Lorah *Large-scale Assess Educ (2018) 6:8*
https://doi.org/10.1186/s40536-018-0061-2

Large-scale Assessments in Education

METHODOLOGY Open Access

# Effect size measures for multilevel models: definition, interpretation, and TIMSS example

Julie Lorah*

*Correspondence:
jlorah@iu.edu
Indiana University, Bloomington, USA

**Abstract**

Effect size reporting is crucial for interpretation of applied research results and for conducting meta-analysis. However, clear guidelines for reporting effect size in multilevel models have not been provided. This report suggests and demonstrates appropriate effect size measures including the ICC for random effects and standardized regression coefficients or $f^2$ for fixed effects. Following this, complexities associated with reporting $R^2$ as an effect size measure are explored, as well as appropriate effect size measures for more complex models including the three-level model and the random slopes model. An example using TIMSS data is provided.

**Keywords:** Multilevel model, Effect size, TIMSS

## Background

The reporting of effect sizes in quantitative research lends interpretability and practical significance to findings and provides comparability across studies (Kelley and Preacher 2012). Further, providing effect sizes in a study can aid future researchers in conducting meta-analysis to synthesize findings from multiple studies (Denson and Seltzer 2011) and power analysis to plan future studies (Kelley and Preacher 2012). Effect size has been defined in various ways, but recent work considers it "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelley and Preacher 2012). More specifically, measures of effect size can broadly be categorized as either measures of variance explained (Tabachnick and Fidell 2007) or measures of standardized effect size (Snijders and Bosker 2012).

Because of these important benefits, organizations such as the American Psychological Association (APA) and American Educational Research Association (AERA) are moving toward stronger language requiring the reporting of effect sizes (Kelley and Preacher 2012; Peng and Chen 2013). Many educational and psychological journals are also following suit in requiring effect sizes in tandem with a de-emphasis on null hypothesis significance testing (NHST; Kelley and Preacher 2012). In response to these guidelines, researchers have found that the prevalence of effect size reporting has increased, although there is room for improvement in both prevalence as well as quality of reporting (Peng and Chen 2013). Although the importance and usefulness of effect size reporting is clear, guidance regarding effect size measures for multilevel models is scarce. Further, multilevel models may be particularly relevant in cross-cultural educational research using international datasets due to the nesting of data (i.e. students within schools within countries, etc.).

This paper provides guidance regarding choice and interpretation of effect size measures for multilevel models. A demonstration using large-scale survey data is provided for each topic. Assessing effect size for random effects is demonstrated using the ICC. Following this, assessing effect size for fixed effects is demonstrated using standardized regression coefficients and $f^2$. Lastly, complexities associated with additional topics, including three-level models, $R^2$ as a measure of variance explained, and models with random slopes will be explored.

## Example analysis

An example is provided based on a multilevel model estimated from IEA's Trends in International Mathematics and Science Study (TIMSS) 2011 fourth grade mathematics data. Note that although efforts were made to ensure the data and models estimated are substantively relevant and realistic, this applied analysis is primarily for demonstration purposes. Ten countries were randomly selected (Bahrain, Czech Republic, Denmark, Iran, New Zealand, Norway, Slovak Republic, Spain, Sweden, Tunisia) and used as the level-2 unit. All data preparation and analysis was done in R (R Core Team 2014) and multilevel models were estimated with the lme4 package (Bates et al. 2015) unless specified otherwise. R syntax for each model is provided in the appendix.

The unit of analysis is students (level 1) and the nesting variable is country (level 2). The sample included a total of 46,475 students nested within 10 countries. The largest country sample size was 5760 and smallest was 3121 with an average country size of 4648. In addition, demonstrations involving 3-level models included school membership as a level. The sample included 1817 schools. The largest within-school sample size was 93 and the smallest was 2 with an average school size of 25.6. Missing data was addressed with default listwise deletion resulting in a final analytic sample of 44,800 students.

