Large-scale Assessments in Education

SOFTWARE ARTICLE

Open Access

Using plausible values when fitting multilevel models with large-scale assessment data using R

Francis L. Huang[1,2*]

*Correspondence:
huangf@missouri.edu

[1] University of Missouri, 16 Hill Hall, Columbia, MO 65211, USA
[2] Missouri Prevention Science Institute, University of Missouri, 16 Hill Hall, Columbia 65211, USA

Abstract

The use of large-scale assessments (LSAs) in education has grown in the past decade though analysis of LSAs using multilevel models (MLMs) using R has been limited. A reason for its limited use may be due to the complexity of incorporating both plausible values and weighted analyses in the multilevel analyses of LSA data. We provide additional functions in R that extend the functionality of the WeMix (Bailey et al., 2023) package to allow for the automatic pooling of plausible values. In addition, functions for model comparisons using plausible values and the ability to export output to different formats (e.g., Word, html) are also provided.

Keywords: R, Multilevel models, Plausible values, Weights

The use of modern large-scale assessments (LSAs) in education has grown dramatically over the years. Based on metadata from the Web of Science database, the annual number of articles published using international LSAs in education has grown from fewer than 10 in 1997 to over 300 articles per year in 2020 (Hernández-Torrano & Courtney, 2021). Commonly-used datasets include PISA (the Programme for International Student Assessment organized by the Organization for Economic Co-operation and Development [OECD]) and TIMSS (the Trends in International Mathematics and Science Study coordinated by the International Association for the Evaluation of Educational Achievement [IEA]) (Hernández-Torrano & Courtney, 2021; Laukaityte & Wiberg, 2018).

Although the LSA public-use datasets are freely and readily available for download from the respective agency websites,[1] many of the tutorials for the analyses of such data have been limited to the use of commercial software such as Mplus (Yamashita et al., 2021) or SAS (Rutkowski et al., 2010). Recent articles (e.g., Caro & Biecek, 2017; Mirazchiyski, 2021) have focused on the use of the open-source R (R Core Team, 2022) statistical software. However, most of the articles do not focus on how to use multilevel models (Raudenbush & Bryk, 2002) for the analysis of LSAs using R.

Multilevel modeling (MLM or hierarchical linear modeling or mixed effects modeling) is a well-known and highly flexible regression-based approach used for the analysis of

[1] For example: https://www.oecd.org/pisa/data/ or https://timss2019.org/international-database/.

Springer Open

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

clustered or nested data. MLM allows the variance in the outcome variable to be appropriately partitioned within and between clusters—which in itself may be a research focus of interest (e.g., how much variability in the outcome is due to the school or the student?). Research questions focusing on the variance partitioning have had a long history in educational policy research such as those found in the Coleman report (1966) which looked at the unique contributions associated with student- and school-level factors related to academic achievement. MLM can also be used for data with more than two levels, commonly used in cross-national studies, allowing researchers to look at the student, school, and country effects all in one model (e.g., Baysu et al., 2023).

Although MLM has grown in popularity over the years (Huang, 2018), using R specifically for the multilevel analyses of LSAs has been limited. This is likely due to several of the following reasons which are specific to the analyses of LSA data. First, the two commonly-used R MLM packages of lme4 (Bates, 2010) and nlme (Pinheiro et al., 2022) do not allow for the use of sampling weights at the different levels of the model. Second, if the plausible values (PVs) of the outcome measures are to be used when analyzed using either lme4 or nlme, there has been no straightforward way (i.e., simply using a function) to properly pool results, aside from manually doing this through syntax (e.g., Lorah, 2022). Third, although packages such as EdSurvey (Bailey et al., 2020) and BIFIEsurvey (Robitzsch & Oberwimmer, 2022) provide features for R users to conduct multilevel analysis using PVs and weighted data analyses, users may have an additional challenge of learning a new package which requires figuring out how to download ILSA data (which is done using the package) and then filter, select, and recode the data specifically using custom-built package functions.[2] In a detailed comparison of five R packages for the analysis of LSA data, Ringiene et al. (2022) indicated that proper data preparation using R functions specific to certain packages can be complex and may lead applied researchers to not use R. Software such as Mplus (Muthén & Muthén, 1998–2017) and HLM (Raudenbush & Congdon, 2021) are popular among LSA researchers using MLM due to their ability to accommodate both the use of weights and plausible values (Karakolidis et al., 2022). Note however that software such as Mplus and HLM still require researchers to perform all of the data management necessary to analyze the data using other software.

As R has evolved over the years, so have its data management capabilities using packages such as dplyr[3] (Wickham et al., 2020) and tidyr (Wickham, 2021), making R much more accessible to applied researchers. Researchers who are already familiar with R, are used to managing their own data, and already know how to fit multilevel models may want to simply fit the models of interest with minimal coding or without having to learn how to use a different package. The R package WeMix (Weighted mixed effect models; Bailey et al., 2023) was specifically designed to allow users to fit multilevel models (both linear and logistic regression models) with weights at different levels (such as those commonly found in LSAs) and uses standard formula notation commonly used

[2] Obtaining LSA data may be challenging for some and may require having access to either SAS or SPSS. The use of the R packages for obtaining the data helps reduce the burden on the users who may not have access to the necessary commercial software.

