Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 22 Number 2, April 2017 ISSN 1531-7714

# Hierarchical Linear Modeling with Maximum Likelihood, Restricted Maximum Likelihood, and Fully Bayesian Estimation

Peter Boedeker, *University of North Texas*

Hierarchical linear modeling (HLM) is a useful tool when analyzing data collected from groups. There are many decisions to be made when constructing and estimating a model in HLM including which estimation technique to use. Three of the estimation techniques available when analyzing data with HLM are maximum likelihood, restricted maximum likelihood, and fully Bayesian estimation. Which estimation technique is employed determines how estimates can be interpreted and the models that may be compared. The purpose of this paper is to conceptually introduce and compare these methods of estimation in HLM and interpret the computer output that results from using them. This is done for the intraclass correlation, parameter estimates, and model fit indices using a simulated dataset that is available online. The statistical program R is utilized for all analyses and syntax is provided in Appendix 1. This paper is written to aid applied researchers who wish to better understand the differences between the estimation techniques and how to interpret their HLM results.

Hierarchical linear modeling (HLM) is an effective tool in social and educational research for analyzing data collected from groups. As with any analytical model, there are many decisions to be made when constructing and estimating a model in HLM (Peugh, 2010). One of these decisions is the estimation technique to be used. Raudenbush and Bryk (2002) detail methods of estimation in HLM, including maximum likelihood (ML), restricted maximum likelihood (REML), and fully Bayesian estimation. ML or REML is typically the default setting for software estimating an HLM while fully Bayesian estimation is not. There are meaningful differences between estimation techniques and if these are not thoughtfully considered a poor choice may be inadvertently made.

The purpose of this paper is to conceptually introduce and compare methods of statistical estimation in HLM and how the computer output resulting from the use of each may be interpreted. The analyses are conducted in R (Version 3.3.1; R Core Team, 2016), a free program available to anyone with an Internet connection. Syntax is provided in Appendix 1 for all analyses conducted in this paper and sample output with references to tables displayed in the paper is available in Appendix 2. The techniques to be compared are maximum likelihood, restricted maximum likelihood, and fully Bayesian estimation. Empirical Bayes is another estimation technique that generally gives "shrunken" estimates compared to ML and REML, further discussion of which is not included in this paper but the curious reader is directed to Raudenbush and Bryk (2002). The output resulting from HLM implemented with ML, REML, and fully Bayesian estimation techniques will be compared for the intraclass correlation (ICC), estimates for intercepts and slopes, and model fit indices. This paper is written to aid applied researchers who wish to better understand the differences between the estimation techniques and how to interpret their HLM results. To begin, an overview of the HLM framework is provided.

## Hierarchical Linear Modeling

HLM in the social and educational setting models the interrelationships between people that live or interact in groups. For example, in a research study students may be selected from many classrooms. Students from the same classroom have common experiences and relationships that influence how they may respond to survey items or influence their measured ability on assessments. Having peers with positive attitudes may make one's own attitude more positive or having an exceptional teacher may make everyone in a given classroom score higher on a math exam. This dependence on the class in which a student is enrolled in regards to the dependent variable violates the assumption of statistical independence. To be statistically independent, the observed responses or scores of individuals in the study must be independent of one another. This assumption is not tenable when students are from groups, such as classrooms or schools. When this violation occurs, the standard errors of parameter estimates in ordinary least squares regression will be underestimated, leading to higher rates of rejecting the null hypothesis (Osborne, 2000). HLM can be used to account for the violation of the independence assumption by modeling the hierarchy of the grouping structure.

In HLM, the hierarchy of the grouping structure is comprised of levels, each with information pertinent to that level. In an educational setting, the first level may contain individual student information. This could include independent variables such as the race, sex, or previously measured ability of each student and a dependent variable such as standardized test score. The second level is a grouping level and can be the classroom in which a student learns. Independent variables in this second level could be the teacher's age, number of years of experience, or class size. The second level accommodates for the dependence of student measurements within the same classroom. Another type of grouping is repeated measures, in which measurements at different time points (first level) are grouped within the individual (second level) who was measured. HLM can be further extended to include higher levels. For instance, students, classrooms, and schools may be three levels of data. For the purposes of this paper the discussion will be limited to two levels. These levels interact through related regression equations.

HLM is a generalization of regression analysis, modeling the intercept and slopes in such a way as to either be constrained to a single value across groups or allowed to vary depending on group membership (Gelman, 2006). This is accomplished by equating the intercept and slope coefficients of the first level equation with equations on the second level. For example, a model with an intercept that is allowed to vary depending on group membership and a single first level predictor with a slope coefficient that does not vary by group membership would be specified as:

$$\begin{aligned}
&\text{Level 1: } y_{ij} = \beta_{0j} + \beta_1 x_{ij} + r_{ij}, && r_{ij} \sim N(0, \sigma_y^2) \\
&\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, && u_{0j} \sim N(0, \sigma_{\beta_0}^2) \\
&\qquad \beta_1 = \gamma_{10}. &&
\end{aligned} \qquad (1)$$

In the first level, the response of person *i* in group *j* is equal to the intercept of group *j* plus the product of the independent variable of person *i* and the coefficient $\beta_1$ (which is the same across groups). The intercept ($\beta_{0j}$) and slope ($\beta_1$) are modeled by second level equations. The intercept is comprised of two terms, $\gamma_{00}$, which is the mean of all of the intercept terms for the groups, and $u_{0j}$, a residual term that represents the deviance of groups from the all-groups mean ($\gamma_{00}$). The residual term is normally distributed with a mean of zero and variance ($\sigma_{\beta_0}^2$). Including the residual term in the intercept equation allows the intercept to vary according to group membership. The coefficient of the predictor is invariant, made evident by the exclusion of a residual term in the second-level equation. An invariant predictor coefficient means a change in the independent variable produces the same change in the dependent variable, regardless of group membership.

Terminology describing the different terms in HLM as "fixed" or "random" is common and understanding their differences and when to use them is necessary when specifying a model. A fixed effect is a single value for all groups. If the intercept is specified as a fixed effect, then a single value is estimated for the intercept of all groups. If the coefficient of an independent variable is specified as a fixed effect, then the coefficient for that independent variable will be the same regardless of group membership. The second-level equation for a fixed effect does not have a residual term. In equations (1), the coefficient of $x_{ij}$ is a fixed effect. Fixed effects can be used when the intercept or slope of all groups are the same. If all terms in the model were fixed effects, then the model would be a standard regression model.

