associated with the variance of the between- and within-group error terms. Adding together τ_{00} and σ^2 will provide the total variance of the outcome variable. An ICC of 1 indicates that differences in the outcome variable are completely dependent on the grouping variable and an ICC of 0 signifies complete observation independence.

Often, the null model is computed initially to determine if an MLM is needed in the first place or to give an indication of how much variance the cluster can account for (cf., Peugh, 2010). Some methodologists, "literally too many to list" (Nezlek, 2008, p. 856), may suggest that with low ICCs, MLM models may not be needed at all and data may be analyzed using much simpler ordinary least squares (OLS) regression (Hayes, 2006). For example, in the absence of a substantial ICC (e.g., $\rho < .05$), Thomas and Heck (2001) indicated that "in such cases where the observations are nearly independent, traditional multiple regression analysis using appropriately weighted data will provide accurate estimates of the parameters and standard errors" (p. 526). Studies may then point to low ICCs and proceed with using more straightforward, single-level analyses.

View Today

It is true that with low ICCs and a low number of observations per cluster, Type I error may not be an issue. Simulation studies have shown that the higher the ICC, the more serious the repercussions on standard error estimates for Level 2 variables which may increase Type I errors (Maas & Hox, 2005; Musca et al., 2011). However, even with ICCs as low as .01 (see the Appendix in the online supplemental materials for an example), the Type I error rate may be as high as .20, four times higher than the conventionally used alpha of .05 (Musca et al., 2011). Best practice today is not to simply ignore the clustering effect, but to account for the clustering effect using MLM or some other alternative means (Huang, 2016).

To understand why even low ICCs may have a large impact, an understanding of design effects (DEFFs; Kish, 1965) is informative. DEFFs are known as the ratio of the operating variance to the sampling variance if a simple random sample were conducted. DEFF is computed as $1 + (n_c - 1)\rho$, where ρ is the ICC and n_c is the average (or the harmonic mean for unbalanced clusters) cluster size. Dividing the nominal sample size used in a study by DEFF will indicate what the effective sample size would be if a random sample were taken so a DEFF greater than 1 would reduce the estimated sample size. The only time when DEFF is equal to 1 is when the ICC = 0 or $n_c = 1$ (i.e., there is only one observation per cluster or, in other words, there is no clustering in effect).

Note that the effect of DEFF combines both the ICC and the average cluster size. Even if ICC is held constant, DEFF increases as cluster size increases. In other words, the ICC is not the sole determinant of the design effect. An ICC of .10 with an average cluster size of 10 would have a DEFF of 1.09 but an ICC of .01 with an average cluster size of 100 (e.g., students within schools) would result in a much higher DEFF of 1.99.

Tip

Often, the ICC, which is needed in the computation of the design effect, is estimated using an unconditional MLM. However, a quick way to approximate the ICC without needing to run an MLM is to perform an OLS regression using only the dummy coded grouping variables as the predictors—also known as a fixed effects model (Huang, 2016). Manually creating $g - 1$ dummy codes, where g is the total number of groups, may sound time consuming but statistical software can automatically create dummy codes using syntax (e.g., factor in R or the class statement in SAS) or drop down menus (transform → create dummy variables in SPSS). The adjusted R^2 (not the standard R^2) from the regression model, which represents the proportion of variance accounted for by the group factor, will approximate the ICC which is the amount of variance accounted for by the grouping variable (Huang, 2016).

Myth 2: When the Design Effect Is Less Than Two, MLM Is Not Needed

Related to Myth 1, another often-cited golden rule is that MLM may not be needed when DEFF is less than 2 (Maas & Hox, 2005; Peugh, 2010). Lai and Kwok (2015) indicated that the rule has often been invoked numerous times in the education, psychology, business, and medical literature (see Lai & Kwok, 2015 for a list of articles using the rule). Often, articles attribute the rule to Muthén and Satorra (1995) who actually did not explicitly state that general rule.[2]

View Today

In a recent study that investigated the DEFF <2 rule using Monte Carlo simulations that tested varying conditions (i.e., DEFFs, cluster size, and number of clusters), Lai and Kwok (2015) found some support for the rule though indicated that it works only in a limited number of situations and caution researchers when applying the rule of thumb. Support for the rule was found only when the number of observations within cluster was at least 10, the relationship between the Level 1 predictors and the outcome were constant (i.e., no random coefficients), and when the predictors were group-mean centered. If the research question focused on Level 2 effects, Lai and Kwok warned that standard errors would

[2] In the Muthén and Satorra (1995) study, DEFF was not a manipulated factor in their simulation but rather it was ICC.

be biased when DEFF $\geq$1.5 thereby increasing the probability of Type I errors.