[3] As a sign of its popularity, based on results from the packageRank package, as of 2023.11.27, dplyr was the 10th most downloaded package out of 19,625 R packages on CRAN.

in other R functions.[4] However, the task of fitting multiple models and pooling the output—which is a due to the use of plausible values (Mislevy et al., 1992)—is still left to the users. To address this, we provide some R functions in a form of an R wrapper (which is a function that wraps around another function in R), that extends the functionality of WeMix to allow for the analysis of multiple datasets, pooling of results, the ability to conduct nested model comparisons, and easily output regression tables in a customizable and exportable format. The functions provided are specifically designed for users who already know how to obtain LSA data (i.e., download the data from the appropriate websites), are familiar with managing their data using R, already have a background on multilevel modeling (there are several primers on the topic) but want to simply analyze their data properly using multilevel modeling using R (i.e., WeMix follows the conventional mixed effects notation already used in lme4) without having to program how to pool results. We compare results as well to output produced using SAS and the EdSurvey package (see Appendices).

The challenge of analyzing LSA data

Two defining characteristics of LSAs involve the use of sampling weights and plausible values. For practical and statistical purposes, the samples used for LSAs are not simple random samples and are drawn for the purpose of making inferences about the population (i.e., population estimates) using multistage sampling. In addition, when students are assessed in a particular subject area (e.g., math, reading, science), students are only assigned portions of the assessment (i.e., certain blocks or booklets) and not the assessment in its entirety. With plausible value (Mislevy et al., 1992) methodology, as students do not complete the entire assessment, student achievement is treated as missing data which needs to be properly accounted for. Statistical analyses must account for these two design characteristics of weights and plausible values to avoid biasing both the point estimates and standard errors (Laukaityte & Wiberg, 2017; Rutkowski et al., 2010). The use of weights and plausible values are briefly described.

Using weights

The specific details of the weighting procedures are explained in the user manuals of the particular LSAs and have been discussed in much detail in several articles (Kim et al., 2013; Meinck, 2015; Rutkowski et al., 2010). The use of sampling weights with survey data though has been “a subject of controversy among theorists” (Pfeffermann, 1993, p. 317) and findings from Monte Carlo simulations (where the true population value is known) have shown different results where the use of weights may (Mang et al., 2021) or may not matter (Laukaityte & Wiberg, 2018). However, the general recommendation in the LSA manuals is to use the weights as the objective of the analyses is to make generalizations to the population and not the sample itself (Fishbein et al., 2021; Herget et al., 2019).

[4] Note that the BIFIEsurvey package (Robitzsch & Oberwimmer, 2022) can fit multilevel linear models but only allows for two-level models.

With multilevel models, weights can be formed at different levels. This corresponds to the sampling design where in some assessments, within a country (or locale), schools[5] (level 2) are first selected (with a probability proportional to size) and then students or teachers (level 1) within schools are sampled. Not all multilevel models may require the use of weights but when working with LSAs that have complex sampling designs and inferential statistics are of interest, weights can be used (Sterba, 2009).[6] To account for the sampling design, in a multilevel framework, using weights at the different levels has been suggested (Rathbun et al., 2021). Another approach would be to use the total student weight (which is a product generally of the school and student weights which also includes some other adjustments) on its own (Zhang et al., 2020). Yet, another alternative and simpler approach when running MLMs is to only use the school-level weight at the second level without the need to specify the level-1 weight (Mang et al., 2021). As the sampling weights account for the sampling design (e.g., any stratification or oversampling) as well as adjustments for nonresponse, the use of weights is recommended (Joncas, 2007). As indicated by Snijders and Bosker, "the reason for using sampling weights is to avoid bias" (2011, p. 221).

When using weights with multilevel models, careful attention must be paid as to what type of weights are being used and what the software is actually doing. Most LSAs may provide an unconditional student weight for use at level 1, however some software (e.g., SAS) may require the unconditional weight to be rescaled by dividing the level-1 weight by the school weight. For a discussion on and examples of how the different weights are computed, see Rutkowski et al. (2010) and as indicated, researchers "should consult their software documentation for the appropriate application of weights at multiple levels" (p. 144).

Using plausible values

To reduce the test burden on the respondents, students participating in LSAs do not complete the entire battery of assessments. For example, with TIMSS, if a student were to take the entire assessment, this would represent more than 10 h of testing time (Rutkowski et al., 2010). Instead, students are assigned certain test booklets to complete and, because of the administration method, individual testing time is reduced to 90 min.

However, as the students do not complete the entire assessment, this can be treated as a missing data problem where missing values can be imputed (Mislevy et al., 1992). Random draws (five or ten depending on the LSA) from an estimated ability distribution are repeatedly taken for every student which are referred to as plausible values (Rutkowski et al., 2010). Different LSAs may use a different number (*m* number) of plausible values and are appropriate for making population- or subpopulation-level estimates and they are not individual scores. These values represent an ability range for each student. As a result, additional measurement error is introduced into the outcome due to the use of multiple plausible values. Thinking of the plausible values as imputed values—as used in multiple imputation to account for missing data—may be helpful.

[5] This also depends on the country. For example, in a small country such as Singapore, all schools are selected so the corresponding school weight is 1.0.

[6] For a discussion on model-, design-, and hybrid-based approaches to analysis, readers can consult Sterba (2009).

Even though each student has m plausible values representing some latent (i.e., unobserved) ability measure, an incorrect way of analyzing the data would be to take the average of all the plausible values or even just taking one of the values and then fitting a model (Aparicio et al., 2021). Doing so will result in generally underestimated standard errors which do not account for the variability resulting from the slightly different results for each m analyses. Instead, models should be fit m number of times, each with one of the plausible values as the outcome. As a result of the differing values, regression coefficients and standard errors will fluctuate slightly from model to model. The results of the m analyses should then be pooled using Rubin's (2004) rules so that in the end, only one set of results are reported.