A random effect allows each group to have a different parameter estimate. If an intercept is a random effect, then a separate intercept is estimated for each group. Likewise, if the coefficient of an independent variable is a random effect then each group will have a different estimate for that coefficient. An estimate is made random by the summation of a mean and a residual term in the second-level equation. In equations (1) the intercept is a random effect. The $\gamma_{00}$ term is the grand mean of the intercepts across all groups and the residual term ($u_{0j}$) is taken to be a value from a normal distribution with mean zero and a variance. Using random effects for both the intercept and the coefficients may mirror reality more accurately, even if differences between groups are small. However, a large sample is necessary when estimating many random effects because each group has a parameter that must be estimated, instead of estimating a single value shared by all groups. This may be a problem if there are many groups and many effects to be estimated. If it is possible to use random effects for all parameter estimates, it is the recommended approach (Gelman & Hill, 2007).

HLM models can be described in different ways. Gelman and Hill (2007) describe models by which of the terms are allowed to vary. For instance, the equations (1) represent a varying-intercepts, fixed-slope model. It is named so because the intercept is the only aspect of the model that is allowed to differ by group. This approach will be taken to describe models presented in this paper. Additionally, because the discussion will often turn to the components of random and fixed effects, the terms “fixed” and “random” will be used to describe the components of each term similarly to the approach taken by Hayes (2006). In discussing the terms by their components, the components that are a single value are considered fixed and those that are normally distributed with a mean of zero and a variance are considered random. In the above set of equations, the intercept has a fixed component ($\gamma_{00}$) and a random component ($u_{0j}$). The slope does not vary but is instead equal to a single fixed component ($\gamma_{10}$).

## Estimation Techniques

Maximum likelihood, restricted maximum likelihood (also called residual maximum likelihood) and fully Bayesian estimation are three methods of estimating the fixed components and variances of the random components in HLM. Each estimation technique has limitations and assumptions that must be taken into consideration when determining which to use. These three techniques are here briefly described.

### Maximum Likelihood

Maximum likelihood estimation yields simultaneous estimation of fixed and random components by maximizing the likelihood function of the data (Corbeil & Searle, 1976). These estimates are those parameter values that were most likely to have produced the observed data (Myung, 2003). This maximization may not be possible in closed form; therefore, an iterative procedure such as expectation-maximization or fisher scoring may be required (Raudenbush & Bryk, 2002). ML works well when sample sizes are large and when there are many groups at the second level. However, when either or both of these are small, the variances are negatively biased (Peugh, 2010; Raudenbush & Bryk, 2002). To account for these limitations, REML can be employed.

### Restricted Maximum Likelihood

The primary difference between ML and REML is in the estimation of variances (Peugh, 2010). In ML, the variances are estimated as if the fixed components were known and therefore measured without error. REML accounts for the fact that fixed components were estimated when estimating variances. By doing so, REML estimates are less biased than ML estimates, particularly when the number of groups is small. The mathematics of REML is beyond the scope of this paper, as it requires matrix algebra with error contrasts, but the process is outlined here. First, an ordinary least squares regression model is fit using only the fixed components. The residuals of this regression are then modeled and variances and covariances are estimated by maximizing the likelihood of the residuals (Searle, Casella, & McCulloch, 2006, pg. 250). This process will require an iterative procedure to determine final variance estimates (Corbeil & Searle, 1976), but the computer does this. Generalized least squares (GLS) estimates for the fixed components are then derived using the variances and covariances estimated in the previous step. The GLS estimates may be the same as the original fixed components regression, but this is not always the case and GLS estimates are retained. REML estimates for variances are typically larger than ML estimates, particularly for higher order variances. When the number of groups is small, the variance estimates when using ML will be smaller than the estimates when using REML approximately by a factor of

$$\frac{(J-F)}{J}, \qquad (2)$$

where *J* is the number of groups and *F* is the number of fixed components (Raudenbush & Bryk, 2002). As the number of groups increases relative to the number of fixed components the difference between REML and ML diminishes in regards to variance estimates. Differences do remain between REML and ML in regards to model fit indices.

Model selection is to be discussed later in this paper, but an important caveat when using REML estimation can be made here. Because of the manner in which REML adjusts for the uncertainty of the fixed components in the estimation of residual variances, models that are fit using REML can be compared if they differ only in their random components (Peugh, 2010). In REML, the random components are estimated so as to explain the variance left after removing the influence of the fixed components with the ordinary least squares regression. If models have different fixed components, then the remaining variance to be explained by the random components is no longer the same across models and comparisons are not sensible. Therefore, caution must be taken when fitting and comparing models using REML. The final estimation technique is Bayesian estimation.

## Bayesian Estimation

Full explanation of Bayesian estimation and its application to various research methods are beyond the scope of this article. A brief introduction is provided here, but resources are available for the curious reader. For article introductions see Kruschke (2013) and Louis (2005). For textbooks on the topic, see Carlin and Louis (2009), Gelman et al. (2013), and Kruschke (2015).

In the application of fully Bayesian estimation, researchers use probability distributions to model the credibility of possible parameter values. In its simplest form, three distributions are considered. The first is the prior distribution, which models the prior belief that each possible parameter value is true before the analysis of new data. The prior belief can be specified based on previous research or expert opinion. The second distribution is the data likelihood, which is the likelihood of parameter values based only on the data collected in a given study. This is the same likelihood as was maximized using ML and REML. The prior and the likelihood are mathematically combined with the use of Bayes' Theorem. The outcome of a Bayesian analysis, the posterior, is the third probability distribution. The posterior models the probability of each possible parameter value being true, given the prior and likelihood. The greatest difference between Bayesian estimation and the other estimation techniques is in the use of prior and posterior distributions. These are further detailed next.