The outcome variable is mathematics achievement. TIMSS provides five plausible values, which are multiple imputations of the latent construct (Wu 2005), for this variable. Only the first plausible value was used for analysis (ASMMAT01). Using only one plausible value is not preferred compared to using all five plausible value (Rogers and Stoeckel 2008); however, since using only one plausible value has been shown to typically recover population parameters (Rogers and Stoeckel 2008; Wu 2005) and since the present analysis is primarily for demonstration purposes, use of only one plausible is used simply to avoid overwhelming the reader with many new topics at once. Further, it should be noted that it is inappropriate to use an average of plausible values for analysis, which was therefore not done in the present analysis (Rogers and Stoeckel 2008). The mean for math achievement was 475.18 and the standard deviation was 93.77.

Independent variables include Female (ITSEX; recoded 0 = boy, 1 = girl); whether the student has an internet connection at home (ASBG05E; recoded 0 = no, 1 = yes); and the students' confidence with math (ASBGSCM, continuous). Overall, the sample was 50% female and 78% of students had an internet connection at home. The mean confidence value was 10.1 and standard deviation was 1.92, although this variable was standardized for most analyses.

For this analysis, the following random intercept models were estimated:

$$Math_{ij} = \beta_0 + u_{0j} + \varepsilon_{ij} \tag{1}$$

$$Math_{ij} = \beta_0 + \beta_1 * Female_{ij} + \beta_2 * Internet_{ij} + \beta_3 * Confidence_{ij} + u_{0j} + \varepsilon_{ij} \tag{2}$$

where $Math_{ij}$ is the outcome for student i within country j; $\beta_0$ is the intercept; $u_{0j}$ is random error at level 2 with estimated variance $\tau^2$; $e_{ij}$ is random error at level 1 with estimated variance $\sigma^2$; all other $\beta$ are slope coefficients.

Although a detailed description of sufficient sample size for multilevel modeling is beyond the scope of this paper, it should be briefly mentioned that researchers differ in their recommendations. For example, one rule of thumb, the 30/30 rule recommends a minimum of 30 groups, with at least 30 members in each group (Hox 2010) whereas other researchers suggest that with as few as 10 groups, modeling with random rather than fixed effects is appropriate (Snijders and Bosker 2012). Another simulation study specifically recommends at least 50 groups to avoid bias is certain parameter estimates (Maas and Hox 2005). Ten groups were used for the present study, consistent with researcher recommendations (Snijders and Bosker 2012). The primary purpose of analyses in the present study is demonstration; however, applied researchers should be cognizant of differences in the probability of biased parameter estimates and standard errors at various sample sizes.

## Random effects

With a multilevel model, the intercept is allowed to vary for each level 2 unit. Within the unconditional model (Eq. 1), the intercept represents the mean math achievement value and the model serves to partition the variance between level 1 and level 2 units. The results can be summarized by the intraclass correlation coefficient (ICC) given as:

$$\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{3}$$

where $\tau^2$ is the between-cluster variance (variance of $u_{0j}$) and $\sigma^2$ is the within-cluster variance (variance of $e_{ij}$; Snijders and Bosker 2012). The ICC can be interpreted as the proportion of variance in the outcome accounted for by the level 2 unit (cluster) membership (Hox 2010; Kirk 2013; Snijders and Bosker 2012) and represents a measure of strength of association (Kirk 2013) since it represents a proportion of variance. In addition, the ICC can be interpreted as the expected correlation between two randomly drawn level 1 units within a given randomly drawn level 2 unit (Hox 2010; Snijders and Bosker 2012) and since the magnitude of this measure can be interpreted in the same way as a correlation coefficient, it represents an effect size index (Snijders and Bosker 2012). The ICC can be interpreted in the context of previous research findings. For example, Hedges and Hedberg (2007) report the average ICC for K-12 academic achievement is about 0.22 for students nested within schools. Although the present demonstration considers students nested within countries and the comparison with ICC = 0.22 may not be particularly helpful; for studies examining achievement of students within schools within the United States, this type of reference may be a useful point of comparison.