Rutkowski et al. (2010) provide an example showing how results can differ using only one value, an average set of values, and a properly pooled set of results. It is likely in this stage of the analysis where applied researchers may have some difficulty as even with software such as SAS, data have to be converted to a long format, analyzed multiple times, and then pooled appropriately. Note that the handling of plausible values is only of importance if the assessment measure is used as some analyses may not focus on the ability measures (e.g., focus is on bullying; Smith & López-Castro, 2017).

Pooling results: estimates and standard errors

Rubin's (2004) method for pooling results has long been used with multiply-imputed data to account for imputation variability. For a regression model, pooling the regression coefficient b is straightforward and merely the average of the coefficients ($\overline{b}$) from the m analyses. The standard errors—which captures the uncertainty of the estimate, takes slightly more work to compute and is not the simple average of the standard errors.

Using formulas adapted from Schafer and Olsen (1998, p. 557), the pooled standard errors are made up of the within ($\overline{U_b}$) and between imputation (B_b) variance for each b coefficient. The within imputation variance of a regression coefficient is $\overline{U}_b = \frac{\Sigma SE_b^2}{m}$ which is the average of the squared standard errors over the m sets of analyses. The between imputation variance is $B_b = \frac{\left(b-\overline{b}\right)^2}{m-1}$ which is the variance of the regression coefficients over the m sets of analyses. Combining the two sources of variance results in $T_B = \overline{U}_b + \left(1 + \frac{1}{m}\right)B_b$ and the pooled standard error is $SE_b = \sqrt{T_b}$.

The estimate (b) is then divided by its standard error to obtain the corresponding t-statistic. The corresponding degrees of freedom (*df*) for the b coefficient is computed as: $df = (m-1)\left(1 + \frac{m\overline{U}_b}{(m+1)B_b}\right)^2$. The p-values can then be evaluated using the t-statistic with the corresponding *df*. The subscript b indicates that this is computed for each b coefficient.

Pooling results: likelihood ratio tests

A common method for evaluating improvements in multilevel model fit uses a likelihood ratio test (LRT) that compares two nested models (i.e., a full and a restricted model) with each other. A restricted model is nested within a full model if the restricted model can be obtained by excluding parameters to be estimated from the full model. The difference in deviance statistics (deviance = − 2 × log-likelihood or − 2LL) between the two models (i.e., $\Delta d = -\ 2\text{LL}_{\text{FULL}} - -\ 2\text{LL}_{\text{REDUCED}}$) is evaluated using a χ^2 statistic with k degrees of freedom where k represents the difference in the number of parameters estimated in the full and the reduced models. A statistically significant result would indicate a better fit of the full model and a nonstatistically significant result would suggest that the simpler, more parsimonious model would suffice. However, there are several pooling approaches for LRTs to choose from and the computation is not as straightforward as the approach to pooling the estimates and standard errors (see Grund et al., 2023 for a comparison of three approaches).

We show the computation for the pooled statistic as proposed by Li et al. (1991) referred to as the D_2 statistic by Schafer (1997, Eq. 4.40) which involves pooling the χ^2 statistic from each m model analyzed using different plausible values. The D_2 statistic is calculated using $D_2 = \frac{\frac{\overline{d}}{k} - \frac{m+1}{m-1} r_m}{1+r_m}$ where $\overline{d}$ is the average Δd statistic from the m models and r_m is the an estimate of the average relative increase in variance as a result of missing (i.e., imputed) values. The formula for $r_m = \left(1 + \frac{1}{m}\right)\left(\frac{\sum_{i=1}^{m} (\sqrt{d_m} - \overline{\sqrt{d}})^2}{m-1}\right)$ where $\overline{\sqrt{d}}$ is the average value of the square root of Δd for each model, $\frac{\sum_{i=1}^{m} \sqrt{d_m}}{m}$. Although the second part of the r_m equation may look complicated, this is merely the variance of the square root of the Δd statistic for each m model.

The D_2 statistic is evaluated using an F distribution with k df for the numerator and v_2 df for the denominator where $\mathrm{v}_2 = k^{-3/m}(m-1)\left(\frac{1}{r_2}\right)^2$ (see Schafer, 1997, Eq. 4.41). The F distribution used for the D_2 corresponds to the χ^2 distribution for LRTs using complete data but accounts for the number of imputations (m plausible values) used (Grund et al., 2023). Although the D_2 statistic has been found to result in somewhat higher levels of Type I errors, this is an issue with smaller sample sizes (e.g., $n = 100$) (Grund et al., 2023) which is not the case when analyzing LSAs which typically have thousands of observations and over a hundred clusters.

Some researchers though may want to use information criterion measures, such as the Akaike information criterion (AIC), to assess the quality of competing statistical models, with lower values indicating better model fit. Combining these measures as a result of multiple models, which are predicated on using the same dataset, is not clear (Grund et al., 2016). Some though have suggested using the average of the AIC measures resulting from m datasets or creating and analyzing an averaged dataset using the m complete datasets (Schomaker et al., 2010 as cited in Consentino & Claeskens, 2010, p. 2294). Using a simulation, Consentino and Claeskens showed that different pooling approaches for the AIC performed similarly. Though commonly done, we caution against the use of information criterion measures which may actually not perform well when selecting the best fitting models (e.g., Ferron et al., 2002; Gelman & Rubin, 1994; Vallejo et al., 2008).