Tip

Researchers may actually use the DEFF values and manually adjust standard errors used in statistical significance testing. For example, in computing for statistical significance, a model with nested data can be run using standard OLS regression and standard errors can be adjusted by multiplying the standard errors by the square root of DEFF (Hahs-Vaughn, 2005; McCoach & Adelson, 2010). Large DEFFs will result in higher standard errors which is why the adjustment has also been referred to as a variance inflation factor (Donner, 1998, p. 10).[3] The DEFF adjustment is an approximation and can be used for various procedures (e.g., structural equation modeling), not just regression (Fan, 2001).

Myth 3: Standard Errors From the OLS Analysis of Clustered Data Will Always Be Underestimated Resulting in Greater Type I Errors

An often cited reason for using MLM is to correct for the underestimated standard errors which may result when OLS regression is used in analyzing clustered data (Bliese & Hanges, 2004). However, MLM standard errors at Level 1 actually may be smaller (i.e., more powerful) compared with OLS standard errors. The myth of underestimated standard errors is partially correct and depends on the level of the variable of interest and the design of the study.

View Today

Level 2 standard errors will often be underestimated (Huang, 2016) though not necessarily so for Level 1 variables. If researchers are interested in Level 2 effects (e.g., intervention administered at the school or classroom level with student level outcomes) and data are analyzed using standard OLS regression, the coefficients will often have underestimated standard errors (depending on the cluster size and number of clusters) as a result of the data not being analyzed with the actual number of observations. For example, in a study of 300 students nested within 30 schools, the predictor variables at the school level will be estimated with an n of 300 (the sample size of the students) instead of $n = 30$ (the actual number of groups). As n increases, standard errors decrease which results in increased power to reject the null hypothesis. In addition, erroneous degrees of freedom will be used when evaluating statistical significance, again increasing the probability of Type I errors (e.g., the critical value for a t test for a study with 30 participants is larger compared with the critical value of a study with 300 participants). The implications for ignoring the clustering can spell the difference between supporting or rejecting certain hypotheses. Baldwin, Murray, and Shadish (2005) reanalyzed 33 studies which administered group-level treatments and ignored the clustered nature of the data. After applying a correction factor, six to 19 out of the 33 studies no longer had statistically significant results.

For Level 1 predictors, Bliese (2000) mentioned that the estimates based on OLS regression can be "too liberal or too conservative" indicating that the bias can go in either direction. Studies that have used secondary data sets as well as Monte Carlo simulations have indicated that Level 1 standard errors for OLS regression may be more conservative (i.e., too small) and at other times more liberal (i.e., too high) as well (Arceneaux & Nickerson, 2009; Astin & Denson, 2009; Harden, 2011; Huang, 2014, 2016; Rocconi, 2013).

To explain this, a multilevel model partitions the variance at the between and within levels as illustrated earlier using τ_{00} and σ^2. However, in analyzing a Level 1 outcome with only a Level 1 predictor using OLS regression, standard error estimates are based on the total variance, which is equal to the sum of τ_{00} and σ^2 and is thus larger than σ^2 alone (i.e., the Level 1 variance) when ICC is greater than zero (Bliese & Hanges, 2004). The result is a more conservative test of statistical significance resulting in more Type II errors and a loss of power. A more accurate characterization then of the standard error estimates using OLS with clustered data for Level 1 variables is that standard errors may be *misestimated* as the direction of the bias may be positive or negative.

Myth 4: MLM and OLS Differ Only in Their Standard Errors and Regression Coefficients Will Be the Same

Studies comparing regression coefficient estimates using OLS regression and MLM have shown the parameter estimates may not differ greatly between the two methods (Astin & Denson, 2009; Huang, 2016; Lai & Kwok, 2015). Others have mentioned that OLS will generally produce unbiased estimates for the regression coefficients suggesting approximately the same estimates regardless of the type of model used. In a Monte Carlo simulation using 15,000 data sets across a range of ICCs (from .00 to .95), Mundfrom and Schultz (2001) compared regression coefficient estimates between OLS and MLM and "showed remarkable similarity when compared with each other" (p. 20) though they noted also that MLM provides better, more accurate estimates of standard errors.

View Today

Although coefficient estimates may often be similar in OLS and MLM models, that may not always the case (Huang, 2018a). When Level 1 predictors are correlated with the higher level group or unit effects which are not included in the model, bias is introduced (Bafumi & Gelman, 2006). The bias is not merely theoretical or a technical issue.

