Journal of Modern Applied Statistical Methods

Volume 18 | Issue 2 Article 24

9-22-2020

Estimating a Multilevel Model with Complex Survey Data: Demonstration using TIMSS

Julie Lorah
Indiana University Bloomington, jlorah@iu.edu

Estimating a Multilevel Model with Complex Survey Data: Demonstration using TIMSS

Cover Page Footnote

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

Journal of Modern Applied Statistical Methods
November 2019, Vol. 18, No. 2, eP3155.
doi: 10.22237/jmasm/1604190360

בס"ד

ISSN 1538 − 9472

Estimating a Multilevel Model with Complex Survey Data: Demonstration using TIMSS

Julie Lorah
Indiana University Bloomington
Bloomington, IN

Analysis of complex survey data is demonstrated for the multilevel model. Description of specific aspects of analysis, including plausible values, sampling weights, and replicate weights is provided. Following this, example TIMSS data and models are described and results are presented.

Keywords: Multilevel model, complex survey data, complex sampling, TIMSS

Introduction

Multitudes of complex survey data is available that researchers can use to answer questions on various topics. Complex survey data are obtained through a more complex sampling plan involving, for example, cluster and/or stratified sampling (Skinner & Wakefield, 2017). A complete treatment of complex sampling techniques is beyond the scope of the present article, but several excellent resources exist (see, for example, Kalton, 1983; Lee et al., 1989; Lumley, 2010) as well as guidance specific to international large-scale assessments (see, for example, Rutkowski, Gonzalez, et al., 2010; Rutkowski, von Davier, & Rutkowski, 2013). Although there is the potential to learn much by analyzing this type of data, the analysis itself can be difficult, particularly for applied researchers who are not trained specifically in complex survey analysis (Skinner & Wakefield, 2017). The implementation within software packages for analyzing complex survey data has been slow (Skinner & Wakefield, 2017), additionally complicating the process.

The multilevel model represents a particularly well-suited model for analyzing complex survey data because it directly models different levels of data

doi: 10.22237/jmasm/1604190360 | Accepted: October 25, 2019; Published: September 22, 2020.
Correspondence: Julie Lorah, jlorah@iu.edu

that can correspond to a cluster sampling design, and as such, the multilevel model is frequently used for the analysis of complex survey data (Laukaityte & Wiberg, 2018). With a cluster sampling design, the sampling plan includes sampling of clusters from a population of clusters rather than random sampling from the population (Lumley, 2010). Clusters sampled at the first stage of sampling are called primary sampling units (PSU; Lumley, 2010). These clusters may violate the assumption of non-independence of observations, which is an assumption for many models, such as linear regression. Ignoring the non-independence is not recommended because it has the potential to severely inflate Type I error rates (Snijders & Bosker, 2012). Therefore, a method for accounting for non-independence is needed. Several options are available including using cluster membership as a fixed effects (i.e. dummy-coded predictors); replicate weights which represent a re-sampling method that can correctly estimate standard errors; multilevel models which directly model a cluster by adding a random error term at the cluster level; and generalized estimating equations for correct estimation of standard errors. When questions related to the connection of variables at multiple levels (such as students and schools) are investigated, multilevel models can be used (Snijders & Bosker, 2012) as well as generalized estimating equations approaches (Gardiner et al., 2009; Graubard & Korn, 1994; McNeish, 2019).

The multilevel model offers several advantages over the single-level model options. For applied researchers, the multilevel model is practical to implement due to the great number of resources available including its widespread inclusion in statistical software. Because multilevel models are commonly used in association with complex survey data (Laukaityte & Wiberg, 2018), audiences may be more familiar with these analyses. In addition, the model itself is extremely flexible, easily allowing for the inclusion of additional grouping variables (for example, with a three-level model); inclusion of cluster-level predictor variables (Laukaityte & Wiberg, 2017); inclusion of random slopes whereby the relationship between an individual-level variable and the outcome variable is allowed to vary randomly by group membership; and investigation of contextual effects by including group-average predictor variables. Further, by including a random effect for group membership, evidence related to the degree of nesting, for example by reporting the intraclass correlation coefficient, can be evaluated.

Estimation of the multilevel model with plausible values, sampling weights, and replicate weights added for analysis are considered here. Although guidance regarding analysis of complex survey data is available (for example, Lumley, 2010; Skinner & Wakefield, 2017), there is little guidance specific to the multilevel model and software options may be more difficult to find and implement. A notable

exceptions to this are two papers examining plausible values (Laukaityte & Wiberg, 2017) and sampling weights (Laukaityte & Wiberg, 2018) for multilevel models. The following treatment is intended primarily as a tutorial for applied researchers.