The prior distribution can take many shapes depending on the credibility the researcher wishes to assign to parameter values a priori. Two broad classifications of prior distributions are uninformative or informative. Uninformative priors are relatively flat compared to informative priors, indicating that any value for the parameter is plausible a priori. For example, an uninformative prior in the context of student ability measured by a test instrument may be a normal distribution with mean zero and standard deviation 100. Such a broad distribution gives nearly equal credibility to all possible (and impossible) parameter values. The posterior is essentially a weighted combination of the prior and likelihood distributions, so an uninformative prior allows the data the greatest role in determining the posterior. HLM is typically used with a large number of subjects and groups, in which case the influence of the prior on the posterior is minimal. The prior has the greatest influence on the posterior when the number of groups or samples sizes within each group is small or when an informative prior is used.

The use of an informative prior is justified when evidence exists indicating that certain parameter values are more likely to be true than others. Instead of assigning equal credibility for all values a priori, an informative prior can be used to assign higher credibility to values that have been found in the literature or are deemed more reasonable by experts. The results of Bayesian estimation would be interpreted in the same manner across prior specifications, with consideration given to the prior and the data likelihood. In the analysis examples provided later in this paper, only uninformative priors will be used. After a prior has been specified and the information from it and the likelihood have been combined, the posterior distribution is used for estimation.

From the posterior distribution point and interval estimates are determined. Using the posterior distribution, the researcher can identify the parameter value that is most likely to be true, based on the prior and likelihood, and make probabilistic statements concerning its credibility. Point estimates can be determined by finding the mean, median, or mode of the posterior distribution. The highest density interval (HDI; Kruschke, 2015) is a range of values with a given probability of containing the true value. Because the posterior is a probability distribution, the researcher need only sum the area under the posterior curve to determine the probability of any range of values. The 95% HDI indicates the range of values in which there is a 95% chance that the true value lays. A confidence interval does not have the same probabilistic interpretation but instead must be understood in the context of replication (Greenland et al., 2016).

In most cases, the posterior distribution is impossible to mathematically derive and instead Markov Chain Monte Carlo (MCMC) simulation techniques must be employed. Samples from the posterior distribution are repeatedly taken, creating a distribution of sampled values. The samples are then compiled into a distribution used as the posterior. The sampling process starts with a single value and iteratively converges to the posterior. Multiple starting values can be used to produce separate "chains" of resampling. These chains are then combined after thousands of iterations. With enough samples the empirical posterior will approach the mathematical posterior. Specialized software has been developed for conducting this procedure, including Bayesian inference Using Gibbs Sampling (BUGS; Gilks, Thomas, & Spiegelhalter, 1994), Just Another Gibbs Sampler (JAGS; Plummer, 2003), and Stan (Stan Development Team, 2016). To determine if enough sampling has occurred, visually monitoring the chains for convergence is recommended. This is accomplished by plotting the sampled values of each chain. If the values all fall within a consistent range, then convergence to the posterior distribution has been achieved. As a result of sampling variability within chains, parameter estimates for the exact same data may not be identical if the same analysis is conducted again. For the interested reader using the syntax in Appendix 1 to replicate the results found later in this paper, parameter estimates that differ somewhat are expected.

### Which Estimation Technique to Use?

Considering the three estimation techniques previously discussed, the next natural question is, "which do I use?" ML and REML are more commonly used whereas fully Bayesian estimation is used less frequently. The lower use of fully Bayesian estimation is likely due to the required use of specialized software and the fact that it is infrequently taught in graduate education programs. Even though it is less frequently used, Bayesian estimation allows for intuitive probabilistic interpretations of results based on the posterior distribution. The author recommends Bayesian estimation in HLM. Apart from this recommendation, decisions concerning which estimation technique to use depend on the structure of the data, particularly the number of groups.

The number of groups is important when deciding which estimation technique to use. When the number of groups is small, REML will produce less biased estimates of variances compared to ML. What number is small? This depends on many aspects of your data and may not be known a priori. Once data is collected, the model can be estimated using both ML and REML. If the variance estimates are very different between the two, then the REML results should be used for interpretation. If the results are similar, then the ML results can be used, allowing for more model comparisons. If using Bayesian estimation, a small number of groups should prompt use of the posterior mode instead of the posterior mean as the variance estimate (Browne & Draper, 2006). When the number of groups is small, the prior has a greater influence on the posterior distribution. An uninformative prior assigns low credibility to an extremely large range of values. Even with extremely low posterior credibility for extreme outliers, the posterior mean will be influenced by those values. Therefore, the posterior mode will yield more accurate results. When the number of groups is large, the mean and mode will render similar estimates. Thus, the posterior mean and posterior mode may be compared and if differences exist, the mode should be interpreted. See Table 1 for a brief summary of the differences between the estimation techniques.

**Table 1.** Comparison Across Estimation Techniques

| | ML | REML | Bayesian |
|---|---|---|---|
| Advantage | Compare models with different fixed components | More accurate variance estimates when the number of groups is small (compared to ML) | Intuitive probabilistic interpretations of point and interval estimates |
| Disadvantage | Poor estimation of variances when the number of groups is small (Compared to REML) | Compare models that differ only in random components | Less frequently used in the literature; Requires use of specialized software packages |

Deciding which estimation technique to use is something that should not be left to software defaults. Data can be analyzed using ML and REML and if higher order variance estimates are different, REML results should be interpreted. Bayesian methods offer probabilistic interpretations for point and interval estimates that ML and REML do not, but require the specification of a prior distribution and use of specialized software. When the number of groups is small, the posterior mode should be interpreted instead of the posterior mean. The remainder of this paper will focus on the ICC, parameter estimates, and fit indices when using ML, REML, and fully Bayesian estimation. An introduction to each is provided and computer output explained. The same data set, described next and available online (see Appendix 1 for downloading instructions), will be analyzed for all examples.

### Example Dataset

Hox (2010) provides a simulated data set constructed for teaching purposes. The complete data set consists of 2,000 students in 100 schools. Because differences between estimation techniques are most obvious when the number of groups is small, only the 101 students in the first 5 schools will be used. If the full data set were used, estimates across techniques would be nearly identical. Strong multilevel effects exist with students (level 1) grouped within schools (level 2). The dependent variable is a student popularity score on a scale from 1-10. The student's sex is included as the only level-1 predictor. No school (level 2) predictors will be used.

