Supplementary Material for:


We specifically do not cite these studies in the main body of the manuscript so as not to question the original study findings. Several of these come from leading journals such as School Psychology Quarterly, School Psychology Review, Prevention Science, the BMJ, and Psychological Assessment (there are more). These references though illustrate the complexity of the issues that exist in the field as they relate to multilevel modeling and reflect how developments may change over time. We use these references as well because their results are robust to any particular violations (i.e., there is nothing wrong with the results of the analysis unless stated otherwise) based on the related myth. However, we believe that myths may be perpetuated by citing studies that have used these rules of thumb. Refer to the main article for the explanation of the myths to provide a context for the quotes.

**Myth 1: When the intraclass correlation is low, multilevel modeling is not needed.**

“Various authors (literally too many to list – this is an ongoing discussion) suggest that multilevel models are not appropriate when something called the intraclass correlation (ICC) is low (or 0).” (Nezlek, 2008, p. 856)

For example, Huang and Invernizzi (2013, p. 15) first tested an unconditional MLM to see if “multilevel modeling was necessary (i.e., the intraclass correlation coefficient or \( \rho \) was greater than .05).” Cornell, Allen, and Fan (2012) indicated in their cluster randomized control trial that ICC “coefficients ranged from 0.04 to 0.07 and therefore were deemed unlikely to cause serious inflation of the Type I error rate” (p. 108). NOTE: Type I errors were accounted for in their analyses as a more conservative alpha of .01 was used in evaluating statistical significance as recommended by Heck, Thomas, and Bauer (2005). The study had 201 students nested within 40 schools (or an average of 5 students per school).

Even with low ICCs, accounting for the clustering effect is extremely important. Alderman, Konde-Lule, Sebuliba, Bundy, and Hall (2006) conducted a cluster randomized control trial focusing on child health (with weight gain as the outcome) and provided deworming medicine to preschool aged children in the treatment group. Deworming, which has been a contentious topic, has been an identified way to improving nutritional status among malnourished children which leads to better school attendance through reduced absenteeism (Miguel & Kremer, 2003). Participants included 27,995 children in 48 parishes (\( \sim \) 582 children per parish). Half of the parishes were assigned to the treatment group and half were assigned to the control group. Results indicated (in Table 2 of the article) that children receiving the deworming medicine pills gained approximately 154 g (CI: 91 – 214, \( p < .01 \)) or approximately 10% of average initial body weight. A few years later, a correction was published in the BMJ\(^1\) (2012) by the authors who indicated that they inadvertently failed to account for the clustering and results were no longer statistically significant 154 g (CI: -19.7 – 330, \( p > .05 \)) once accounted for. Based on our

---

\(^1\) In 2016, the BMJ had an impact factor of 20.79.
calculations, using an estimated design effect approximated from the change in standard errors together with the average cluster size, we estimated the ICC to be a low .014.

**Myth 2: When the design effect is less than two, multilevel modeling is not needed.**

See Lai and Kwok (2015) for an extensive list of articles from several disciplines such as education, psychology, business, and medicine that cite this rule.

**Myth 3: Standard errors from the OLS analysis of clustered data will always be underestimated resulting in greater Type I errors.**

Huang (2016) indicated that even with moderate ICCs, “OLS standard errors may be underestimated” (without specifying the level where the underestimation is occurring) though he later also indicates that standard errors may be too large. O’Malley, Voight, Renshaw, and Eklund (2015) wrote that the multilevel model used in their analysis rendered “standard error estimates more conservative by accounting for the common variance in the outcome variable” (p. 148) even though their analyses focused on only student-level (level one) variables with no school-level predictors.

Methodologists may indicate that “standard multivariate models are not appropriate for the analysis of such hierarchical systems, even if the analysis includes only variables at the lowest (individual) level, because…standard errors are negatively biased, which results in spurious ‘significant’ effects” (Maas & Hox, 2004, p. 428).

**Myth 4: MLM and OLS differ only in their standard errors and regression coefficients will be the same.**

Several studies cited in text show that the MLM and OLS results may be similar (see main article). However, this may not always be the case. In these cases, “absence of evidence is not evidence of absence.”

**Myth 5: There is no overall R² when using multilevel linear models.**

“Although OLS regression generates biased parameter estimates and standard errors when analyzing multilevel data, it does provide an adequate approximation of effect-size estimates (i.e., R²) for the overall variance in individual-level outcomes that is explained by individual-level and group-level predictors … Thus, we reported R² results from OLS regression analyses as a way of conveying effect sizes.” (Wallace, Edwards, Arnold, Frazier, & Finch, 2009, p. 258)

“Rather than using pseudo-R² estimates, the effect size assessments for self-presentation and helping behavior were derived from ordinary least squares (OLS) regression.” (Vandenberghhe et al., 2007).
Myth 6: Multilevel modeling is not necessary with factor analysis.

Recent studies on measures of school climate, which take student-level measures to make decisions about school level factors, despite recognizing the clustered nature of the data, are often analyzed at only the student-level (Bear, Yang, Pell, & Gaskins, 2014; Zullig et al., 2015). Factor analytic studies may also invoke the different golden rules specified earlier as a basis for ignoring the clustering effect. For example, a factor analytic study using nested data indicated that DEFF was not greater than two for most variables and indicated that clustering did not need to be accounted for (Yang et al., 2013). For school climate, conducting factor analytic studies on student level information only is also important as at times, level one measures are used as outcomes (e.g., Datta, Cornell, & Huang, 2017). NOTE: studies may also correctly account for the clustering effect by demeaning the data (e.g., Bear, Gaskins, Blank, & Chen, 2011)—but do not test for the factor structure at the higher level. The level 1 factor structure may be reasonable, but if not tested, the level 2 factor structure is unknown so generalizations to the higher level construct may be difficult. The need for MCFA is probably even more important for student evaluations of teaching as the unit of interest is the teacher and students are merely informants (Huang, Bergin, Tsai, & Chapman, 2016).

Myth 7: Clustering can always be accounted for properly using the “type = complex” option in Mplus.

The following examples merely indicate that Mplus was used—but do not actually indicate what specific statistical procedure was performed.

“Therefore, we used the ‘TYPE = COMPLEX’ procedure in Mplus to calculate standard errors and chi-square values …” (Breevaart et al., 2014).

“Since randomization took place within the school level and children were nested within these schools, we used Mplus 6.1 (Muthen and Muthen 1998) to control for potential clustering effects.” (Starrenburg, Kuijpers, Kleijnjan, Hutschemaekers, & Engels, 2017).

Myth 8: At least 30 or 50 clusters/groups are needed to use a multilevel model.

“Another challenge in using HLM is that a 2- or 3-level model requires a larger sample size than might be available for some research questions. For example, in a simulation study using a 2-level model, Maas and Hox (2005) suggested that it would be necessary to have at least 50 cases at level-2 in order to achieve unbiased standard errors. Although 50 programs or institutions may be readily available in large, national datasets, this may be beyond many smaller studies.” (Niehaus, Campbell, & Inkelas, 2014)

“According to Maas and Hox (2005), a minimum of 30 cases at the highest, team level of analysis is needed for adequate power in multilevel modelling. Following this rule of thumb, we do not have a sufficient amount of cases at the highest, third level (N = 8) required for robust estimations” (Breevaart et al., 2014).
“Our 24 job groups approach the preferred number of (at least) 30 groups on the second level prescribed by some authors (Kreft, 1996; Maas and Hox, 2004) for multilevel analysis” (Cambré, Kippers, van Veldhoven, & De Witte, 2012, p. 211)
References