In an analysis using the PISA 2012 dataset in Thailand, Huang (2018a) used an indicator variable if a student spoke another language at home ($1 = yes$, $0 = no$) to predict reading achievement. Analyzed using OLS regression, results showed a statistically significant and negative relationship ($B = -10.9$, $p < .001$) indicating that if a student spoke another language at home, this was associated with poorer reading outcomes. However, using the same variables but analyzed using MLM, results were the opposite and students who spoke another language at home had higher reading scores ($B = 8.1$, $p < .001$). Using different model specifications, Huang illustrated how models can be estimated to produce the same point estimates using either method.

[3] Not to be confused with the regression diagnostic used to test for multicollinearity.

Researchers should keep in mind that MLM does not control for variables at the higher level if the variable is not included in the model. Not including variables at the higher level may result in omitted variable bias at Level 1 where coefficient estimates are higher or lower than they should actually be. In prior simulations that have examined the differences between OLS and MLM models (e.g., Mundfrom & Schultz, 2001), Level 1 and Level 2 predictors were generated orthogonally (i.e., not correlated) so no bias was present, thereby producing comparable results.

Tip

Fortunately though, if the researcher is interested in the Level 1 coefficients (e.g., student socioeconomic status [SES]), getting unbiased coefficients can be done several ways—and these methods are applicable when using either OLS or MLM. One way is to include the contextual effect or the Level 2 aggregation of the Level 1 variable (in this case, the school average of SES; Huang, 2018a). Another way is to include the group-mean centered Level 1 variable, also known as centering within context (CWC). If the researcher is interested in the association of the Level 1 predictor (X_{ij}) on the outcome (Y_{ij}), group-mean centering is the best option because group-mean centering removes all between-group variation (Dalal & Zickar, 2012). Group-mean centering or using demeaned data (i.e., subtracting the group mean from variables) effectively eliminates the group-level effect from the variable and reduces the ICC of the predictor variable to zero as all of the clusters will have a mean of zero for the centered variable. Enders and Tofighi (2007) stated that analyses using grand-mean centered variables result in an ambiguous mixture of Level 1 and Level 2 associations with the *X* and *Y* variables and that CWC results in coefficients that were "pure estimates" (p. 127) of the association between the level one variable with *Y*. Finally, a third way to remove bias resulting from missing higher-level variables is to run a fixed-effect model which merely includes the dummy coded cluster variables as Level 2 predictors (see Huang, 2016, 2018a).

Myth 5: There Is No Overall R^2 When Using Multilevel Linear Models

Frequently, studies using MLM show a reduction in variance at the between and within levels using a pseudo R^2 as a global effect size measure. Indeed, this is in advantage of MLM where variance can be partitioned at both the between and within groups allowing researchers to indicate the proportion of variance explained at either or both levels. However, researchers may want an overall R^2 statistic just like in OLS regression, which explains how much overall variance is accounted for in the outcome variable by the independent variables. Some may indicate that "the [R^2] statistics are computed in different ways, there is no straightforward comparison of variance explained statistics between OLS regression and HLM analysis" and that the "variance explained statistics are not directly comparable between analyses" (Rocconi, 2013, p. 456). Though it is true that proportion of variance reduced at different levels are not directly comparable to the R^2 statistic in an OLS regression (which is also why they are often referred to as pseudo R^2s and at times may be negative), an overall R^2, which means the same thing in both OLS and MLM models, can be computed. The challenge is that R^2 values are automatically provided in OLS regression output whereas computing the R^2 in an MLM requires some additional, though straightforward, computations since they are not routinely provided. Some authors may reestimate MLMs using OLS regression, assume that parameter estimates are the same as with an MLM, and report the OLS R^2 instead (see the Appendix in the online supplemental materials).

View Today

In a regression model, the R^2 can be conceptualized as the squared correlation between the predicted $((\widehat{Y}_{ij})$ values and the actual observed Y_{ij} values (Agresti & Finlay, 1997). Instead of simply viewing R^2 as a percentage of variance accounted for, R^2 can be viewed as the proportion reduction of prediction error (Luke, 2004). As a measure of global effect size, R^2 can be computed, in both an MLM and OLS regression model, by correlating the predicted scores and the observed scores and squaring that coefficient (Peugh, 2010). Roberts (2004) also showed the computation in an MLM setting using the sums of squares (regression) divided by the sums of squares (total) which yields the same results (i.e., $R^2 = \text{SS}_{\text{reg}}/\text{SS}_{\text{total}}$). Although the proportion of variance reduced at the different levels is useful, an overall R^2 may also be informative with regard to the overall variance explained by the dependent variables which is readily understood. The logic follows that if one model performs better than another model, it should be more accurate, have lower residual error, and thus have a higher R^2. For a comparison of different explained variance measures, readers can consult LaHuis, Hartman, Hakoyama, and Clark (2014).