The analyses for examples in this paper were conducted in R (Version 3.3.1; R Core Team, 2016). For ML and REML estimation, the packages lme4 (Version 1.1-12; Bates, et al., 2015) and sjstats (Version 0.7.1; Ludecke, 2016) were used. Bayesian estimation was conducted using the package R2jags (Version 0.5-7; Su & Yajima, 2015). All of the programs are free and code is provided in Appendix 1 for readers to replicate results. Additionally, running the first eleven lines of code in Appendix 1 will load the complete dataset (of 2,000 students) and reduce the dataset to the same as what will be used for the remainder of this paper.

For Bayesian estimation, three chains were run for 21,000 iterations (samples) per chain and a burn-in period of 1,000 iterations. A burn-in period accounts for the fact that MCMC is an iterative process that may take several samples before converging to the actual posterior. By removing the first 1,000 samples the posterior approximated by the remaining 20,000 is more likely to be reflective of the actual posterior and not influenced by those values that existed only because the algorithm was attempting to converge. The chains can be monitored, by plotting, to ensure convergence was achieved. Convergence can be visually identified when the iterated values all fall within a consistent range.

## Intraclass Correlation

By employing HLM, the researcher is recognizing the potential that variability is occurring at both the individual level and the group level. Whether or not variability is occurring at the group level and if so, how much of the total variability can be attributed to the grouping level, is determined in the calculation of the intraclass correlation (ICC). A higher ICC indicates that a greater amount of variability is occurring at the group level, meaning a greater violation to the assumption of independence and justifying the use of HLM.

An unconditional model is used to calculate the initial ICC. The unconditional model is a varying intercept model with no predictors at any level. The equations for the unconditional model are:

$$\begin{array}{lll} \text{Level 1: } y_{ij} = \beta_{0j} + r_{ij}, & r_{ij} \sim N(0, \sigma_y^2) & \\ \text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, & u_{0j} \sim N(0, \sigma_{\beta_0}^2) & \end{array} \quad (3)$$

The formula for the ICC is

$$\frac{\sigma^2_{\beta_0}}{\sigma^2_{\beta_0} + \sigma^2_y} \qquad (4)$$

The numerator of the ICC is the residual variance on the second level and the denominator is the total residual variance in the model. The ICC is the proportion of the total residual variance that can be attributed to the grouping level.

As a proportion, the ICC ranges from 0 to 1. An ICC equal to zero indicates that there is zero variability on the grouping level. If this is the case, then there is no justification for employing HLM and a less complex regression model can be used. An ICC of one indicates that the difference in scores is only found between groups and not within. Neither of these extremes is very likely. There is no set rule for what ICC would necessitate the use of HLM, but values as low as 0.05 may be sufficient (Kreft & de Leeuw, 1998).

### Comparison

The ICC can be calculated for the popularity data. Table 2 shows the estimated ICCs when using ML, REML, and Bayesian methods. Across all ICC values there is strong evidence that variability is occurring between the groups, supporting the use of HLM. For instance, using the ML estimate, 78% of the variability between student popularity scores can be attributed to differences between schools.

**Table 2**. Intraclass Correlation by Estimation Technique

| | Estimation Technique | | | |
|---|---|---|---|---|
| | ML (SE) | REML (SE) | Bayesian Mean (95% HDI) | Bayesian Mode |
| ICC | 0.78<br>(0.029) | 0.81<br>(0.0322) | 0.86<br>[0.63, 0.99] | 0.94 |

Note. The 95% HDI is the same for the Bayesian Mean and Mode.

Using lme4 with ML or REML, the ICC can be calculated from summary output. Table 3 shows a portion of the output when using REML to estimate the unconditional model. The intercept residual variance is $\sigma^2_{\beta_0}$ and the first-level residual variance is $\sigma^2_y$. The values for $\sigma^2_{\beta_0}$ and $\sigma^2_y$, 2.1242 and 0.4898, respectively, can be used in equation 4 to calculate the ICC. The standard error, however, cannot be estimated simply from this output. Instead, the se() function in the R package sjstats can be used to find both the ICC estimate and the bootstrapped standard error estimate.

**Table 3**. Random Effects Summary Statistics for the Unconditional Model Fit with REML

| Random effects: | | | |
|---|---|---|---|
| Groups | Name | Variance | Std. Dev. |
| school | (Intercept) | 2.1242 | 1.4574 |
| Residual | | 0.4898 | 0.6998 |

Note. Table presents a portion of the output as it appears in R using the lmer command in lme4

Table 4 shows typical summary output for a Bayesian analysis using R2jags. Recall that point estimates and HDIs are derived from a posterior distribution. Therefore, the mean of the posterior is presented as a point estimate and the mode can be determined by further functions in R. For this model and data, the posterior mean for the ICC is estimated to be 0.86, indicating that 86% of the variability in the dependent variable can be attributed to differences between groups. The posterior mode is 0.94, indicating that an even higher proportion of the variability can be attributed to school enrollment. The area between the 2.5 and 97.5 percentile values captures 95% of the area under the curve. The 95% HDI ranges from 0.63 to 0.99, indicating that there is a 95% chance that the true value of the ICC falls within that range given the prior and likelihood. The HDI is the same regardless of using the posterior mean or mode as the parameter estimate.

**Table 4**. Summary Statistics for the Unconditional Model Fit with Fully Bayesian Estimation

| | mean | sd | 2.5% | 97.5% |
|---|---|---|---|---|
| Deviance | 215.62 | 3.57 | 210.67 | 224.25 |
| icc | 0.86 | 0.095 | 0.63 | 0.99 |
| mu.a | 5.89 | 1.39 | 3.43 | 8.33 |
| sigma.a | 2.39 | 2.05 | 0.94 | 6.60 |
| sigma.y | 0.71 | 0.05 | 0.62 | 0.82 |

Note. Elements of the full R2jags output have been excluded. "Deviance" is used for model fit, to be discussed later. The "icc" is intraclass correlation, of interest here. "mu.a" and "sigma.a" are the fixed and random components, respectively, for the intercept. "sigma.y" is the residual of the first level.

The ICC is important for justifying the use of HLM. Across the three estimation techniques the estimated ICC values differed. However, whether using ML, REML, or Bayesian estimation, the ICC made evident the need for HLM to appropriately model the relationship between the dependent and independent variables. Once the use of HLM has been justified the parameter estimates are of interest.

