Sample Size Issues and Power

There are two sample size issues to be concerned about. One issue has to do with the minimum number of cases needed for using multilevel regression to avoid biases. The second issue concerns sufficient statistical power needed for obtaining significance. Generally, having more groups is more important than having more cases per group for either of these concerns (see Scherbaum & Ferreter, 2009, for a review).

Convergence and estimation bias

It is important to consider the minimum number of cases needed to ensure the model converges and that sample estimates are not biased. Most simple models with 50 or more groups and approximately 5-10 cases per group will not have convergence problems (e.g., McNeish & Stapleton, 2016). More cases may be needed for convergence if the model is more complex, when there is more missing data (unbalanced n_j), and when more slope variances are estimated (Raudenbush, 2008). For binary outcomes, the likelihood that a model will not converge can be quiet high when the number of groups is low (e.g., < 30-50), the number of cases per group is small (e.g., 5-20), and the proportion of events is low (Moineddin, Matheson & Glazier, 2007).

Hox (2002; 2010; Hox, Moerbeek, & van de Schoot, 2018) and McNeish and Stapelton (2016) provide the best overviews of sample size issues with regard to minimum sample sizes needed. Under most conditions, fixed effects and their standard errors are unbiased. With fewer than 5 cases per group and fewer than 50 groups, standard errors for fixed effects will be too small (increased Type I errors), and random effects (variances) and their standard errors may be underestimated. More recent evidence suggests that, with Kenward-Roger corrections (based on Satherthwaite degrees of freedom and adjustments to standard errors) or the bias reduced linearization (BRL; Bell & McCaffrey, 2002; aka cluster robust or CR2; Huang & Li, 2022; Huang, Wiederman, & Zhang, 2023), this bias in fixed effects tests can largely be addressed for a smaller number of clusters, particularly when there are more units within each cluster (Elff, Heisig, Schaeffer, & Shikano, Huang & Li, 2022; Hox & McNeish, 2020; McNeish, 2017). Thus, it seems prudent to make sure to use the Kenward-Roger or the CR2 degrees of freedom option[1] whenever there are less than 50 clusters.[2] Bayesian estimation may also be a potential solution when the number of clusters is small, although its advantages over other methods depends on the choice of priors (Bolin, Finch, & Stenger, 2019; McNeish, 2016b) and may be limited to less complex models with few random effects (Elff et al., 2021; Yamamoto & Miyazaki, 2024).

The poor estimates obtained with small group sizes may be offset by a very large number of groups (e.g., 450 or more; Theall, et al., 2011). Maas and Hox (2004, 2005) distinguish between fixed effects tests and random effects test with respect to sample size requirements for adequate significance tests. Their results suggest that for variance tests, 100 or more groups will be needed to achieve nominal alpha levels. In their more recent review, McNeish and Stapleton (2016) came to the conclusion based on the Maas and Hox findings and others, stating "a minimum of 50 clusters with a cluster size of 50 are suggested with 100 clusters being a more conservative figure, especially if FML estimation is utilized" (p.304). In general, REML should be the preferred method with smaller sample sizes (e.g., < 100 or higher), however.

Models with noncontinuous outcomes seem to require additional groups to avoid convergence problems and for adequate power (Austin, 2010; Bauer & Sterbin, 2011; Moineddin et al., 2007; Paccagnella, 2011). Moineddin and colleagues found fixed effects statistical tests were generally correct for adaptive quadrature estimation if there are 30 cases per group and 30 groups or 50 groups and 5 cases per group. The Kenward-Roger df adjustment seems to help considerably with bias in statistical tests of fixed effects with small number of groups or cases per group (Bell, Ene, Smiley, & Schoeneberger, 2013; McNeish & Stapleton, 2014). PQL regression estimates have been notoriously poor (Breslow & Lin, 1995), but a study by McNeish (2106a) suggests that estimates using a restricted version of PQL (RPQL) may work well with

[1] Kenward Roger degrees of freedom and standard errors are now available in SPSS or in R `lme4`. The CR2 corrections are available in R with the `MLMusingR` (which can be used with `lme4`) or the `clubsandwich` packages, https://cran.r-project.org/web/packages/clubSandwich/index.html.
[2] While the Kenward-Roger and biased reduced/cluster robust adjustments may be comparable in most circumstances, results from Zhang and Lai (2024) suggest that the cluster robust estimates were less affected by heteroscedasticity.

smaller number of groups and perform better than adaptive quadrature or Laplace methods in this case (see the handout "Multilevel Models with Binary and other Noncontinuous Dependent Variables" for more information on these estimators).