Tip

A simple way—without having to correlate the observed and predicted values—is to compare the reduction in total variance from the null to the full model. For example, in the null model, the variance of the outcome variable is 100. In the full model, the variance (i.e., the within- plus the between-level variance) is 70. Then, the 30-point reduction in variance is equal to an R^2 of .30 or the predictors explained 30% of the variance in the outcome—which means the same thing as in standard OLS regression.

Although R^2 may be informative as a measure of effect size, for an evaluation study where it is important to show how meaningful the difference is between a treatment and control conditions, the R^2 does not communicate this magnitude. For example, in one of the most influential experimental studies involving class size and student achievement, the Tennessee Project STAR (Mosteller, 1995), our own analyses show the R^2 measure to be .02 based on the treatment assignment variable alone. A more meaningful measure is a standardized mean difference such as Cohen's (1992) *d*. Estimating effect sizes may be done for binary predictors if the outcome variable is standardized (i.e., *z* scored) so that the regression coefficients for the binary predictors can be interpreted in standard deviation units.

Myth 6: MLM Is Not Necessary With Factor Analysis

Actually, this myth has not been explicitly stated as such but is evident in several factor analytic studies which do not account for the clustered nature of the data when nesting is present. The violations of nested designs in factor analytic work are quite common with measures related to school climate or teacher evaluations where results from the individual respondents are factor analyzed but in actuality, the higher level construct (e.g., the school climate or the teacher rating, not the individual response) is of interest. This however may not always be an issue but is dependent on the level of interest.

View Today

The majority of parametric statistical procedures used, which includes factor analysis, are part of the general linear model (Graham, 2008) which assumes the statistical independence of observations. Factor analytic studies that ignore the clustered nature of the data are still the norm, despite that over a decade ago, Julian (2001) wrote about the consequences of ignoring the nested structure present with multilevel data. Julian (2001) indicated that as ICC increased, model fit indices, χ^2 statistics, parameter estimates, and standard errors all exhibited estimation problems. Older studies have also indicated the problems associated with not accounting for the clustered nature of the data (Kaplan & Elliott, 1997; Muthén & Satorra, 1995). Konold et al. (2014) suggested several reasons why this may be the case: (a) a limited number of software packages that can perform multilevel factor analysis, (b) estimation and convergence issues, or (c) a failure to recognize the nested data structure when present. Indeed, Heck and Thomas (2008) indicated that years ago, getting software to estimate multilevel factor analytic models were "programming nightmares for even simple within- and between-group factor models" (p. 114).

Given the findings of several methodological studies (Julian, 2001; Kaplan & Elliott, 1997; Muthén & Satorra, 1995), the clustered nature of the data should be accounted for, especially if the factors of interest are higher level constructs (e.g., school climate). However, unlike multilevel regression models, multilevel factor analysis has an additional complication that the factor structures, which are often the focus of the studies, may differ at the individual and at the group level (Bliese, 2000; Dyer, Hanges, & Hall, 2005; Huang & Cornell, 2016; Huang, Cornell, & Konold, 2015). The implication of the invariance in factor structures at different levels is large, especially when the unit of interest is at the group level. If individual-level data are aggregated to form group-level composites (e.g., an evaluation of teacher effectiveness based on individual student feedback), and the factor structures at both levels differ, results will be misleading as the variables may load on different factors at the different levels.

Group-level factor structures have been found to often be simpler compared with individual-level factor structures (Huang & Cornell, 2016). Schweig (2014) demonstrated in an analysis of a school climate measure and a teacher evaluation measure, that when factor structures differ at both levels, scales formed based on factor loadings can be highly misleading. The problem of invariant factor structures though cannot simply be solved by adjusting standard errors or applying a correction procedure but requires multilevel factor analysis or factor analyzing the properly estimated correlation matrix at the different levels (Schweig, 2014; Stapleton, 2006). In addition, reliability estimates (e.g., Cronbach's alpha, omega) at Level 1 are not necessarily the same as the reliability estimates at Level 2 (Geldhof, Preacher, & Zyphur, 2014) though relatively straightforward to compute (Huang, 2017).