## Parameter Estimates

Parameter estimates are derived for both fixed components and the variance or standard deviations of the random components. The fixed component is the average for all groups on the intercept or slope coefficient while the random component indicates the variability in intercepts and slope coefficients that exists across groups. If the intercept does not vary, then in the model a single intercept is estimated for all groups. If a slope coefficient does not vary, then in the model the estimated relationship between the independent variable and the dependent variable does not depend on group membership.

When the intercept or the slopes of a model are allowed to vary, the second level equations will contain both fixed and random components. Consider first a varying intercept. The fixed component is the average of all of the estimated intercepts. If the random component has a large residual variance, then the intercepts estimated across the groups vary widely or there may be outliers. If the residual variance is small, then the intercepts for the different groups are relatively similar to one another. Likewise, a varying slope has a fixed component, representing the average slope value across all groups, and a random component that shows the deviation of the estimated slope coefficients from that average. Allowing more aspects of a model to vary increases the complexity of the model because more parameters must be estimated. For instance, for the current example, allowing the intercept to vary by group means that a separate intercept must be estimated for each group.

What follows are the parameter estimates for the varying-intercept and varying-slope model with a single first level predictor. The dependent variable is popularity score, the first level predictor is the sex of the student, and the grouping variable is the school that the student attends. The two second-level residuals are allowed to correlate, a relationship that is assumed when using lme4 but must be specified in the R2jags model. The equations for the varying-intercept and varying-slope model are:

$$\begin{aligned}
&\text{Level 1: } y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + r_{ij}, \quad r_{ij} \sim N(0, \sigma_y^2)\\
&\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}\\
&\qquad\quad \beta_{1j} = \gamma_{10} + u_{1j}, \qquad (5)\\
&\qquad\quad \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00}^2 & \tau_{01} \\ \tau_{01} & \tau_{11}^2 \end{pmatrix}\right)
\end{aligned}$$

The popularity score of student *i* in school *j* is equal to the intercept of school *j* plus the product of the sex indicator for student *i* in school *j* and the slope coefficient for school *j*. This model differs from the varying-intercepts only model (see equations 1) by including a random component for the slope coefficient of sex, thereby allowing that coefficient to vary by school. The correlation between second-level residuals allows for the relationship to be estimated between the deviations of the school from the all-school average for the intercept and the all school-average for the coefficient of sex.

### Comparison

Table 5 shows the estimates of the fixed and random components, where the random component values are the residual standard deviations instead of variances. The estimates using ML and REML are accompanied by bootstrapped 95% confidence intervals and fully Bayesian estimates by 95% HDIs.

**Table 5.** Parameter Estimates Using ML, REML, and Fully Bayesian Estimation

| | | Estimation Techniques | | | |
|---|---|---|---|---|---|
| Component | | ML [95% CI] | REML [95% CI] | Bayesian Mean [95% HDI] | Bayesian Mode |
| Fixed | | | | | |
| | Intercept | 6.17 [5.12, 7.34] | 6.17 [4.87, 7.46] | 6.16 [3.32, 8.97] | 6.15 |
| | Sex | -0.57 [-0.85, -0.30] | -0.56 [-0.92, -0.27] | -0.57 [-1.14, -0.01] | -0.56 |
| Random | | | | | |
| | Intercept | 1.28 [0.38, 1.88] | 1.44 [0.46, 2.41] | 2.70 [0.98, 7.49] | 1.58 |
| | Sex | 0.12 [0.01, 0.38] | 0.20 [0.02, 0.54] | 0.41 [0.01, 1.50] | 0.19 |
| | Residual | 064 [0.54, 0.73] | 0.64 [0.54, 0.72] | 0.65 [0.56, 0.75] | 0.64 |
| Correlation | | 0.14 [-1, 1] | 0.05 [-1, 1] | 0.01 [-0.90. 0.90] | -.017 |

Note. Information is consolidated from output using lme4 and R2jags. Bootstrapped 95% confidence intervals were derived using the confint() function. The Bayesian estimates show the posterior mean as the point estimate and accompanying 95% HDI. The 95% HDI is the same for the Bayesian Mean and Mode.

There are two fixed components, one for the intercept were estimated for each the intercept ($u_{0j}$), sex ($u_{1j}$), and first level residual ($r_{ij}$).

When using lme4, the estimates for ML and REML results are presented without p-values. This is because the null distribution and degrees of freedom necessary to derive p-values can only be approximated, if at all determined, when using HLM. While commands exist in R for p-value approximations, they are only as good as the accuracy of the approximation. Hence, the author of the lme4 package chose to exclude such calculations from standard output (see Bates, 2006).

Interpretations of the REML results follow. The average intercept value for student popularity across all schools was 6.17 (recall, the dependent variable was on a scale of 1 to 10). An intercept term was estimated for each school and deviations of these values from the average intercept of all schools were assumed to be normally distributed. The standard deviation of the intercept residuals was estimated to be 1.44. By making the coefficient of sex random, the difference between the popularity score of boys and girls was allowed to be dependent on the school in which the student was enrolled. This means that, in regards to popularity score, being a boy in one school does not necessarily mean the same thing as being a boy in another school. On average, boys were 0.56 points lower in popularity than girls, although this also varied across schools with a standard deviation of 0.20. The residual term in the random components output shows that the error on the first level was distributed with a standard deviation of 0.64. Finally, the second level terms were slightly positively correlated (0.05), although this estimate is extremely uncertain with a bootstrapped 95% confidence interval ranging between -1 and 1.

The ML and REML results differ in the estimates of second level residuals. For the intercept and the sex variable the estimates are larger for REML than for ML. This is to be expected given the negative bias of ML when estimating variances, particularly when the number of groups is small. In the reduced dataset analyzed here, the number of groups was only five, a sufficiently small number to cause these notable differences in estimates. Estimates of the fixed effects and the standard deviation of the first-level residual were similar, if not identical across the two techniques. The total number of students was 101, sufficiently large ($\gamma_{00}$) and one for the slope ($\gamma_{10}$). Random components for the estimates at the student level to be similar with both ML and REML.