Standard penalized quasi-likelihood (PQL) variance estimates, however, appear to be unbiased in this range of groups and cases per group (Austin, 2010). Variance estimates for adaptive quadrature were generally biased when the ratio of groups to cases per group was below 30:30 and were still biased even with many more groups (200) if there are few cases per group (5-10; Clarke, 2008). Statistical problems are likely to be worsened with unbalanced n_j (Moineddin, Matheson, & Glazier, 2007), lower event frequency, and more complex models with more random effects, but more simulation work is needed on the variety of conditions that might exist in practice.

Power

The second important issue has to do with whether or not there is sufficient statistical power to find significance. One should consider power with regard to the particular hypotheses of interest. Fixed effects, in general, require fewer cases to have sufficient power, with "main effects" requiring fewer cases than cross-level interactions. Random effects generally require more cases for sufficient power. Significance tests of intercept variances require fewer cases than variance tests of random slopes (reliability is generally lower for slopes). Because cross-level interactions involve predicting variability of slopes, the sample size requirements for adequate power of tests of cross-level interactions may even surpass what is needed for adequate power of tests of random effects. Cross-level interactions appear to require more cases in each group as well (Mathieu, Aguinis, Culpepper, & Chen, 2012).

Although some authors have suggested a minimum of 100 groups with 10 cases per group is needed for sufficient power to test fixed effects (Kreft, 1996), Hox (2010) concludes that 50 groups with 5 cases per group may be sufficient.[3] For random effects (variances) and cross-level interactions, 100 to 200 groups with approximately 10 cases per group is likely to be needed for sufficient power to test these effects. But these are general recommendations, and power depends on several factors.

Naturally, the best way to plan for the appropriate sample size is to compute power estimates. Hox (2010), Snijders and Bosker (2012), and Scherbaum and Ferreter (2009) have the best coverage of power computations. I do not present any power computation examples here, but Snijders and Bosker (2012) and Hox (2010) illustrate with several examples. I typically rely on software for computer power estimates. The PinT, https://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT, and the Optimal Design (Spybrook, Raudenbush, Congdon, & Martinez, 2011), https://wtgrantfoundation.org/optimal-design-with-empirical-information-od, programs are free online. I find the Optimal Design program to be the most user friendly and flexible for power computations.

Growth curves and power

Power is too infrequently discussed in the context of longitudinal growth curves, but the principles described above are likely to translate into the longitudinal application of multilevel models as well. That is, although one really only needs 3 time points (or even 2) for growth curve models theoretically, practically speaking, there may be issues with convergence, bias in random effects tests, or power issues with fewer than 5 cases. Raudenbush (2008) makes the point that power in growth curve models depends on spacing between intervals and intraclass correlation coefficients in addition to the number of time points. Greater spacing (increased variance on X) and higher intraclass correlation coefficients (associated with more reliable estimates) will both increase power. Curvilinear models may need far more time points for sufficient power. Simulations by Muthén and Curran (1997) and power analyses I have conducted for my research suggest that 3 or 4 time points are nowhere near sufficient for power to test random effects for curvilinear coefficients (Diallo, Morin, & Parker, 2014). If curvilinear models are of interest, plan for perhaps twice as

[3] Note also that the impact of degrees of freedom and standard error corrections (e.g., KR or BRL/CR2) on power has not been studied.

many cases and time points and a very rough guide.

Power Analysis
Power analyses can be conducted to determine whether an analysis already completed had sufficient power to find significance (sometimes referred to as *post hoc* power analysis) or it can be conducted when planning a study (*a priori*). The latter use of power analysis is by far the most common use of power analysis, so I will focus on that here. Typically a researcher is interested in determining whether a given sample size from an existing study will have sufficient power or in determining a sample size that will have sufficient power. In either case, one needs to know the effect size expected, but a range of standard values can always be used to obtain a range of power estimates or sample sizes.

Effect Size. The effect size could be calculated from prior research in a related area using similar measures or a range of effect sizes may be used to estimate power at each sample size. When I have conducted power analyses, I have usually just used a range of effect sizes and calculate power with varying assumptions for each effect size. Although authors sometimes discuss unstandardized effect size estimates, it is often more convenient to work with standardized effect size estimates. The formula for calculating the standardized effect for nested data is (Raudenbush & Liu, 2000):