For the example analysis, the ICC was 0.27 (computed from Eq. 3 based on estimates from Eq. 1) indicating that the proportion of variance in math scores explained by country membership is 0.27. It is unclear what a typical ICC for achievement might be when considering nesting of students within countries, but understanding the extent to which countries differ may be an important first step for further investigating differences in academic achievement between countries.

## Fixed effects

Fixed effects, such as intercept or slope coefficients, depend on the scale of the independent variable, and so are not comparable across studies or among multiple variables within a single study. Some researchers suggest Cohen's *d*, which is a measure of the standardized mean difference between two categories in a binary variable, as a measure of effect size for a binary covariate in a multilevel model (Snijders and Bosker 2012; Spybrook 2008) and by analogy, one could imagine representing the effect size for a continuous covariate as the correlation between that covariate and the outcome variable (for example, between confidence and math achievement in the TIMSS example). However, if Cohen's *d* or a bivariate correlation coefficient were to be simply computed (i.e. based on the one-level model with no covariates), it would not be recommended as an effect size measure since it represents the relationship between the two variables without controlling for level-2 unit membership and other associated covariates.

Instead, the standardized coefficients can be used (Ferron et al. 2008; Snijders and Bosker 2012). These can be obtained by standardizing (i.e. M = 0; SD = 1) each variable before analysis (Ferron et al. 2008) or by standardizing each regression coefficient by multiplying it by the standard deviation of X and dividing by the standard deviation of Y (Snijders and Bosker 2012).

For the example analysis, Eq. 2 was estimated, and the standardized coefficient for Female is 0.004 ($p > 0.05$); for Internet is 0.186 ($p < 0.05$) and for Confidence is 0.262 ($p < 0.05$). This indicates that the coefficient for Female is not significantly different from zero; that one standard deviation increase in the Internet variable is related to 0.186 expected standard deviations increase in math achievement; and that one standard deviation increase in the Confidence variable is related to 0.262 expected standard deviations increase in math achievement, controlling for associated covariates.

Although these measures are now comparable, the interpretation for binary covariates, such as Female, may still be somewhat difficult; instead, the researcher can dummy code the binary covariate and standardize the outcome variable resulting in partially standardized coefficients. For the example analysis, this results in a coefficient of 0.008 ($p > 0.05$) for Female and 0.45 ($p < 0.05$) for Internet indicating that females are not significantly different from males on math achievement, when controlling for associated measures, and that students with internet connection are expected on average to score 0.45 standard deviations higher on math achievement, controlling for associated measures.

Both standardized and partially standardized coefficients provide information about the magnitude of the effect (after controlling for other covariates and nesting) and these measures are generally comparable amongst themselves and across studies with similar populations (and analogously similar analytic samples). One limitation is that these measures are dependent on the sample standard deviation of each variable, which will be particular to the given sample, and may vary from sample to sample. When working with large samples however, this sampling variability, or sample to sample variation, should be quite small meaning that comparability should typically not be a problem.

Another possibility for effect size of a given fixed effect is (Aiken and West 1991):

$$f^2 = \frac{R_2^2 - R_1^2}{1 - R_2^2} \tag{4}$$

where $R_2^2$ represents the variance explained for a model with the given effect and $R_1^2$ represents the variance explained for a model without the given effect and the measure can be interpreted as the proportion of variance explained by the given effect relative to the proportion of outcome variance unexplained (Aiken and West 1991) and is considered small at a value of 0.02, medium at a value of 0.15, and large at a value of 0.35 (Cohen 1992). In the present example, $f^2$ for Confidence is 0.07; for Internet is 0.14; and for Female is $< 0.01$, indicating a small-medium effect, a medium effect, and a negligible effect, respectively. These results indicate that confidence explains about 7% of the variance in math scores relative to unexplained variance and internet access explains about 14% of variance in math scores relative to unexplained variance. Note that computation of $R^2$ is covered in a proceeding section of the present study.