Myth 7: Clustering Can Always Be Accounted for Properly Using the "Type = Complex" Option in Mplus

The availability of Mplus has greatly helped applied researchers in dealing with clustered data. Numerous articles mention handling the clustered nature of their data by using the type = complex option in Mplus or even at times merely indicating clustering was automatically accounted for by using Mplus without even indicating the procedure used. However, analysts should understand what the option is actually doing and how the clustering is handled as it may not be appropriate in some situations. The use of type = complex is not a statistical approach in itself.[4] Based on Mplus documentation, the type = complex option applies the well-known Huber White[5] standard error adjustments and retains the parameter estimates.

The standard error adjustment uses a sandwich estimation procedure (Berger, Graham, & Zeileis, 2017) which may account for the clustering when the number of groups is approximately 25 or more (Huang, 2014, 2016). With few clusters however, the standard errors may still be misestimated (Bell & McCaffrey, 2002; Cameron & Miller, 2015). This is recognized in the Mplus discussion board as well (see Footnote 5; Muthén, March 10, 2005). So, as long as there are a reasonable number of groups and the assumption that the relationship of the variables at Level 1 and higher are the same, using type = complex may account for misestimated standard errors.

Myth 8: At Least 30 or 50 Clusters/Groups Are Needed to Use a Multilevel Model

A commonly cited rule of thumb that MLMs require at least 30 groups with 30 individuals per group (i.e., the 30/30 rule) can be attributed to Kreft's (1996) unpublished article. Based on a review of MLM studies, Tonidandel, Williams, and LeBreton (2014) indicated that this 30/30 rule was the most widely cited guideline for required sample sizes using MLMs. However, Tonidandel et al. (2014) pointed out that Kreft's (1996) study was based on a review of other unpublished articles, focused on fixed effects estimation, and were for obtaining power for cross-level interactions.

Another often-cited reference for MLM sample sizes is a simulation study (using various individual and group sample size conditions) of Maas and Hox (2005) who indicated in their study abstract that ". . . a small sample size at level two (meaning a sample of 50 or less) leads to biased estimates of the second-level standard errors" (p. 86). However, Maas and Hox (2005) were specifically referring to estimates for the residual variance components and indicated, in conclusion, that "both the regression coefficients and the variance components are all estimated without bias, in all of the simulated conditions. The standard errors of the regression coefficients are also estimated accurately, in all of the simulated conditions." (p. 90). Several studies, though, erroneously reference Maas and Hox as a reason to require at least 50

[4] Similar procedures with R require the lavaan.survey package (Oberski, 2014) in a latent variable framework or the survey package (Lumley, 2014) in a regression framework.

[5] http://www.statmodel2.com/discussion/messages/12/587.html?1376493089

clusters in order to use an MLM even if only interested in the fixed effects.

View Today

Even with a small number of clusters, MLMs may result in unbiased estimated for the regression coefficients and standard errors. Several simulation studies have shown that MLM may be used even with as little as 10 groups (Bell, Morgan, Schoeneberger, Kromrey, & Ferron, 2014; Huang, 2016, 2018b; McNeish & Stapleton, 2016). However, with a smaller number of clusters, restricted maximum likelihood is recommended compared with the use of maximum likelihood estimation (Goldstein, 2011; Huang, 2016; Meijer, Busing, & Van der Leeden, 1998) together with a Kenward & Roger (1997) degrees of freedom adjustment or Satterthwaite approximation (see McNeish & Stapleton, 2016 for detailed explanation).

Often, cluster randomized trials (CRTs) may operate with a limited number of groups (which is a practical limitation). A review of 285 CRTs in the health sciences indicated that the median number of clusters used in studies was 21 (Ivers et al., 2011), far less than the 30 clusters or even 50 clusters often cited. However, to determine the number of clusters required for MLM studies, we strongly recommend conducting actual power analyses using freely available software rather than simply relying on rules of thumb. Free and readily available software such as Optimal Design (Spybrook et al., 2011) or PowerUp! (Dong & Maynard, 2013) were specifically developed for that purpose.

Implications for Practice

Numerous developments and methodological studies related to MLM have been conducted within the past decade alone. Although years ago, unfamiliarity with MLM was "commonplace within the field of school psychology" (Graves & Frohwerk, 2009, p. 91), MLM today remains an important analytic tool, especially with school- or group-based studies that randomize intact groups to treatment or control conditions often used in school-based intervention studies (Resnicow et al., 2010). What has become apparent, though, with the availability of various MLM tutorials, access to software, and the presence of nested data, is that the clustered data structure should not be ignored, but rather it should be accounted for properly—and MLM is not necessarily the only technique that can be used (for alternatives such as using fixed effect models and cluster robust standard errors, see Huang, 2016). Ignoring the clustering effect, even with an ICC as low as .01, can have practical, real-world implications (see the Appendix in the online supplemental materials). In the case of factor analytic work, this multilevel analysis may be even more important if the unit of interest is the higher level unit (e.g., teacher evaluations by students).