The current study

To reduce the complexity in the analysis of LSA data using multilevel models, we provide several freely downloadable functions (available at https://github.com/flh3/pubdata/tree/main/mixPV), to be used together with the WeMix package (Bailey et al., 2023) for R. The following functions are provided:

- `mixPV`: for the analysis using plausible values using the `mix` function in WeMix.

- `summary`: for generating the pooled output resulting from `mixPV`.
- `summary_all`: for viewing the MLM output for each plausible value.
- `lrtPV`: for conducting model comparisons using models fit using plausible values.
- `glance`: for viewing summary statistics.

In addition, "helper" functions[7] are provided that make the output readily exportable—in formats such as Word or html, using the modelsummary (Arel-Bundock et al., 2022) package.

Data analysis

We extend the example in the WeMix vignette[8] that used PISA 2012 data from the United States (USA) but only used the first plausible value for math (`pv1math`). For the current analyses, five plausible values (i.e., `pv1math`, `pv2math`, `pv3math`, `pv4math`, `pv5math`) will be used. Data are available at https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm and requires researchers to select only data from the USA and merge both the student and school data files using the `schoolid` variable.[9] The files are merged as required by standard multilevel modeling software and we do so in this example in order to use both student- and school-level predictors. The predictors used for the current example are shown in Table 1.

The student- and school-level weights at level one and level two are `w_fstuwt` and `w_fschwt`, respectively. These weights are provided in the PISA dataset and are referred to as unconditional weights which can be used directly (i.e., without alteration) with the `mix` function when specifying the weights at two levels. To compute conditional student weights (which are used by some software), the total student weight (`w_fstuwt`) can be divided by the school weight (`w_fschwt`).[10] Of the observations without missing data, there were 3,136 students nested in 157 schools.

Using composite notation, the random intercepts model can be expressed as: $Y_{ij} = \gamma_{00} + \gamma_{01}LMT_j^{VL} + \gamma_{02}LMT_j^{SE} + \gamma_{03}LMT_j^{AL} + \gamma_{10}LFM_{ij}^{A} + \gamma_{20}LFM_{ij}^{D} + \gamma_{30}LFM_{ij}^{SD} + \gamma_{40}MALE_{ij}$ $+\gamma_{50}ESCS_{ij} + u_{0j} + r_{ij}$, where Y_{ij} is the outcome for student i in school j; LMT_j^{VL-AL} represent three dummy codes for the school-level variable for the lack of qualified math teachers; LFM_{ij}^{A-SD} represent three dummy codes for the student-level variable for looking forward to math lessons; $MALE_{ij}$ is a dummy code for student gender; and $ESCS_{ij}$ is the continuous student-measure of socioeconomic status. The error term u_{0j} captures the variability of the outcome between schools and r_{ij} is the student-level error term.

[7] A function that allows one function to work with other functions.

[8] https://cran.r-project.org/web/packages/WeMix/vignettes/Introduction_to_Mixed_Effects_Models_With_WeMix.pdf

[9] When importing SPSS files into R, users can use the rio::import() function. Although the haven::read_sav() function may work, WeMix may have issues with the labels used in.
haven. The variable labels may be removed using the haven::zap_labels() function.
The combined R data file can also be accessed using.
> data(pisa2012, package = 'MLMusingR') # from package version 0.3.2 or.
> pisa2012 <—rio::import("https://github.com/flh3/pubdata/raw/main/mixPV/pisa2012.rds").

[10] By default, this does not have to be done when using the mix function. However, if conditional weights are used, this option can be set by using the mix function and including the option cWeights = TRUE. The conditional weight in the dataset is variable pwt1.

Table 1 Descriptive statistics

Variable	Description	
Level 1 (student level; *n* = 3,316)		
escs	A continuous socioeconomic status variable (an index of economic, social and cultural status)	*M* = 0.20 *SD* = 0.98

st04q01	Student gender (male or female) with female as the reference group	Female = 1570 (50%)
		Male = 1566 (50%)

st29q03	"I look forward to mathematics lessons" with response options strongly agree (the reference group), agree, disagree, and strongly disagree	Strongly agree = 387 (12%)
		Agree = 1037 (33%)
		Disagree = 1231 (39%)
		Strongly disagree = 481 (15%)

Level 2 (school level; *n* = 157)		
sc14q02	"Is your school's capacity to provide instruction hindered by any of the following... A lack of qualified mathematics teachers" with response options a lot, to some extent, very little, and not at all (the reference group)	Not at all = 116 (74%) Very little = 25 (16%) To some extent = 13 (8%) A lot = 3 (2%)