Bootstrapped 95% confidence intervals were computed for parameters. Bootstrapping is a non-parametric resampling procedure that can be used to calculate confidence intervals for many statistics, including regression coefficients and effect sizes (Banjanovic & Osborne, 2016; Yu, 2003). This, and any other confidence interval, is best understood in the context of replication. If this study were repeated with a new sample 100 times, assuming all of the assumptions of the study were true, then 95 of the resulting intervals would be expected to capture the true value. A single confidence interval does not have a probabilistic interpretation but can provide a range of plausible values (Cumming & Finch, 2004) and information concerning replicability (Cumming, Williams, & Fidler, 2010). To say that a 95% interval has a 95% chance of containing the true parameter, Bayesian methods must be used (Greenland et al., 2016).

The Bayesian mean or mode are similar to ML and REML for point estimates of the fixed components and the first-level residual; however, differences are evident in the second-level estimates of residual standard deviations for the intercept and sex. The random component standard deviation for the intercept has a posterior mean of 2.70 and a posterior mode of 1.58 with a 95% HDI from 0.98 to 7.49. The HDI indicates that there is a 95% chance that the true value of the residual standard deviation for the intercept lays between 0.98 and 7.49. The HDI is the same regardless of whether the posterior mean or posterior mode is used because the HDI represents the area under the posterior curve and is therefore independent of the estimate used. For sex, the posterior mean is 0.41 and the posterior mode is 0.19 with a 95% HDI from 0.01 to 1.50. Comparing the HDI range to the CI range for ML and REML, the HDI is wider for estimates of fixed components and second-level variances. This is the result of both the use of uninformative priors and a small number of groups. The uninformative prior gave credibility to extreme values, yielding a wider HDI. As the number of groups increases, the prior will have less influence on the posterior and HDIs will become increasingly narrow. The final aspect of HLM reviewed is model fit indices.

## Model Fit Indices

When using HLM, models can vary by the number of independent variables as well as how many independent variables are allowed to vary by group. Considering both the fit and complexity of models is important when determining if one model is superior to another (Spiegelhalter, Best, Carlin, & van der Linde, 2002). Fit is typically defined by a deviance measure and complexity by the number of parameters estimated in the model. A more complex model may prove to have a better fit, but models that are too complex may not be valid for making out-of-sample predictions. In HLM, both fit and complexity are taken into consideration in the calculation of many standard model fit indices. Model fit indices should be used comparatively to evaluate which of two or more models has the best combination of fit and complexity. When comparing between two models, the model with the index closest to zero is deemed to be the best fitting model and provides the least out-of-sample prediction error. The estimation technique will determine which indices should be used.

When using lme4 to estimate a model with ML, the log-likelihood, deviance, Akaike Information Criterion (AIC; Akaike, 1987), and Bayesian Information Criterion (BIC; Schwarz, 1978) are readily produced. The log-likelihood and deviance are measures of fit but do not account for complexity. Deviance is -2 times the log-likelihood and AIC and BIC are adjustments to the deviance. AIC and BIC are penalized deviance measures, adding to the deviance based on the number of predictors in the model. In this way, more parsimonious models are "rewarded" with smaller penalizations. To see this, the formulas for AIC and BIC are given:

$$AIC = d + 2p \tag{6}$$

$$BIC = d + p\ln(n), \tag{7}$$

where *d* is the deviance, *p* is the number of predictors in the model, and *n* is the sample size. Smaller values of deviance, AIC, and BIC indicate overall better model fit and lower out-of-sample predictive error. Because the sample size in HLM will differ at different levels, Hox (2010) recommends the use of AIC for its straightforward calculation.

Model fit indices when using REML must be considered carefully. Models fit by REML can only be compared if they have identical fixed components, for reasons described earlier. Using lme4, a REML convergence criterion is produced instead of the deviance previously mentioned with ML. Evaluation with the REML convergence criterion is the same, with a value closer to zero indicating better model fit. Although lme4 does not immediately produce the AIC and BIC for models fit using REML, these values can be called using functions found in Appendix 1. However, if the models being compared differ in their fixed effects, then using these measures to assess model fit does not make sense.

Bayesian model fit indices include the deviance and the deviance information criterion (DIC; Spiegelhalter et al., 2002). The DIC functions similarly to AIC by penalizing the deviance for complexity (Gelman & Hill, 2007). The use of DIC to evaluate model fit is the same as other indices; a smaller DIC indicates a superior model in terms of fit and complexity.

### Comparison

Presented in Table 6 are fit indices for two models. The two models being compared are a varying-intercepts only model (see equations 1) and a varying-intercepts and varying-slopes model (see equations 5). Note that the two models only differ in their random components, thereby making comparisons using REML appropriate.

The deviance and REML criterion were lower for the more complex model across estimation techniques, indicating that the more complex model was a better fit. However, the AIC and BIC for both ML and REML and the DIC for Bayesian estimation had lower values for the simpler model. While including more parameters in the varying-intercepts and varying-slopes model improved fit, the increased complexity of the model made it less attractive in terms of out-of-sample prediction. The more parsimonious model was rewarded with lower values of AIC, BIC, and DIC.

The point needs to be made that although one model yields a better set of fit indices than another, it may not be the best model. Instead, when one model is deemed superior to another model, the superior model should be considered as a member of several possible models still to be compared. This requires thoughtful consideration by the researcher and a willingness to test all reasonable models.

**Table 6**. Mode Fit Indices Using MIL, REML, and Bayesian Estimation

| | Estimation Technique | | | | | |
|---|---|---|---|---|---|---|
| | ML | | REML | | Bayesian | |
| Fit Index | Varying Intercept | Varying Int/Slope | Varying Intercept | Varying Int/Slope | Varying Intercept | Varying Int/Slope |
| Log-Likelihood | -109.1 | -109 | | | | |
| Deviance | 218.2 | 218.1 | | | 198.26 | 196.81 |
| REML Criterion | | | 219.6 | 219.2 | | |
| AIC | 226.2 | 230.1 | 227.6 | 231.2 | | |
| BIC | 236.7 | 245.8 | 238.0 | 246.9 | | |
| DIC | | | | | 205.9 | 207.7 |

Note. Bayesian posterior mean values are shown only. Posterior mode values were similar, yielding the same interpretation of fit.