## Variance explained

Variance explained for a multilevel model is more complex compared with a single level model, since there are now multiple residual terms. A formula for $R^2$ specific to multilevel models is provided by Snijders and Bosker (2012) and represents proportional reduction in prediction error at the individual level:

$$R^2 = 1 - \frac{\sigma_F^2 + \tau_F^2}{\sigma_E^2 + \tau_E^2} \tag{5}$$

where $\sigma_F^2$ represents the level-one random error variance (variance of $e_{ij}$) for the full model (i.e. the model of interest); $\tau_F^2$ represents the level-two random error variance (variance of $u_{0j}$) for the full model; $\sigma_E^2$ represents the level-one random error variance for the empty model; and $\tau_E^2$ represents the level-two random error variance for the empty model.

In the present study, variance components based on the empty model (Eq. 1) and the full model (Eq. 2) were estimated and used to compute $R^2 = 0.19$ which can be interpreted as the proportion of variance in math achievement explained by the covariates (female, internet, confidence). The effect size measure related to variance explained for the overall model is $f^2$ which can be computed as (Cohen 1992):

$$f^2 = \frac{R^2}{1 - R^2} \tag{6}$$

In the present study, $f^2 = 0.23$ for the overall model indicating female, internet access, and confidence explain 23% of variance in math achievement relative to the unexplained variance in math achievement. Guidelines for interpretation of $f^2$ indicate that 0.02 is a small effect, 0.15 is a medium effect, and 0.35 is a large effect (Cohen 1992), indicating that the present effect is medium to large.

## Three-level models

Three-level random intercept models include an additional hierarchically nested level (Snijders and Bosker 2012). In the present example, another level for school membership is added, implying a structure of students nested within schools nested within countries. Effect size measures for fixed effects used with a standard two-level multilevel model can be used analogously in a three-level model. The three-level model, however, implies additional random effects, so although ICC can still be used as an effect size measure, multiple different ICC statistics are defined for this model. The empty model can now be defined in the present example as:

$$Math_{ijk} = \beta_0 + v_{00k} + u_{0jk} + \varepsilon_{ijk}. \tag{7}$$

For student i within school j within country k where $\beta_0$ is the overall intercept, and the remaining terms are random error terms at the country, school, and individual levels respectively such that $Var(v_{00k}) = \phi^2$, $Var(u_{0jk}) = \tau^2$, $Var(e_{ijk}) = \sigma^2$. Note that while the standard two-level model partitions variance between level 1 and level 2, the present model is an extension of this where the total variance is partitioned among three levels. Three ICC measures can now be defined as follows (Hox 2010):

$$\text{ICC, L3} = \frac{\phi^2}{\phi^2 + \tau^2 + \sigma^2} \tag{8}$$

$$\text{ICC, L2.1} = \frac{\tau^2}{\phi^2 + \tau^2 + \sigma^2} \tag{9}$$

$$\text{ICC, L2.2} = \frac{\tau^2 + \phi^2}{\phi^2 + \tau^2 + \sigma^2} \tag{10}$$

In the present example with a random effect for school membership added, there are 46,475 students nested within 1817 schools nested within 10 countries. Based on random effect estimates from the empty model (Eq. 7: $\phi^2 = 2376$; $\tau^2 = 1906$; $\sigma^2 = 4778$), all three ICC values were computed as measures of effect size for random effects. Results indicate that ICC at level 3 (Eq. 8) is 0.26, the first version of ICC at level 2 (Eq. 9) is 0.21 and the second version of ICC at level 2 (Eq. 10) is 0.47. Each of these three measures provides different information and is interpreted in a different way. The present results indicate that 26% of variance in math scores is accounted at the country level, 21% is accounted at the school level; and 47% at the level of schools nested within countries.