$$\delta = \frac{\left|\gamma_{qq}\right|}{\sqrt{\tau_q^2 + \sigma^2}}$$

where δ is the standardized effect, γ_{qq} is a fixed effect estimate, τ_q^2 is the variance estimate of that parameter, and σ^2 is the within-group variance. Thus, one can use this formula to estimate the effect size for the intercept using γ_{00} and τ_0^2 or the slope using γ_{10} and τ_1^2, for instance. Raudenbush and Liu (2000) suggest a standardized effect size of .2 represents a small effect, .5 represents a medium effect, and .8 represents a large effect for the fixed effects. For random effects, they suggest .05, .10, and .15 should be used for small, medium, and large effect sizes (based on variance values for a standard normal variable). Note that power may differ considerably for a level-2 predictor because the design effect will tend to be much larger and this leads to lower power. Dziak and colleagues illustrate this difference in the context of multisite trials in which the intervention is applied at the group rather than the individual level (Dziak, Nahum-Shani, & Collins, 2012). Raudenbush and Liu (2001) give a slightly different formula for estimating standardized effects for growth curve models:

$$\delta = \frac{\left|\gamma_{qq}\right|}{\sqrt{\tau_q^2}}$$

With values defined as above, except that the within-group variance is omitted.[4]

Noncentrality parameter. Noncentrality parameter, which can be viewed as an estimate of the degree of difference the true alternative hypothesis value is from the null hypothesis value, illustrates what factors power depends on (Spybrook, et al., 2011, p. 33):

$$\lambda = \frac{N\delta^2}{\sigma_\delta^2 + 4/n_j}$$

N is the number of groups and σ_δ^2 is the effect size variability, which is essentially an estimate of the variance of the slope taking into account sampling variability, defined as:

$$\sigma_\delta^2 = \frac{\tau_1^2}{\sigma^2}$$

[4]The authors omit the within-person variance based on a rationale that the slope is not impacted by measurement error and estimates "true change." (p. 391, Footnote 4)

Standard errors. Because power of the Wald ratio is a function of the standard error (smaller standard errors lead to more power), it is instructive to look at the standard error formula, here for the slope, γ_{10}.

$$S.E.\gamma_{10} = \sqrt{\frac{n_j\tau_1^2 + \sigma^2}{n_j N}}$$

And can be restated in terms of the ICC (or also in terms of the design effect).

$$S.E.\gamma_{qq} = \sqrt{\frac{4\left(\rho + (1-\rho)/n_j\right)}{N}}$$

In the above equations, γ_{10} is the fixed effect for the level-1 predictor, ρ is the ICC, n_j is the number of cases per group, and N is the number groups (Raudenbush & Liu, 2000). Standard errors will increase and power will decrease for larger ICC and design effects. The standard error estimate for random effect, τ_{00}, is:

$$S.E._{\tau_{00}} = \sqrt{\text{var}\left(\tau_0^2\right)} = \sqrt{\frac{2\sigma^4}{n_j N}\left(\frac{1}{n_j - 1} + 2\left(\frac{\tau_0^2}{\sigma^2}\right) + n\left(\frac{\tau_0^2}{\sigma^2}\right)^2\right)}$$

The formula hinges largely on the ratio τ_0^2/σ^2 and can also be stated in terms of the ICC.

$$S.E._{\tau_0^2} = \sqrt{\text{var}\left(\tau_0^2\right)} = \sqrt{\frac{2\sigma^4}{n_j N}\left(\frac{1}{n_j - 1} + 2\left(\frac{\rho}{1-\rho}\right) + n\left(\frac{\rho}{1-\rho}\right)^2\right)}$$

Within-group variance is nearly always larger than the between-group variance, usually many times larger.

Snijders and Bosker approach to power computation. Snijders and Bosker (1995, 1999, 2012) explain power analyses in terms of standard errors in reference to the standard normal distribution (see also Scherbaum & Ferreter, 2009, for a nice overview). Using z-values from the normal distribution, their formula can be used to estimate the standard error for a standardized effect.

$$standard\ error \leq \frac{effect\ size}{Z_{1-\alpha/2} + Z_{1-\beta}}$$

With a little algebra, one can also estimate the power using values of the effect size and standard error using 1 - probability value associated with the calculated z-value for the power estimate.

$$Z_{1-\beta} \leq \frac{effect\ size}{standard\ error} - Z_{1-\alpha/2}$$

Optimal Design software (Spybrook et al., 2011, link given above) focuses on clustered randomized trials (i.e., experimental studies in which data are nested), so the focus is on the comparison of two groups and testing of a regression coefficient that represents the difference between two groups. Remember that with a binary predictor the regression coefficient is the difference between two groups:

$$\beta_{1j} = \overline{Y}_{E_J} - \overline{Y}_{C_J}$$

where $\overline{Y}_{E_J}$ is the mean of the experimental group and $\overline{Y}_{C_J}$ is the mean of the control group.