Plausible Values

When investigating achievement in a large population, it can be more efficient to use a matrix sampling design, where each subject responds to relatively few items, rather than creating long assessments for each participant. This matrix sampling procedure is used in several studies, including TIMSS. Although this design does not allow for making precise statements about individuals, it does allow for the more efficient estimation of population characteristics.

The implication of this design is that individual scores contain a large amount of uncertainty. In order to model this uncertainty, plausible values are used. Note that this score uncertainty may be due to matrix sampling designs and/or other source of uncertainty. The plausible values are often represented with 5 scores per student (although some datasets may include 20 scores per student; Laukaityte & Wiberg, 2017); each score representing a random draw from the student's posterior distribution which is a function of that student's item responses as well as background characteristics (Martin & Mullis, 2012). In other words, the plausible values represent multiple imputations of the latent construct (Wu, 2005).

The procedure to conduct analyses using a variable measured with plausible values is given by Martin and Mullis (2012, p. 5). First, the statistic of interest should be computed with each of M plausible values (for TIMSS 2011 $M = 5$). The formula for the imputation variance is given as $\text{Var,imp} = (1 + 1/M)*\text{Var}(t_1,\ldots, t_m)$. This can then be added to the sampling variance to find the correct standard error for the statistic. It should also be noted that it may be possible to recover population parameters based on only one plausible value (Rogers & Stoeckel, 2008; Wu, 2005) although this is not recommended. Further, it is important to know that plausible values should never be averaged for analysis (Rogers & Stoeckel, 2008). For a more in-depth treatment regarding use of plausible values for multilevel models with TIMSS, please see Laukaityte and Wiberg (2017).

Sampling Weights

TIMSS data also includes sampling weights to adjust for unequal probability of selection. Sampling weights are included in analysis to avoid bias; however, failure to model with sampling weights does not necessarily produce bias in parameter

estimates (Snijders & Bosker, 2012). Although guidance typically indicates that sampling weights should be included in all analyses conducted with a non-random sample, researchers still disagree as to whether or not and under what conditions sampling weights should be included (Snijders & Bosker, 2012). In addition to the impact on bias of point estimates, inclusion of sampling weights in large-scale assessment data has been shown to decrease sampling precision by about 10%, thereby slightly increasing standard errors (Meinck & Vanderplas, 2012). Sampling weights are available at multiple levels (for example, students and schools) and these weights additionally need to be scaled appropriately (Laukaityte & Wiberg, 2018). Scaling should be applied only for level 1 (student) weights; for a more in-depth discussion of sampling weights, scaling, and when sampling weights may impact results with multilevel models using TIMSS, see Laukaityte and Wiberg (2018).

With a multilevel model, these sampling weights can be included in the analysis and there are software options that can do this automatically for the researcher, such as the BIFIEsurvey package in R, which will be explored in the subsequent demonstration section (BIFIE, 2017). Note that when these weights are included in the likelihood, the estimation proceeds using a psudo-likelihood (Rabe-Hesketh & Skrondal, 2006). Other options for including sampling weights are available in R, including the WeMix package which allows inclusion of weights at every level of a multilevel model and the RStan package which allows R users to interface with the Bayesian analyses available in Stan.

Replicate Weights

Many datasets using complex survey designs include replicate weights, which can be used to adjust for cluster sampling and the implied non-independence of individual observations. Failure to account for non-independence of observations could induce downwardly biased standard errors which would inflate Type I error rates (Snijders & Bosker, 2012). Use of replicate weights essentially represents a resampling method that can empirically derive unbiased standard error estimates (Martin & Mullis, 2012). However, multilevel models already account for the non-independence of data explicitly, so if the grouping variables (i.e., random effects) corresponding to the multi-stage sampling design are included, use of replicate weights may be unnecessary (Snijders & Bosker, 2012). It should be noted that when complex sampling plans are used, it is unlikely that researchers will be able to directly model the groups associated with the multi-stage sampling due to the

complexity of the sampling plan. In the present demonstration with TIMSS data, replicate weights are used.

Data and Model

The data used for the present demonstration analysis is from IEA's Trends in International Mathematics and Science Study (TIMSS) 2011 based on the fourth-grade mathematics data. The models estimated in the present study have been examined in previous work related to reporting effect size measures for multilevel models (Lorah, 2018). The reader is referred to Lorah (2018) for more in-depth description of the data and models, but a brief overview is provided here.