## Random slopes

The multilevel model implies the addition of a random effect to allow the intercept to vary randomly by level 2 unit (this describes the random intercept model that has been examined thus far). Additionally, this model can further be generalized to allow the slope for a given level 1 covariate to vary randomly by level 2 unit resulting in a random slopes model (Snijders and Bosker 2012). This model could be thought of as including an interaction effect between the level 1 covariate and the random effect (level 2 unit). In the present analysis, a random slope for Internet was added, and all predictors were standardized (M = 0, SD = 1) for ease of interpretation, resulting in the following model:

$$Math_{ij} = \beta_0 + \beta_1 * Female_{ij} + \beta_2 * Internet_{ij} + \beta_3 * Confidence_{ij} + u_{1j} * Internet_{ij} + u_{0j} + \varepsilon_{ij} \quad (11)$$

Four random effects will be estimated: $Var(e_{ij}) = \sigma^2$; $Var(U_{0j}) = \tau_0^2$; $Var(U_{1j}) = \tau_1^2$; $Cov(U_{0j}, U_{1j}) = \tau_{01}$. Interpretation can proceed in multiple ways. First, since the model assumes that the distribution of $U_{1j}$ terms (the difference between the overall slope and group-specific slope for each group) is normally distributed with variance $\tau_1^2$, the range within which 95% of level 2 units would fall can be computed (Snijders and Bosker 2012). In the present example, the average slope for internet ($\beta_2$) was estimated as 0.17 ($t = 8.77$, $p < 0.05$) and $\tau_1^2 = 0.003$ (SD = 0.06). So a hypothetical country with a "high" slope could be computed by starting with the average slope and adding two times the square root of $\tau_1^2$. Analogously, a hypothetical country with a "low" slope could be computed similarly, but by subtracting two times the square root of $\tau_1^2$. In the present example, this would imply that 95% of countries would have a slope for Internet between 0.05 and 0.29. Since these represent standardized coefficients, these slopes can be interpreted in the same way as fixed effects based on standardized covariates. Since there is no clear effect size measure to aid in interpretation of the random slope, this interpretation may offer a helpful alternative in the spirit of presenting the scope of the effect.

Researchers may also want to interpret the covariance term, $\tau_{01}$. As with any covariance, this term can be standardized to report the correlation and interpreted according to effect size criteria for *r*: 0.10 is small; 0.30 is medium; and 0.50 is large (Cohen 1992). In the present example, the correlation is 0.07 indicating that countries with higher slopes tend to have slightly higher intercepts, but that this relationship is fairly weak, according to Cohen's (1992) criteria. However, caution regarding the intercept interpretation should be applied. The intercept will be specific only to the case where all predictors (X) are equal to zero (Snijders and Bosker 2012). In the present example, this implies that the relationship between Internet and Math is stronger for countries with higher average math scores for average Internet connectivity, although the effect is relatively small. If Internet had not been grand mean centered, the intercept-slope covariance term would have taken a different value based on the fact that the intercept would take a different meaning. Thus, centering must be carefully considered within the context of a random slopes model due to the fact that interpretation of the intercept-slope covariance parameter depends on how covariates are centered. Centering does not, however, necessarily directly contribute to issues of comparability across studies.

Since the effect sizes based on computing hypothetical fixed effects and standardizing the covariance term may still be difficult to interpret, and since guidance regarding