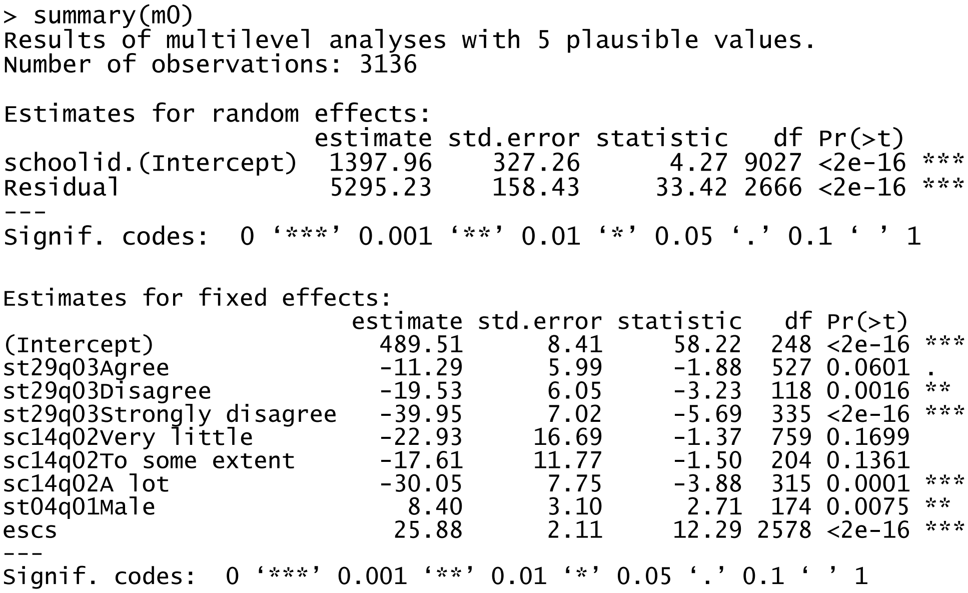
Unweighted statistics shown.

The standard method of fitting a random intercepts model using the `mix` function (without plausible values) in WeMix can be done with the following specification:

```
> nopv <- mix(pv1math ~ st29q03 + sc14q02 + st04q01 + escs +
(1|schoolid), weights = c('w_fstuwt', 'w_fschwt'), data = pisa2012)
> summary(nopv)
```

However, to use the five plausible values in the analysis, we specify each of the plausible values as dependent variables (on the left-hand side of the equation) in one model using the new `mixPV` function. There is no need to reshape the data from a wide to long format as may be required by other software (e.g., SAS). To generate results and to have the output properly formatted, the broom package (Robinson et al., 2022) needs to be installed as well using `install.packages(’broom’)`. The newly introduced functions can be loaded using the `source` function:

```
> source("https://raw.githubusercontent.com/flh3/pubdata/main/mixPV/
mixPV.R")
> m0 <- mixPV(pv1math + pv2math + pv3math + pv4math + pv5math ~
      st29q03 + sc14q02 + st04q01 + escs + (1|schoolid), weights =
      c('w_fstuwt', 'w_fschwt'), data = pisa2012)
Analyzing plausible value: pv1math
Analyzing plausible value: pv2math
Analyzing plausible value: pv3math
Analyzing plausible value: pv4math
Analyzing plausible value: pv5math
```



```
> summary(m0)
Results of multilevel analyses with 5 plausible values.
Number of observations: 3136

Estimates for random effects:
                       estimate std.error statistic   df Pr(>t)
schoolid.(Intercept)   1397.96    327.26      4.27 9027 <2e-16 ***
Residual               5295.23    158.43     33.42 2666 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimates for fixed effects:
                           estimate std.error statistic   df Pr(>t)
(Intercept)                  489.51      8.41     58.22  248 <2e-16 ***
st29q03Agree                 -11.29      5.99     -1.88  527 0.0601 .
st29q03Disagree              -19.53      6.05     -3.23  118 0.0016 **
st29q03Strongly disagree     -39.95      7.02     -5.69  335 <2e-16 ***
sc14q02Very little           -22.93     16.69     -1.37  759 0.1699
sc14q02To some extent        -17.61     11.77     -1.50  204 0.1361
sc14q02A lot                 -30.05      7.75     -3.88  315 0.0001 ***
st04q01Male                    8.40      3.10      2.71  174 0.0075 **
escs                          25.88      2.11     12.29 2578 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the only difference in the specification is the use of multiple values for the dependent variable (i.e., `pv1math + pv2math + pv3math + pv4math + pv5math`) together with the `mixPV` function. The output shows both the combined random and fixed effects using the point estimates, standard errors, the *t*-statistics, degrees of freedom, and the *p*-values computed and combined using Rubin's (2004) rules. By default, the standard errors reported are also the robust standard errors (Liang & Zeger, 1986) which account for heterogeneity of variance violations (Huang et al., 2022). For a more detailed discussion on robust standard errors and their computation in the context of mixed models, readers can consult Huang et al. (2022).

Although the `mixPV` function is run once, the model is fit five times using `mix`, once for each plausible value specified. If the user wants to see the result of each analysis separately, `summary_all(m0)` can be used. The `glance(m0)` function can also be used to view the number of observations, the number of plausible values used, and the average AIC and BIC statistics:

```
> glance(m0)
              Nobs N.pv   AICbar   BICbar
Number of obs 3136    5 25651569 25651599
```

Following the original WeMix vignette, a random slope for `escs` can also be specified in the standard manner (as done in `lmer` and `lme`) and in this case the variable `escs` is allowed to randomly vary by school by including `(escs|schoolid)`.

```
> m1 <- mixPV(pv1math + pv2math + pv3math + pv4math + pv5math ~
      st29q03 + sc14q02 + st04q01 + escs + (escs|schoolid), weights =
      c('w_fstuwt', 'w_fschwt'), data = pisa2012)
(...output omitted...)