Researchers should also take care in citing various references (e.g., Maas & Hox, 2005; Muthén & Satorra, 1995) which may not actually state the cited rules and lead to further perpetuation of certain myths. In addition, care should be taken in citing applied studies that use these guidelines as the more modern view with regard to the rules of thumb, informed by simulations and newer studies, may have changed. Many of the cited myths have much truth in them—though at times, researchers may not be aware of the exceptions to the rules that prevent their overall generalization.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Alderman, H., Konde-Lule, J., Sebuliba, I., Bundy, D., & Hall, A. (2006). Effect on weight gain of routinely giving albendazole to preschool children during child health days in Uganda: Cluster randomised controlled trial. *British Medical Journal, 333,* 122. http://dx.doi.org/10.1136/bmj.38877.393530.7C

Alderman, H., Konde-Lule, J., Sebuliba, I., Bundy, D., & Hall, A. (2012). Correction to "Effect on weight gain of routinely giving albendazole to preschool children during child health days in Uganda: Cluster randomised controlled trial." *British Medical Journal, 345,* e8724. http://dx.doi.org/10.1136/bmj.e8724

Arceneaux, K., & Nickerson, D. W. (2009). Modeling certainty with clustered data: A comparison of methods. *Political Analysis, 17,* 177–190. http://dx.doi.org/10.1093/pan/mpp004

Astin, A. W., & Denson, N. (2009). Multi-campus studies of college impact: Which statistical method is appropriate? *Research in Higher Education, 50,* 354–367. http://dx.doi.org/10.1007/s11162-009-9121-3

Bafumi, J., & Gelman, A. (2006). Fitting multilevel models when predictors and group effects correlate. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1010095

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology, 73,* 924–935. http://dx.doi.org/10.1037/0022-006X.73.5.924

Bear, G. G., Yang, C., Pell, M., & Gaskins, C. (2014). Validation of a brief measure of teachers' perceptions of school climate: Relations to student achievement and suspensions. *Learning Environments Research, 17,* 339–354. http://dx.doi.org/10.1007/s10984-014-9162-1

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). Blood pressure goals: How low should you go? Lowering pressure to 140/90 is a reasonable goal for otherwise healthy men at any age. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10,* 1–11. http://dx.doi.org/10.1027/1614-2241/a000062

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

Berger, S., Graham, N., & Zeileis, A. (2017). Various versatile variances: An object-oriented implementation of clustered covariances in R. Retrieved from https://cran.r-project.org/web/packages/sandwich/vignettes/sandwich-CL.pdf

Bliese, P. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. Bollen & J. Long (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods, 7,* 400–417. http://dx.doi.org/10.1177/1094428104268542

Breevaart, K., Bakker, A., Hetland, J., Demerouti, E., Olsen, O. K., & Espevik, R. (2014). Daily transactional and transformational leadership and daily employee engagement. *Journal of Occupational and Organizational Psychology, 87,* 138–157. http://dx.doi.org/10.1111/joop.12041

Cambré, B., Kippers, E., van Veldhoven, M., & De Witte, H. (2012). Jobs and organisations. *Personnel Review, 41,* 200–215. http://dx.doi.org/10.1108/00483481211200033

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Cornell, D., Allen, K., & Fan, X. (2012). A randomized controlled study of the Virginia Student Threat Assessment Guidelines in kindergarten through Grade 12. *School Psychology Review, 41,* 100–115.

Dalal, D. K., & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods, 15,* 339–362. http://dx.doi.org/10.1177/1094428111430540

Datta, P., Cornell, D., & Huang, F. (2017). The toxicity of bullying by teachers and other school staff. *School Psychology Review, 46,* 335–348. http://dx.doi.org/10.17105/SPR-2017-0001.V46-4

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6,* 24–67. http://dx.doi.org/10.1080/19345747.2012.673143

Donner, A. (1998). Some aspects of the design and analysis of cluster randomization trials. *Journal of the Royal Statistical Society Series C, Applied Statistics, 47,* 95–113. http://dx.doi.org/10.1111/1467-9876.00100

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly, 16,* 149–167. http://dx.doi.org/10.1016/j.leaqua.2004.09.009

Elsevier. (2015). Most cited Journal of School Psychology articles. Retrieved from http://www.journals.elsevier.com/journal-of-school-psychology/most-cited-articles/

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

Fan, X. (2001). Parental involvement and students' academic achievement: A growth modeling analysis. *Journal of Experimental Education, 70,* 27–61. http://dx.doi.org/10.1080/00220970109599497

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72–91. http://dx.doi.org/10.1037/a0032138

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Hoboken, NJ: Wiley.