The data used for the present analysis includes 46,475 students (level 1) nested within ten randomly selected countries (level 2). The outcome variable was mathematics achievement which is measured with five plausible values. Three predictor variables included Female (binary measure, 0 = boy & 1 = girl), whether the student has internet connection at home (binary measure, 0 = no & 1 = yes), and student confidence with math (continuous).

The country level is modeled as a random effect in the present treatment, but could also be treated as a fixed effect, due particularly to the large sample size within each country. Depending on the goals of the researchers, either modeling choice could be valid; however, in the present model for the purpose of demonstration, and given that the interest is more in the distribution of countries rather than individual countries, a random effect was used.

Similar to Lorah (2018), the empty multilevel model, and a multilevel model with predictors were estimated. These models are:

$$Math_{ij} = \beta_0 + u_{0j} + \varepsilon_{ij} \quad (1)$$

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

where $Math_{ij}$ is the outcome for student i within country j; β_0 is the intercept; u_{0j} is random error at level 2 with estimated variance τ^2; ε_{ij} is random error at level 1 with estimated variance σ^2; all other β are slope coefficients.

Example Demonstration

In order to implement and demonstrate the preceding complexities associated with complex surveys, the BIFIEsurvey package (BIFIE, 2017) was used with R (R Core Team, 2014). The R syntax is provided in the Appendix. The BIFIEsurvey package can automatically handle multiple imputed datasets for each plausible value, incorporate sampling weights, and incorporate replicate weights and it is customized particularly for select datasets, including TIMSS. For the present analysis, both the empty model (1) and full model (2) were estimated adjusting for these complexities. For analyses using plausible values, all 5 plausible values for the math achievement outcome are used and kept in their unstandardized form.

For comparison purposes, two analyses were demonstrated without the use of plausible values, and those used the first math achievement variable only. Sampling weights are incorporated ("TOTWGT") and scaled so that the sum of the weights is equal to the total sample size, in order to produce correct standard error estimates. The BIFIE.data.jack() function is used to first create a dataset with data, sampling weights, plausible values, and replicate weight settings specified and then the BIFIE.twolevelreg() function is used to estimate the multilevel model. By specifying jktype="JK_TIMSS" and keeping the replicate weight variable names initially used by TIMSS, this function automatically uses the correct variables and procedures associated with the replicate weights. Results from this analysis are displayed in Table 1 (last two columns) along with four other incorrect analyses displayed for comparison purposes and sample R syntax is provided in the Appendix.

Analysis 1 in Table 1 does not include weights and only uses one plausible value; analysis 2 additionally incorporates all 5 plausible values; analysis 3 includes sampling weights (but not plausible values); analysis 4 includes both sampling weights and plausible values; and finally, analysis 5 add replication weights in addition to sampling weights and plausible values.

Comparison of analyses 1 and 2 indicates that the addition of plausible values increases the standard error estimates for fixed effects. This is expected as the incorporation of plausible values adds a measure of uncertainty regarding student achievement scores in addition to sampling variability. In addition, in this example, the addition of plausible values doesn't show much impact on the intercept and slope coefficient parameter estimates. This is consistent with the literature, since using just one plausible value has been shown to typically recover population parameters (Rogers & Stoeckel, 2008; Wu, 2005).

Table 1. Comparison of Results from Five Models

	1. Simple analysis		2. Add PV only		3. Add weights only		4. Add PV + weights		5. Add replication method	
	Estimate	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
ICC	0.25		0.25		0.33		0.33		0.33	
Intercept	477.30	0.12	477.30	0.37	465.00	0.13	465.00	0.38	465.00	1.68
Female	0.31	0.11	0.62	0.24	0.08	0.11	0.35	0.26	0.35	1.62
Internet	32.23	0.11	32.28	0.24	26.95	0.10	26.98	0.29	26.98	1.19
Confidence	21.48	0.11	21.37	0.33	21.27	0.11	20.91	0.40	20.91	1.02

Comparison of analyses 1 and 3 shows the impact of including sampling weights. Although the sampling weights are expected to slightly increase standard error estimates (Meinck & Vanderplas, 2012), in this particular example some standard error estimates are larger and some smaller than for the model excluding sampling weights, although all estimates are fairly similar. A comparison of the point estimates for fixed effects shows that inclusion of sampling weights results in changes to these point estimates. The intercept estimate decreases, possibly indicating that higher-achieving students were oversampled; analogously, the coefficient for Internet decreases, possibly indicating that students showing a stronger relationship between internet access and achievement were oversampled. The coefficient for Confidence remains about the same, and the coefficient for Female remains non-significant.