> summary(m1)
Results of multilevel analyses with 5 plausible values.
Number of observations: 3136

Estimates for random effects:
                     estimate std.error statistic     df Pr(>t)
schoolid.(Intercept)  1380.68    325.42      4.24 2151.8 <2e-16 ***
schoolid.escs          324.39     63.89      5.08  350.5 <2e-16 ***
Residual              5021.04    106.04     47.35   81.6 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimates for fixed effects:
                           estimate std.error statistic   df Pr(>t)
(Intercept)                  486.44      8.27     58.84  245 <2e-16 ***
st29q03Agree                 -10.54      5.89     -1.79  370 0.0746 .
st29q03Disagree              -17.40      6.08     -2.86  108 0.0050 **
st29q03Strongly disagree     -36.88      7.05     -5.23  250 <2e-16 ***
sc14q02Very little           -23.18     15.78     -1.47  706 0.1422
sc14q02To some extent        -13.75     11.88     -1.16  480 0.2478
sc14q02A lot                 -35.74      8.59     -4.16  258 <2e-16 ***
st04q01Male                    8.88      3.09      2.88  192 0.0045 **
escs                          27.12      2.44     11.11 1412 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Although the random slope (`schoolid.escs`) shows the p-value of the associated Wald test (i.e., $p < .001$), the use of a likelihood ratio test (LRT) is often recommended when testing variance components (Berkhof & Snijders, 2001). To conduct a model comparison between a model using a likelihood ratio test with and without a random slope, the `lrtPV` function can be used by specifying the fitted full and the reduced model (note that the order has be the full model first and the reduced model second):

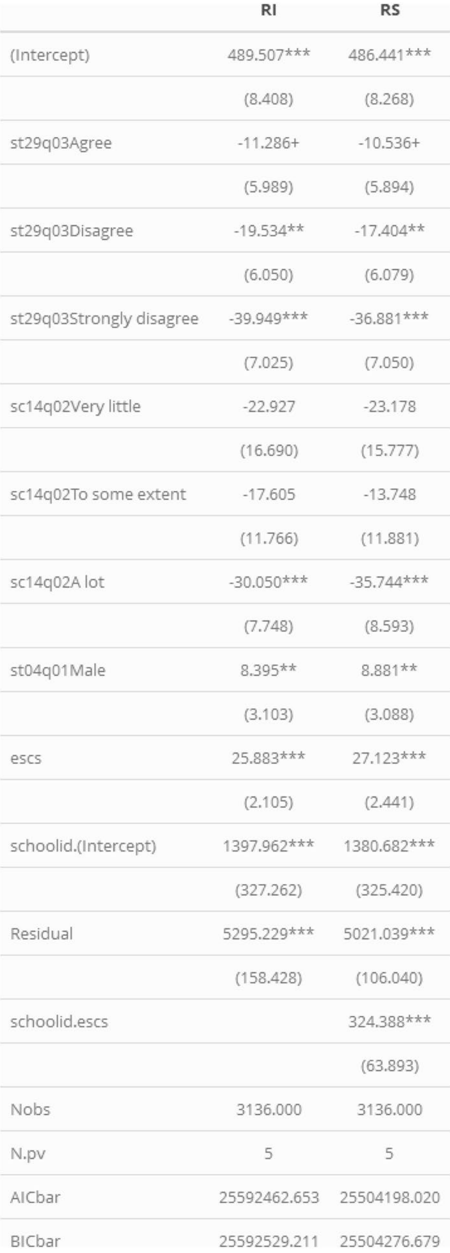
```
> lrtPV(m1, m0)
         F df1      df2        r     pv
1 98.27685   2 2.650988 441.3453 0.0033
```

The likelihood ratio test, based on the D_2 statistic (Li et al., 1991), indicates that the model fits better with the random slope, $F(2, 2.65) = 98.3, p < .01$.

Finally, several model results can be shown side-by-side using the `modelsummary` function (from the package of the same name). The results can be shown using:

```
> library(modelsummary)
> modelsummary(list("RI" = m0, "RS" = m1), stars = TRUE)
```

Note that the output table in Fig. 1 is shown 'as-is' and needs some editing to get it ready for publication (e.g., indicating that these are robust standard errors in parenthesis, indicating the reference group for the categorical variables, separating the random effects, adding other notes). The `modelsummary` function has many useful options for formatting the output (e.g., showing confidence intervals, having estimates and standard errors beside other, controlling the number of digits to show). There is extensive documentation for the use of the `modelsummary` function available at: https://vincentarelbundock.github.io/modelsummary/articles/modelsummary.html. If the random effects are to be hidden (so that only fixed effects are shown),



	RI	RS
(Intercept)	489.507***	486.441***
	(8.408)	(8.268)
st29q03Agree	-11.286+	-10.536+
	(5.989)	(5.894)
st29q03Disagree	-19.534**	-17.404**
	(6.050)	(6.079)
st29q03Strongly disagree	-39.949***	-36.881***
	(7.025)	(7.050)
sc14q02Very little	-22.927	-23.178
	(16.690)	(15.777)
sc14q02To some extent	-17.605	-13.748
	(11.766)	(11.881)
sc14q02A lot	-30.050***	-35.744***
	(7.748)	(8.593)
st04q01Male	8.395**	8.881**
	(3.103)	(3.088)
escs	25.883***	27.123***
	(2.105)	(2.441)
schoolid.(Intercept)	1397.962***	1380.682***
	(327.262)	(325.420)
Residual	5295.229***	5021.039***
	(158.428)	(106.040)
schoolid.escs		324.388***
		(63.893)
Nobs	3136.000	3136.000
N.pv	5	5
AICbar	25592462.653	25504198.020
BICbar	25592529.211	25504276.679

Fig. 1 Output using the `modelsummary` function

the `coef_omit='schoolid|Residual'` option can be added (the characters within the quotations and separated by the pipe operator [|] are matched and hidden). By default, the output is displayed onscreen but if instead the user wants to output the file to a Word file, the option `out='results.docx'` can be specified (other options include jpg, html, tex). As a basis for comparison, model results using plausible values and weights analyzed using both SAS proc glimmix and the EdSurvey package are shown in the appendix and results are similar.