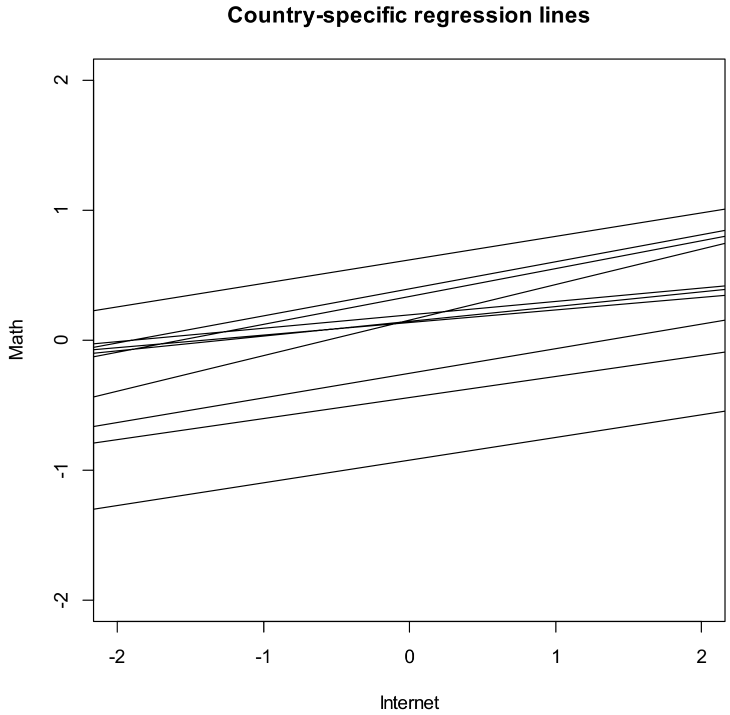

**Fig. 1** Demonstration of random slopes

**Table 1 Model comparison for random slopes**

| Model | df | AIC | BIC | LogLik | Deviance | $\chi^2$ | $\chi^2$ df | p |
|---|---|---|---|---|---|---|---|---|
| No random slope | 6 | 106,446 | 106,498 | − 53,217 | 106,434 | | | |
| With random slope | 8 | 106,368 | 106,437 | − 53,176 | 106,352 | 82.498 | 2 | < 0.001*** |

interaction effect interpretation emphasizes the importance of plotting results (Aiken and West 1991), this guidance is applied here in the context of random slopes. Figure 1 shows the country-specific regression line regressing Math on Internet for each of the 10 countries in the dataset (based on Eq. 11). For a larger number of level 2 units, a random sample of these units can be plotted so that the lines remain distinct. For the plot, it can be seen that some countries display a weak relationship between individual internet connectivity and math achievement whereas other countries display a slightly stronger relationship.

It is clear from the plot that if Internet were recoded with a different substantive meaning for the value zero, the intercept variance could change, as well as the intercept-slope covariance. For example, if 2 was subtracted from each value for Internet, the y-intercept would be further right on the plot (Fig. 1) at the location where Internet = 2. In that case, the y-intercept for each country would be different and the variability of these intercepts ($\tau_0^2$) as well as the correlation between these intercepts and the slopes ($\tau_{01}$) would therefore differ as well. As a researcher interprets and attempts to explore the scope of these effects, their conditional nature should be considered and emphasized accordingly.

Significance (although not necessarily effect size) can be assessed through model comparison procedures, which are demonstrated here to clarify the limitations regarding effect size reporting. Estimating a model with and without the random effect for slope can provide insight into whether the slope should be allowed to vary randomly. For the present study, Table 1 shows the comparison between the model with and without random slopes (Eqs. 2, 11).

Lower BIC indicates a better fit, and a difference of greater than 10 indicates "very strong" evidence for the more complex model (Raftery 1995) which is provided in the present example. Although this model comparison can aid in selecting a model, and even possibly assess the strength of that evidence (i.e. BIC), the evidence cannot be used as a measure of effect size. Each of the model fit indices (AIC, BIC, and log-likelihood) are a function of deviance which is, among other things, a function of sample size (McCoach and Black 2008) and so therefore would not be appropriate as measures of effect size.

Measures of variance explained are also generally inappropriate for random slopes, because allowing the slope to differ for each level 2 unit does not necessarily explain additional variance. It can be noted that the scope of the effect of random slopes is represented by the variance of those slopes. Although there is not strictly a measure of standardized effect size or variance explained for variance terms, the square root of this variance (square root of $\tau_0^2$) is a standard deviation which is considered an interpretable measure of a distribution's spread (Darlington and Hayes 2017).

One other possibility would be to approximate the scope of the effect by modeling the level 2 unit as a fixed effect (if this is substantively plausible) and including the interaction between the covariate and each dummy indicator for group membership. Based on this conceptualization, any method appropriate for effect size of a set of fixed effects (i.e. the interaction terms) would be feasible.