Analysis 4 incorporates both plausible values and sampling weights and, as expected, produces larger standard error estimates than analysis 2 (just plausible values) or analysis 3 (just sampling weights). Finally, analysis 5 incorporates replication weights, in addition to the already incorporated plausible values and sampling weights. Replication weights may be unnecessary if the level two variable corresponds exactly to the nesting structure in the data (Snijders & Bosker, 2012). However, in the present analysis, the sampling structure for the data includes, for example, school membership as a nesting factor, which is not explicitly included in the multilevel model and therefore replication weights are incorporated. Examination of analysis 5 indicates that incorporation of replication weights does not affect the parameter estimates but do increase the standard errors. Since ignoring the nesting of students within schools represents a violation of the assumption of independence, it is logical that ignoring this aspect of the data would result in downwardly biased standard errors. The replication weights should correct for this and should more accurately represent the precision of these estimates.

Conclusion

A description and explanation are provided of plausible values, sampling weights, and replicate weights as they apply to analysis of multilevel models. A demonstration using TIMSS data was provided. Although TIMSS data was used in the present analysis, the process of estimating a multilevel model with complex survey data would be analogous for other complex survey data, such as the National Assessment of Educational Progress (NAEP) or the Program for International Student Assessment (PISA) and could be applied similarly.

Acknowledgements

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

References

BIFIE. (2017). BIFIEsurvey: Tools for survey statistics in educational assessment [R package, version 1.13-24]. Retrieved from https://cran.r-project.org/package=BIFIEsurvey

Gardiner, J. C., Luo, Z., & Roman. L. E. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine, 28*(2), 221-239. doi: 10.1002/sim.3478

Graubard, B. I., & Korn, E. L. (1994). Regression analysis with clustered data. *Statistics in Medicine, 13*(5-7), 509-522. doi: 10.1002/sim.4780130514

Kalton, G. (1983). *Introduction to survey sampling*. Newbury Park, CA: Sage.

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics – Theory and Methods, 46*(22), 11341-11357. doi: 10.1080/03610926.2016.1267764

Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics – Theory and Methods, 47*(20), 4991-5012. doi: 10.1080/03610926.2017.1383429

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. Newbury Park, CA: Sage.

Lorah, J. A. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6*, 8. doi: 10.1186/s40536-018-0061-2

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, NJ: Wiley.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/methods/

McNeish, D. (2019). Effect partitioning in cross-sectionally clustered data without multilevel models. *Multivariate Behavioral Research, 54*(6), 906-925. doi: 10.1080/00273171.2019.1602504

Meinck, S., & Vandenplas, C. (2012). *Sample size requirements in HLM: An empirical study*. Princeton, NJ: IERInstitute. Retrieved from https://www.ierinstitute.org/dissemination-area.html#accordion-55

R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.r-project.org/

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805-827. doi: 10.1111/j.1467-985X.2006.00426.x

Rogers, A. M., & Stoeckel, J. J. (2008). *NAEP 2008 arts: Music and visual arts restricted-use data files data companion* (NCES no. 2011470). Washington, DC: National Center for Education Statistics.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151. doi: 10.3102/0013189X10363170

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. New York: CRC Press.

Skinner, C., & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science, 32*(2), 165-175. doi: 10.1214/17-STS614

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

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128. doi: 10.1016/j.stueduc.2005.05.005