Conclusion

The current manuscript demonstrates additional functions that extend the use of the WeMix package to allow for the pooling of MLM results from models using plausible values. Such a feature is required for the proper analysis of LSA data with outcomes that use plausible values. In addition, functions are introduced that allow for model comparisons using likelihood ratio tests and allow results to be exported into other formats for easier editing.

Appendix A. Two-level multilevel model results using five plausible values and weights analyzed using SAS proc glimmix ($n = 3136$)

	RI	RS
(Intercept)	489.520***	486.730***
	(8.446)	(8.237)
st29q03Agree[1]	− 11.287 +	− 10.533 +
	(5.988)	(5.904)
st29q03Disagree[1]	− 19.526**	− 17.441**
	(6.058)	(6.093)
st29q03Strongly disagree[1]	− 39.929***	− 36.910***
	(7.036)	(7.077)
sc14q02Very little[2]	− 22.549	− 22.702
	(16.804)	(15.494)
sc14q02To some extent[2]	− 17.508	− 16.158
	(11.845)	(12.326)
sc14q02A lot[2]	− 27.257***	− 44.538***
	(11.777)	(9.891)
st04q01Male[3]	8.384**	8.853**
	(3.107)	(3.089)
escs	25.918***	27.433***
	(2.109)	(2.537)
schoolid.(Intercept)	1331.751***	1375.054***
	(356.490)	(321.389)
Residual	5288.857***	5019.517***
	(159.913)	(151.381)
schoolid.escs		339.073***
		(67.403)
$\overline{AIC}$	25,592,616	25,504,872
$\overline{BIC}$	25,592,649	25,504,908

RI random intercepts model. *RS* random slope model. Robust standard errors within parenthesis. [1]Strongly agree is the reference level. [2]Not at all is the reference level. [3]Female is the reference level. + $p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Appendix B. Two-level multilevel model results using five plausible values and weights analyzed using the EdSurvey package in R ($n = 3136$)

	RI	RS
(Intercept)	489.507	486.441
	(8.408)	(8.268)
st29q03Agree[1]	− 11.286	− 10.536
	(5.989)	(5.894)
st29q03Disagree[1]	− 19.534	− 17.404
	(6.050)	(6.079)
st29q03Strongly disagree[1]	− 39.949	− 36.881
	(7.025)	(7.050)
sc14q02Very little[2]	− 22.927	− 23.178
	(16.690)	(15.777)
sc14q02To some extent[2]	− 17.605	− 13.748
	(11.766)	(11.881)
sc14q02A lot[2]	− 30.050	− 35.744
	(7.748)	(8.593)
st04q01Male[3]	8.395	8.881
	(3.103)	(3.088)
escs	25.883	27.123
	(2.105)	(2.441)
schoolid.(Intercept)	1398	1380.7
	(327.5)	(303.89)
Residual	5295	5021.0
	(152.5)	(137.82)
schoolid.escs		324.4
		(63.53)

RI random intercepts model. *RS* random slope model. Robust standard errors within parenthesis. [1]Strongly agree is the reference level. [2]Not at all is the reference level. [3]Female is the reference level. *p*-values are not shown when using the EdSurvey package

Acknowledgements
No further known acknowledgements should be stated.

Author contributions
The author read and approved the final manuscript.

Funding
No funding was received for this work.

Availability of data and materials
I used PISA public use data which is available from https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm. The data may also be loaded in R by entering: `data(pisa2012, package='MLMusingR')` as long as the MLMusingR package is installed.

Declarations

Ethics approval and consent to participate
The present study worked with previously collected PISA data. Therefore, the source data is already anonymized, free, and publicly available. Consequently, ethics approval for this study was not requested.

Consent for publication
Not applicable.

Competing interests
The author reports no competing interests.

Received: 25 May 2023 Accepted: 29 January 2024
Published online: 07 March 2024

References

Aparicio, J., Cordero, J. M., & Ortiz, L. (2021). Efficiency analysis with educational data: how to deal with plausible values from international large-scale assessments. *Mathematics, 9*(13), 1579.

Arel-Bundock, V., Gassen, J., Eastwood, N., Huntington-Klein, N., Schwarz, M., Elbers, B., McDermott, G., & Wallrich, L. (2022). modelsummary: Summary tables and plots for statistical models and data: Beautiful, customizable, and publication-ready (1.2.0) [Computer software]. https://CRAN.R-project.org/package=modelsummary

Bailey, P., Kelley, C., Nguyen, T., & Huo, H. (2023). WeMix: Weighted mixed-effects models using multilevel pseudo maximum likelihood estimation. https://CRAN.R-project.org/package=WeMix

Bailey, P., Lee, M., Nguyen, T., & Zhang, T. (2020). Using EdSurvey to analyse PIAAC data. In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment* (pp. 209–237). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_9

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Springer.

Baysu, G., Agirdag, O., & De Leersnyder, J. (2023). The association between perceived discriminatory climate in school and student performance in math and reading: A cross-national analysis using PISA 2018. *Journal of Youth and Adolescence, 52*(3), 619–636. https://doi.org/10.1007/s10964-022-01712-3

Berkhof, J., & Snijders, T. A. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics, 26*(2), 133–152.

Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software, 81*, 1–44. https://doi.org/10.18637/jss.v081.i07

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F., & York, R. (1966). *Equality of educational opportunity*. Government Printing Office.