## Conclusion

Three methods of estimation have been introduced and discussed in the context of HLM. The estimated values using ML or REML are those that were most likely to produce the data. REML restricts the types of models that can be compared to those which differ only in random components. Estimates of residual variances when using REML are less biased compared to ML, particularly when the number of groups is small. With fully Bayesian estimation, researchers use probability distributions in a hierarchical scheme of priors and likelihood to determine posterior distributions. From the posterior distributions, parameter estimates and intervals may be derived. The posterior mode should be used as the parameter estimate, particularly when the number of groups is small, and the 95% HDI can be interpreted to have a 95% chance of containing the true value. The choice of which technique to use will depend on the statistical framework the researcher is willing to work within and the number of groups in the dataset. Considering its importance, which estimation technique to use is a decision best made by the researcher and not to be left to the default settings of statistical software.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium on Information Theory, ed. B. N. Petrov and F. Csaki, 267-281. Budapest: Akademiai Kiado. Reprinted in Breakthroughs in Statistics, ed. S. Kotz, 610-624. New York: Springer-Verlag, 1992.

Banjanovic, E. S., & Osborne, J. W. (2016). Confidence intervals for effect sizes: Applying bootstrap resampling. *Practical Assessment, Research & Evaluation*, *21*(5). Available online: http://pareonline.net/getvn.asp?v=21&n=5.

Bates, D. (2006, May 19). [R] lmer, p-values and all that [Blog post]. Retrieved from https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., …, & Green, P. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 3*, 473-514.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis*. Boca Raton, FL: CRC Press.

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics, 18*(1), 31-38.

Cumming, G., & Finch, S. (2004). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.

Cumming, G., Williams, J., & Fidler, F. (2010). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics, 48*(3), 432-435. doi: 10.1198/004017005000000661

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gilks, W. R., Thomas A, Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician, 43*, 169-178.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence interval, and power: A guide to misinterpretations. *European Journal of Epidemology, 31*, 337-350. doi: 10.1007/s10654-016-0149-3

Hayes, A. (2006). A primer on multilevel modeling. *Human Communication Research, 32*(4), 385-410. doi: 10.1111/j.1468-2958.2006.00281.x

Hox, J. J. (2010). *Multilevel analysis: techniques and applications*. New York, NY: Routledge.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573-603. doi: 10.1037/a0029146

Kruschke, J. K. (2015). *Doing bayesian data analysis* (2nd ed.). Cambridge, MA: Academic Press.

Louis, T. A. (2005). Introduction to Bayesian methods II: Fundamental concepts. *Clinical Trials, 2*, 291-294. doi: 10.1191/1740774505cn099oa

Ludecke, D. (2016). sjstats: Statistical function for regression models. Retrieved from https://CRAN.R-project.org/package=sjstats

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*(1), 90-100. doi: 10.1016/S0022-2496(02)00028-7

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. Practical Assessment, Research & Evaluation, 7(1). Retrieved from http://pareonline.net/getvn.asp?v=7&n=1

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*(1), 85-112. doi: 10.1016/j.jsp.2009.09.002

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), Proceedings of the 3rd international workshop on distributed statistical computing (pp. 1-10).

R Core Team. (2016). R: A language and environment for statistical computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464. doi:10.1214/aos/1176344136

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. Hoboken, NJ: John Wiley.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64*(4), 583-639. doi: 10.1111/1467-9868.00353

Stan Development Team. (2016). Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0. http://mc-stan.org

Su, Y., & Yajima, M. (2015). R2jags: Using R to run 'JAGS'. Retrieved from https://CRAN.R-project.org/package=R2jags

Yu, Chong Ho (2003). Resampling methods: Concepts, applications, and justification. *Practical Assessment, Research & Evaluation, 8*(19). Retrieved from http://PAREonline.net/getvn.asp?v=8&n=19.