Appendix A: R Syntax

```
library(BIFIEsurvey)
#level 2 weights all equal; sum of weights equals level 2 sample size
mydata2$L2WGT<-rep(.1,nrow(mydata2))
#Scale weight variable to sum up to total sample size
mydata2$TOTWGTscale<-mydata2$TOTWGT/(sum(mydata2$TOTWGT)/nrow(mydata2))

#create BIFIE.data objects for each scenario within Table 1
bdat1 <- BIFIE.data(data=mydata2) #scenario 1
bdat2<-BIFIE.data(data=mydata2,pv_vars=c("math")) #scenario 2
bdat3 <-BIFIE.data(data=mydata2,wgt="TOTWGTscale") #scenario 3
bdat4 <-BIFIE.data(data=mydata2,wgt="TOTWGTscale",pv_vars=c("math"))
   #scenario 4
bdat5 <-BIFIE.data.jack(data=mydata2,wgt="TOTWGTscale",
   pv_vars=c("math"),jktype="JK_TIMSS") #scenario 5

#Empty model to compute ICC, scenario 1-5
M1a<-BIFIE.twolevelreg(BIFIEobj=bdat1,dep="math1",formula.fixed=~1,
   formula.random=~1,idcluster="IDCNTRY",wgtlevel2="L2WGT",se=FALSE)
M2a<-BIFIE.twolevelreg(BIFIEobj=bdat2,dep="math",formula.fixed=~1,
   formula.random=~1,idcluster="IDCNTRY",wgtlevel2="L2WGT",se=FALSE)
M3a<-BIFIE.twolevelreg(BIFIEobj=bdat3,dep="math1",formula.fixed=~1,
   formula.random=~1,idcluster="IDCNTRY",wgtlevel2="L2WGT",se=FALSE)
M4a<-BIFIE.twolevelreg(BIFIEobj=bdat4,dep="math",formula.fixed=~1,
   formula.random=~1,idcluster="IDCNTRY",wgtlevel2="L2WGT",se=FALSE)
M5a<-BIFIE.twolevelreg(BIFIEobj=bdat5,dep="math",formula.fixed=~1,
   formula.random=~1,idcluster="IDCNTRY",wgtlevel2="L2WGT")

#Full model, scenario 1-5
M1b<-
   BIFIE.twolevelreg(BIFIEobj=bdat1,dep="math1",formula.fixed=~1+Female
   Scale +InternetScale+ ConfScale, formula.random=~1,
   idcluster="IDCNTRY", wgtlevel2="L2WGT", se=FALSE)
M2b<-BIFIE.twolevelreg(BIFIEobj=bdat2,dep="math",formula.fixed=~1+
   FemaleScale +
```

```
InternetScale+ConfScale, formula.random=~1, idcluster="IDCNTRY",
   wgtlevel2="L2WGT", se=FALSE)
M3b<-
   BIFIE.twolevelreg(BIFIEobj=bdat3,dep="math1",formula.fixed=~1+Female
   Scale+
InternetScale+ConfScale, formula.random=~1,
   idcluster="IDCNTRY",wgtlevel2="L2WGT", se=FALSE)
M4b<-
   BIFIE.twolevelreg(BIFIEobj=bdat4,dep="math",formula.fixed=~1+FemaleS
   cale +
InternetScale+ConfScale, formula.random=~1, idcluster="IDCNTRY",
   wgtlevel2="L2WGT", se=FALSE)
M5b<-
   BIFIE.twolevelreg(BIFIEobj=bdat5,dep="math",formula.fixed=~1+FemaleS
   cale +InternetScale+ConfScale, formula.random=~1,
   idcluster="IDCNTRY", wgtlevel2="L2WGT")
```

Appendix B: First Six Rows of Dataset

```
> head(mydata2)
  X IDCNTRY FullSchID FullStuID    math1    math2    math3    math4
1 1      48      bhr1  bhr10301 386.7395 433.1434 473.8001 422.6886
2 2      48      bhr1  bhr10302 555.8601 542.7640 505.6931 520.5698
3 3      48      bhr1  bhr10303 416.9616 476.2041 412.7749 432.9446
4 4      48      bhr1  bhr10304 362.2361 308.9201 323.0889 338.3065
5 5      48      bhr1  bhr10305 370.1522 421.9536 430.0027 394.4468
6 6      48      bhr1  bhr10306 523.9933 396.5503 423.8978 390.1867
     math5 female internet ConfCont   TOTWGT JKZONE JKREP WGTFAC1
1 450.3125      0        1  7.05574 5.747126      1     1       1
2 507.1065      0        1 10.26444 5.747126      1     1       1
3 451.6919      0        1  3.34399 5.747126      1     1       1
4 362.2385      0        1  8.64408 5.747126      1     1       1
5 350.2854      0        1  8.41228 5.747126      1     1       1
6 391.1212      0        1 10.26444 5.747126      1     1       1
  WGTFAC2 WGTFAC3  WGTADJ1 WGTADJ2  WGTADJ3   HOUWGT FemaleScale
1       5       1 1.111111       1 1.034483 1.621471  -0.9964237
2       5       1 1.111111       1 1.034483 1.621471  -0.9964237
3       5       1 1.111111       1 1.034483 1.621471  -0.9964237
4       5       1 1.111111       1 1.034483 1.621471  -0.9964237
5       5       1 1.111111       1 1.034483 1.621471  -0.9964237
6       5       1 1.111111       1 1.034483 1.621471  -0.9964237
  InternetScale   ConfScale L2WGT TOTWGTscale
1     0.5335917 -1.58162556   0.1   0.1306992
2     0.5335917  0.08803909   0.1   0.1306992
3     0.5335917 -3.51305503   0.1   0.1306992
4     0.5335917 -0.75512412   0.1   0.1306992
5     0.5335917 -0.87574253   0.1   0.1306992
6     0.5335917  0.08803909   0.1   0.1306992
```