Consentino, F., & Claeskens, G. (2010). Order selection tests with multiply imputed data. *Computational Statistics & Data Analysis, 54*(10), 2284–2295.

Ferron, J., Dailey, R., & Yi, Q. (2002). Misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*(3), 379–403. https://doi.org/10.1207/S15327906MBR3703_4

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database* (2nd edn). TIMSS & PIRLS International Study Center. https://timss2019.org/international-database/downloads/TIMSS-2019-User-Guide-for-the-International-Database-2nd-Ed.pdf

Gelman, A., & Rubin, D. B. (1994). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25*, 165–173.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the r package pan. *SAGE Open, 6*(4), 2158244016668220. https://doi.org/10.1177/2158244016668220

Grund, S., Lüdtke, O., & Robitzsch, A. (2023). Pooling methods for likelihood ratio tests in multiply imputed data sets. *Psychological Methods*. https://doi.org/10.1037/met0000556

Herget, D., Dalton, B., Kinney, S., Smith, W. Z., Wilson, D., & Rogers, J. (2019). US PIRLS and ePIRLS 2016 technical report and user's guide. NCES 2019-113. National Center for Education Statistics.

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education, 9*(1), 17. https://doi.org/10.1186/s40536-021-00109-1

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly, 33*(3), 492–499. https://doi.org/10.1037/spq0000272

Huang, F. L., Wiedermann, W., & Zhang, B. (2022). Accounting for heteroskedasticity resulting from between-group differences in multilevel models. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2022.2077290

Joncas, M. (2007). PIRLS 2006 sampling weights and participation rates. In M. Martin, I. Mullis, & A. Kennedy (Eds.), *PIRLS 2006 Technical report* (pp. 105–130). TIMSS & PIRLS International Study Center.

Karakolidis, A., Pitsia, V., & Cosgrove, J. (2022). Multilevel modelling of international large-scale assessment data. In M. S. Khine (Ed.), *Methodology for multilevel modeling in educational research* (pp. 141–159). Springer Singapore. https://doi.org/10.1007/978-981-16-9142-3_8

Kim, J.-S., Anderson, C. J., & Keller, B. (2013). Multilevel analysis of assessment data. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 389–425.

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. https://doi.org/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. https://doi.org/10.1080/03610926.2017.1383429

Li, K.-H., Meng, X.-L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 65–92.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22.

Lorah, J. (2022). Analyzing large-scale assessment data with multilevel analyses: Demonstration using the Programme for International Student Assessment (PISA) 2018 data. In M. S. Khine (Ed.), *Methodology for multilevel modeling in educational research* (pp. 121–139). Springer Singapore. https://doi.org/10.1007/978-981-16-9142-3_7

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education, 9*(1), 6. https://doi.org/10.1186/s40536-021-00099-0

Meinck, S. (2015). Computing sampling weights in large-scale assessments in education. *Survey Methods: Insights from the Field*, 1–13.

Mirazchiyski, P. V. (2021). RALSA: The R analyzer for large-scale assessments. *Large-Scale Assessments in Education, 9*, 1–24.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Muthén, L., & Muthén, B. (1998). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/revue Internationale De Statistique*. https://doi.org/10.2307/1403631

Pinheiro, J., Bates, D., & R Core Team. (2022). *nlme: Linear and nonlinear mixed effects models*. https://CRAN.R-project.org/package=nlme

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rathbun, A., Huang, F., Meinck, S., Park, B., Ikoma, S., & Zhang, Y. (2021). *Multilevel modeling with large-scale international datasets*. American Educational Research Association, Virtual conference.

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

Raudenbush, S., & Congdon, R. (2021). *HLM 8: Hierarchical linear and nonlinear modeling* (Version 8) [Computer software]. Scientific Software International, Inc.

Ringienė, L., Žilinskas, J., & Jakaitienė, A. (2022). ILSA data analysis with R packages. *Modelling, Computation and Optimization in Information Systems and Management Sciences: Proceedings of the 4th International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences-MCO 2021 4*, 271–282.

Robinson, D., Hayes, A., & Couch, S. (2022). *broom: Convert statistical objects into tidy tibbles*. https://CRAN.R-project.org/package=broom

Robitzsch, A., & Oberwimmer, K. (2022). *BIFIEsurvey: Tools for survey statistics in educational assessment*. https://CRAN.R-project.org/package=BIFIEsurvey

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). Wiley.

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.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545–571. https://doi.org/10.1207/s15327906mbr3304_5

Smith, P. K., & López-Castro, L. (2017). Cross-national data on victims of bullying: How does PISA measure up with other surveys? *International Journal of Developmental Science, 11*(3–4), 87–92. https://doi.org/10.3233/DEV-170227

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44*(6), 711–740. https://doi.org/10.1080/00273170903333574

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifying the error covariance structure in linear mixed models for longitudinal data. *Methodology, 4*(1), 10–21. https://doi.org/10.1027/1614-2241.4.1.10

Wickham, H. (2021). *tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Yamashita, T., Smith, T. J., & Cummins, P. A. (2021). A practical guide for analyzing large-scale assessment data using Mplus: A case demonstration using the program for international assessment of adult competencies data. *Journal of Educational and Behavioral Statistics, 46*(4), 501–518. https://doi.org/10.3102/1076998620978554

Zhang, T., Bailey, P., & Lee, M. (2020). *Using EdSurvey to analyze TIMSS data*. https://www.air.org/sites/default/files/edsurvey-TIMSS-pdf.pdf

Publisher's Note

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