Graham, J. M. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics, 33,* 485–506. http://dx.doi.org/10.3102/1076998607306151

Graves, S. L., Jr., & Frohwerk, A. (2009). Multilevel modeling and school psychology: A review and practical example. *School Psychology Quarterly, 24,* 84–94. http://dx.doi.org/10.1037/a0016160

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *Journal of Experimental Education, 73,* 221–248. http://dx.doi.org/10.3200/JEXE.73.3.221-248

Harden, J. J. (2011). A bootstrap method for conducting statistical inference with clustered data. *State Politics & Policy Quarterly, 11,* 223–246. http://dx.doi.org/10.1177/1532440011406233

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research, 32,* 385–410. http://dx.doi.org/10.1111/j.1468-2958.2006.00281.x

Heck, R. H., & Thomas, S. L. (2008). *An introduction to multilevel modeling techniques* (2nd ed.). New York, NY: Routledge.

Huang, F. (2014). Analyzing group level effects with clustered data using Taylor series linearization. *Practical Assessment, Research & Evaluation, 19,* 1–9.

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

Huang, F. (2017). *Conducting multilevel confirmatory factor analysis using R* [Working paper]. Retrieved from http://dx.doi.org/10.13140/RG.2.2.12391.34724

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

Huang, F. L. (2018b). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and Psychological Measurement, 78,* 297–318. http://dx.doi.org/10.1177/0013164416678980

Huang, F., Bergin, C., Tsai, C., & Chapman, S. (2016, April). *Multilevel factor structure of a middle-school teacher effectiveness survey*. Paper presented at the American Education Research Association, Washington, DC.

Huang, F. L., & Cornell, D. G. (2016). Using multilevel factor analysis with clustered data: Investigating the factor structure of the Positive Values scale. *Journal of Psychoeducational Assessment, 34,* 3–14.

Huang, F. L., Cornell, D. G., & Konold, T. R. (2015). Aggressive attitudes in middle schools: A factor structure and criterion-related validity study. *Assessment, 22,* 497–512. http://dx.doi.org/10.1177/1073191114551016

Huang, F. L., & Invernizzi, M. A. (2013). Birthday effects and preschool attendance. *Early Childhood Research Quarterly, 28,* 11–23. http://dx.doi.org/10.1016/j.ecresq.2012.03.002

Ivers, N. M., Taljaard, M., Dixon, S., Bennett, C., McRae, A., Taleban, J., . . . Donner, A. (2011). Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: Review of random sample of 300 trials, 2000–8. *BMJ: British Medical Journal, 343,* d5886. http://dx.doi.org/10.1136/bmj.d5886

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8,* 325–352. http://dx.doi.org/10.1207/S15328007SEM0803_1

Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling, 4,* 1–24. http://dx.doi.org/10.1080/10705519709540056

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

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Konold, T., Cornell, D., Huang, F., Meyer, P., Lacey, A., Nekvasil, E., . . . Shukla, K. (2014). Multilevel multi-informant structure of the authoritative school climate survey. *School Psychology Quarterly, 29,* 238–255. http://dx.doi.org/10.1037/spq0000062

Kreft, I. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17,* 433–451. http://dx.doi.org/10.1177/1094428114541701

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

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985147

Lumley, T. (2014). *Survey: Analysis of complex survey samples. R package version 3.30*. Retrieved from https://cran.r-project.org/web/packages/survey/survey.pdf

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

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

McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): Understanding the effects of clustered data. *Gifted Child Quarterly, 54,* 152–155. http://dx.doi.org/10.1177/0016986210363076

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28,* 295–314. http://dx.doi.org/10.1007/s10648-014-9287-x

Meijer, E., Busing, F., & Van der Leeden, R. (1998). Estimating bootstrap confidence intervals for two-level models. In J. J. Hox & J. De Leeuw (Eds.), *Assumptions, robustness, and estimation methods in multivariate modeling* (pp. 35–48). Amsterdam, the Netherlands: Publikaties.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 72,* 159–217. http://dx.doi.org/10.1111/j.1468-0262.2004.00481.x