## Complex survey designs

Many large-scale datasets involve complex sampling plans and other additional complexities which may need to be considered during analysis, including use of plausible values for achievement measures, inclusion of sampling weights due to non-equal probability of selection, and inclusion of replicate weights to account for multi-stage sampling. These aspects of analysis are beyond the scope of the present examination of effect size, but they should not be ignored. Several resources are available to learn more about these topics (Martin and Mullis 2012; Meinck and Vandenplas 2012; Rogers and Stoeckel 2008; Snijders and Bosker 2012; Wu 2005) and specific software is available to aid in analysis, such as the BIFIE package (BIFIE 2017) which is implemented in R.

## Conclusions

Reporting measures of effect size is a crucial part of interpretation for applied multilevel modeling studies. Researchers can use the ICC to represent the magnitude of random effects which could represent country and/or school effects and standardized regression coefficients or $f^2$ to represent the magnitude of fixed effects which may represent relationships of interest when examining substantive questions using international datasets.

Complexities associated with three-level models, reporting $R^2$, and random slopes have been explored. The topics in the present study have been demonstrated using TIMSS data, but the suggestions provided could be applied to any multilevel analysis of primary or secondary data.

### Authors' contributions

The author read and approved the final manuscript.

### Acknowledgements

The author would like thank Chao-Ying Joanne Peng and Patricia Martinkova for their comments on prior versions of this manuscript.

This research is based on work presented at the IEA International Research Conference (IRC) in June 2017.

### Competing interests

The author declares that she has no competing interests.

### Availability of data and materials

TIMSS data is publicly available from the IEA. R syntax used for the analyses in this paper is provided in Appendix.

### Funding

Not applicable

## Appendix

R Syntax

```
library(lme4) #package for lmer function used to estimate multilevel model
lmer(Math~(1|IDCNTRY),data=mydata) #equation 1
lmer(Math~Female+Internet+Confidence+(1|IDCNTRY),data=mydata) #equation 2
lmer(Math~(1|FullSchID) +(1|IDCNTRY),data=mydata) #equation 4
lmer(Math~Female+Internet+Confidence    +(1+Internet|IDCNTRY),data=mydata)
  #equation 10
```

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 2 January 2018 Accepted: 17 July 2018
Published online: 23 July 2018

### References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park: Sage.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

BIFIE. (2017). BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 1.13-24. https://CRAN.R-project.org/package=BIFIEsurvey. Accessed 29 Jan 2018.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. New York: Guilford Press.

Denson, N., & Seltzer, M. H. (2011). Meta-analysis in higher education: An illustrative example using hierarchical linear modeling. *Research in Higher Education, 52,* 215–244.

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Charlotte: Information Age Publishing Inc.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Great Britain: Routledge.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152. https://doi.org/10.1037/a0028086.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences*. Los Angeles: Sage.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Charlotte: Information Age Publishing Inc.

Meinck, S. & Vandenplas, C. (2012). *Sample size requirements in HLM: An empirical study*. IERI monograph series issues and methodologies in large-scale assessments. IER institute, special issue 1, educational testing service and international association for the evaluation of educational achievement.

Peng, C. Y. J., & Chen, L. T. (2013). Beyond Cohen's *d*: Alternative effect size measures for between-subject designs. *Journal of Experimental Education, 82,* 22–50. https://doi.org/10.1080/00220973.2012.745471.

R Core Team. (2014). R: A language and environment for statistical computing (Computer software), R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. Accessed 29 Jan 2018.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163.

Rogers, A. M., & Stoeckel, J. J. (2008). *NAEP 2008 arts: Music and visual arts restricted-use data files data companion (NCES 2011–470)*. Washington, D.C: US Department of Education Institute of Education Sciences, National Center for Education Statistics.

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

Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273–311). Charlotte: Information Age Publishing Inc.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31,* 114–128.