Optimal study design: Efficient allocation of resources. Authors writing about multilevel power analysis frequently focus on the costs of sampling multiple groups (or sites in a clinical trial) vs. sampling more cases per site (e.g., Raudenbush & Lui, 2000; Snijders & Bosker, 1999). The number of groups seems to have a more dramatic effect on power after a certain minimum group size has been achieved. Raudenbush and Liu summarize the contribution to the total cost with the following formula:

$$T \geq \left(C_1 n_j + C\right) N$$

Where *C* is the cost of sampling an additional site (or group group) and C_1 is the cost of sampling an additional individual within a site. Total cost then is a function of the sum of the contribution of these the site cost and the costs of all individuals within a site multiplied by the total number of sites. The researcher can then estimate feasible sample size by integrating the costs and the power estimation.

References

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *The international journal of biostatistics*, *6*(1), 1-18.

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

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological methods, 16(*4), 373-390.

Bell, B., Ene, M., Smiley, W., & Schoeneberger, J. (2013). *A multilevel primer using SAS Proc Mixed, SAS Global Forum*.

Bolin, J. H., Finch, W. H., & Stenger, R. (2019). Estimation of random coefficient multilevel models in the context of small numbers of level 2 clusters. *Educational and Psychological Measurement, 79*(2), 217-248.

Breslow, N., & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika, 82*, 81-91.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single level models with sparse data. *Journal of Epidemiology and Community Health, 62*, 752–758.

Diallo, T. M., Morin, A. J., & Parker, P. D. (2014). Statistical power of latent growth curve models to detect quadratic growth. *Behavior Research Methods, 46*, 357-371.

Dziak, J.J., Nahum-Shani, I., & Collins, L.M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods, 17*, 153-175.

Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2021). Multilevel analysis with few clusters: Improving likelihood-based methods to provide unbiased estimates and accurate inference. *British Journal of Political Science, 51*(1), 412-426.

Hox, J. (2002). *Multilevel analysis. Techniques and applications* . Mahwah, NJ: Lawrence Erlbaum

Hox, J. (2010). *Multilevel analysis: Techniques and applications, second edition.* New York: Routledge.

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

Hox, J.J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis. Techniques and applications, third edition.* New York: Routledge.

Hox, J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications, third edition.* New York: Routledge.

Huang, F. L., & Li, X. (2022). Using cluster-robust standard errors when analyzing group-randomized trials with few clusters. *Behavior Research Methods, 54*, 1181–1199 1-19.

Huang, F. L., Wiedermann, W., & Zhang, B. (2022). Accounting for heteroskedasticity resulting from between-group differences in multilevel models. *Multivariate Behavioral Research, 58*(3), 637-657.

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

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127-137.Mathieu, J., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Improving the accuracy of inferences about cross-level interaction tests in random coefficient modeling. *Journal of Applied Psychology, 97*, 951-966

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.

McNeish, D. (2016a). Estimation methods for mixed logistic models with few clusters. *Multivariate Behavioral Research*, *51*(6), 790-804.

McNeish, D. (2016b). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750-773.

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate behavioral research, 52*(5), 661-670.

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(*2), 295-314.

Moineddin, R., Matheson, F., & Glazier, R. (2007). A simulation study of sample size for multilevel logistic regression models. BMC Medical Research Methodology, 7, 34.

Muthen, B.O., & Curran, P.J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371-402.

Raudenbush, S.W. (2008). Many Small Groups. In J. de Leeuw & E. Meijer (Eds.), *Handbook of Multilevel Analysis.* (pp. 207-236). New York, NY. Springer.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199–213.

Raudenbush, S.W., & Liu, X. (2001). Effects of Study Duration, Frequency of Observation, and Sample Size on
Power in Studies of Group Differences in Polynomial Change. *Psychological Methods, 6(4*), 387.401.

Scherbaum, C.A. & Ferreter, J.M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*, 347-367.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Spybrook, Raudenbush, Congdon, & Martinez (2011). Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0.

Theall, K.P., Scribner, R., Broyles, S, Yu, Q., Chotalia, J., Simonsen, N., Schonlau, M., & Carlin, B.P. (2011). Impact of small group size on neighbourhood influences in multilevel models. Journal of Epidemiology and Community Health, 65, 688–695.

Yamamoto, Y., & Miyazaki, Y. (2024). Small Sample Methods in Multilevel Analysis. *The Journal of Experimental Education*, 1-32.

Zhang, Y., & Lai, M. H. (2024). Evaluating two small-sample corrections for fixed-effects standard errors and inferences in multilevel models with heteroscedastic, unbalanced, clustered data. Behavior Research Methods, 1-17.