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5,* 113–127. http://dx.doi.org/10.2307/1602360

Mundfrom, D. J., & Schultz, M. R. (2001). A comparison between hierarchical linear modeling and multiple linear regression in selected data sets. *Multiple Linear Regression Viewpoints, 27,* 3–11.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology, 2,* 74. http://dx.doi.org/10.3389/fpsyg.2011.00074

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

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2,* 842–860. http://dx.doi.org/10.1111/j.1751-9004.2007.00059.x

Niehaus, E., Campbell, C. M., & Inkelas, K. K. (2014). HLM behind the curtain: Unveiling decisions behind the use and interpretation of HLM in higher education research. *Research in Higher Education, 55,* 101–122. http://dx.doi.org/10.1007/s11162-013-9306-7

Oberski, D. L. (2014). lavaan.survey: An R package for complex survey analysis of structural equation models. *Journal of Statistical Software, 57,* 1–27. http://dx.doi.org/10.18637/jss.v057.i01

O'Malley, M., Voight, A., Renshaw, T. L., & Eklund, K. (2015). School climate, family structure, and academic achievement: A study of moderation effects. *School Psychology Quarterly, 30,* 142–157. http://dx.doi.org/10.1037/spq0000076

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48,* 85–112. http://dx.doi.org/10.1016/j.jsp.2009.09.002

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Resnicow, K., Zhang, N., Vaughan, R. D., Reddy, S. P., James, S., & Murray, D. M. (2010). When intraclass correlation coefficients go awry: A case study from a school-based smoking prevention study in South Africa. *American Journal of Public Health, 100,* 1714–1718. http://dx.doi.org/10.2105/AJPH.2009.160879

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities, 2,* 30–38.

Rocconi, L. (2013). Analyzing multilevel data: Comparing findings from hierarchical linear modeling and ordinary least squares regression. *Higher Education, 66,* 439–461. http://dx.doi.org/10.1007/s10734-013-9615-y

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36,* 259–280. http://dx.doi.org/10.3102/0162373713509880

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

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" software. Retrieved from http://www.wtgrantfoundation.org/resources/optimal-design

Stapleton, L. (2006). Using multilevel structural equation modeling techniques with complex sample data. In G. Hancock & R. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 345–383). Greenwich, CT: Information Age.

Thomas, S. L., Heck, R. H., & Bauer, K. W. (2005). Weighting and adjusting for design effects in secondary data analyses. *New Directions for Institutional Research, 2005,* 51–72. http://dx.doi.org/10.1002/ir.155

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42,* 517–540.

Tonidandel, S., Williams, E. B., & LeBreton, J. M. (2014). Size matters . . . just not in the way that you think: Myths surrounding sample size requirements for statistical analyses. In C. Lance & R. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 162–183). New York, NY: Routledge.

Vandenberghe, C., Bentein, K., Michon, R., Chebat, J.-C., Tremblay, M., & Fils, J.-F. (2007). An examination of the role of perceived support and employee commitment in employee-customer encounters. *Journal of Applied Psychology, 92,* 1177–1187. http://dx.doi.org/10.1037/0021-9010.92.4.1177

van Starrenburg, M. L. A., Kuijpers, R. C. M. W., Kleinjan, M., Hutschemaekers, G. J. M., & Engels, R. C. M. E. (2017). Effectiveness of a cognitive behavioral therapy-based indicated prevention program for children with elevated anxiety levels: A randomized controlled trial. *Prevention Science, 18,* 31–39. http://dx.doi.org/10.1007/s11121-016-0725-5

Wallace, J. C., Edwards, B. D., Arnold, T., Frazier, M. L., & Finch, D. M. (2009). Work stressors, role-based performance, and the moderating influence of organizational support. *Journal of Applied Psychology, 94,* 254–262. http://dx.doi.org/10.1037/a0013090

Yang, C., Bear, G. G., Chen, F. F., Zhang, W., Blank, J. C., & Huang, X. (2013). Students' perceptions of school climate in the U.S. and China. *School Psychology Quarterly, 28,* 7–24. http://dx.doi.org/10.1037/spq0000002

Zullig, K. J., Collins, R., Ghani, N., Hunter, A. A., Patton, J. M., Huebner, E. S., & Zhang, J. (2015). Preliminary development of a revised version of the School Climate Measure. *Psychological Assessment, 27,* 1072–1081. http://dx.doi.org/10.1037/pas0000070

Received March 22, 2018
Revision received June 2, 2018
Accepted June 7, 2018 ■