## Appendix 1

Import and set up the data

```
install.packages("foreign")
library(foreign)

popdata <- read.dta("http://www.ats.ucla.edu/stat/stata/examples/mlm_ma_hox/popular.dta")

#This limits the dataset to the 101 students in the first 5 schools
popdata <- popdata[1:101,]

J <- length(unique(popdata$school))
school <- as.numeric(popdata$school)
sex <- ifelse(popdata$sex=="girl",0,1)
y <- popdata$popular
n <- length(y)

#Install necessary packages for ML and REML

install.packages("lme4")
library(lme4)
install.packages("sjstats")
library(sjstats)

# Maximum Likelihood, Unconditional Model, Table 1
fit.ML.unconditional <- lmer(y ~ 1 + (1|school), REML = FALSE)

#To find both ICC and standard error estimates:
se(icc(fit.ML.unconditional))

#Maximum Likelihoood, Varying Intercept, Table 5
fit.ML.Int <- lmer(y ~ sex + (1|school), REML = FALSE)

#Maximum Likelihood, Varying Intercept, Varying Slope, Table 4 & 5
fit.ML.Int.Slope <- lmer(y ~ sex + (1 + sex|school), REML = FALSE)

#REML, Unconditional Model, Table 1 & 2
fit.REML.unconditional <- lmer(y ~ 1 + (1|school))

#To find both ICC and standard error estimates:
se(icc(fit.REML.unconditional))

#REML, Varying Intercept, Table 5
fit.REML.Int <- lmer(y ~ sex + (1|school))

#REML, Varying Intercept, Varying Slope, Table 4 & 5
fit.REML.Int.Slope <- lmer(y~ sex + (1 + sex|school))
#Bootstrap confidence intervals for varying intercept, varying slope REML model, Table 4
```

```
confint(fit.REML.Int.Slope, method = "boot")


#AIC and BIC, Table 5
AIC(fit.REML.Int.Slope)
BIC(fit.REML.Int.Slope)


# Bayesian Estimation using R2jags

#Necessary packages and functions

install.packages("R2jags")
library(R2jags)

#Unconditional model used for the ICC in Table 1 & 3

#Define the model.
cat("model {
   for(i in 1:n) {
   y[i]~dnorm(y.hat[i],tau.y)
   y.hat[i]<-a[school[i]]
   }
   tau.y<-pow(sigma.y,-2)
   sigma.y~dunif(0,100)
   for(j in 1:J){
   a[j]~dnorm(a.hat[j],tau.a)
   a.hat[j]<-mu.a
   }
   mu.a~dnorm(0,0.0001)
   tau.a<-pow(sigma.a, -2)
   sigma.a~dunif(0,100)

   #This bit is included to find the ICC
   sigma2.a<-1/tau.a
   sigma2.y<-1/tau.y
   icc<-sigma2.a/(sigma2.y+sigma2.a)
   }", file="Uncon.txt")


#Tell R the data, the parameters to be monitored, and initial values for the chains
unconDat <- list("n", "J", "y", "school")
unconParams<-c("a", "sigma.y", "mu.a",
            "sigma.a", "icc")
unconInits <- function() list(a=rnorm(J),
                     sigma.y=runif(1,0,1), mu.a=rnorm(1,0,1),
                     sigma.a=runif(1,0,1))

unconResults=jags(data=unconDat, inits=unconInits, parameters.to.save=unconParams,
                n.iter=21000, n.burnin=1000, n.thin=1, model.file="Uncon.txt")

#Use "traceplot" to check convergence of the chains. All three chains are plotted on top of one another.
traceplot(unconResults)

#To find the posterior mode of the ICC
# mode
estimate_mode <- function(x) {
 d <- density(x)
 d$x[which.max(d$y)]
}

icc.sims <- unconResults$BUGSoutput$sims.matrix[,7]

estimate_mode(icc.sims)


#Varying-intercepts, varying-slopes model, Table 4

cat("model{
    for(i in 1:n) {
    y[i]~dnorm(y.hat[i],tau.y)
    y.hat[i] <- a[school[i]]+
    b[school[i]]*sex[i]
    }
    tau.y <- pow(sigma.y, -2)
    sigma.y~dunif(0,100)
    for(j in 1:J){
    a[j] <- B[j,1]
    b[j] <- B[j,2]
    B[j,1:2] ~ dmnorm(B.hat[j,],
    Tau.B[,])
    B.hat[j,1] <- mu.a
    B.hat[j,2] <- mu.b
    }
    mu.a~dnorm(0,0.0001)
    mu.b~dnorm(0,0.0001)
    Tau.B[1:2,1:2] <- inverse(Sigma.B[,])
    Sigma.B[1,1] <- pow(sigma.a,2)
    sigma.a ~ dunif(0,100)
    Sigma.B[2,2] <- pow(sigma.b,2)
    sigma.b ~ dunif(0,100)
    #corrlation
    Sigma.B[1,2] <- rho*sigma.a*sigma.b
    Sigma.B[2,1] <- Sigma.B[1,2]
    rho ~ dunif(-1,1)
    }", file="vivc.txt")

#Tell R the data, the parameters to be monitored, and initial values for the chains
```

```
vivcDat <- list("n", "J", "y", "school", "sex")
vivcParams <- c("a", "b", "sigma.y", "mu.a",
            "sigma.a", "mu.b", "sigma.b", "rho")
vivcInits <- function() {list(B=array(rnorm(J*2), c(J,2)),
                      sigma.y=runif(1,0,1), mu.a=rnorm(1,0,1),
                      sigma.a=runif(1,0,1), mu.b=rnorm(1,0,1),
                      sigma.b=runif(1,0,1), rho=runif(1,-1,1))}
vivcResults <- jags(data=vivcDat, inits=vivcInits, parameters.to.save=vivcParams,
               n.iter=21000, n.burnin=1000, n.thin=1, model.file="vivc.txt")

#Use "traceplot" to check convergence of the chains. All three chains are plotted on top of one another.
traceplot(vivcResults)

#To find the posterior mode of
# mode
estimate_mode <- function(x) {
  d <- density(x)
  d$x[which.max(d$y)]
}

#mode of posterior for the fixed component of the intercept
mu.a.sims <- vivcResults$BUGSoutput$sims.matrix[,12]
estimate_mode(mu.a.sims)
#mode of the posterior for the fixed component of the coefficient of Sex
mu.b.sims <- vivcResults$BUGSoutput$sims.matrix[,13]
estimate_mode(mu.b.sims)
#mode of posterior for the random component of the intercept
sigma.a.sims <- vivcResults$BUGSoutput$sims.matrix[,15]
estimate_mode(sigma.a.sims)
#mode of the posterior for the random component of the coefficient of Sex
sigma.b.sims <- vivcResults$BUGSoutput$sims.matrix[,16]
estimate_mode(sigma.b.sims)
#mode of the posterior for the residual
sigma.y.sims <- vivcResults$BUGSoutput$sims.matrix[,17]
estimate_mode(sigma.y.sims)
#mode of the posterior for the correlation
rho.sims <- vivcResults$BUGSoutput$sims.matrix[,14]
estimate_mode(rho.sims)
#mode of the posterior for the deviance (used in table 5)
deviance.sims.1 <- vivcResults$BUGSoutput$sims.matrix[,11]
estimate_mode(deviance.sims.1)

#Varying-intercept model, used in Table 5

cat ("model{
   for (i in 1:n){
   y[i] ~ dnorm(y.hat[i],tau.y)
   y.hat[i] <- a[school[i]] + b*sex[i]
```

```
    }
    b ~ dnorm(0,0.0001)
    tau.y <- pow(sigma.y, -2)
    sigma.y ~ dunif(0,100)

    for (j in 1:J) {
    a[j] ~dnorm(mu.a, tau.a)
    }
    mu.a ~ dnorm(0,0.0001)
    tau.a <- pow(sigma.a, -2)
    sigma.a ~ dunif(0,100)
    }", file="vi.txt")

#Tell R the data, the parameters to be monitored, and initial values for the chains
viDat <- list("n", "J", "y", "school", "sex")
viParams<-c("a", "b", "sigma.y", "mu.a", "sigma.a")
viInits <- function() {list("b"=rnorm(1,0,1), "mu.a"=rnorm(1,0,1),
                        "sigma.y"=runif(1,0,1), "sigma.a"=runif(1,0,1))}

#Run with a burn-in period included
viResults <- jags(data=viDat, inits=viInits, parameters.to.save=viParams,
                n.iter=21000, n.burnin=1000, n.thin=1, model.file="vi.txt")

#Use "traceplot" check convergence of the chains. All three chains are plotted on top of one another.
traceplot(viResults)

#To find the posterior mode of
# mode
estimate_mode <- function(x) {
  d <- density(x)
  d$x[which.max(d$y)]
}
```

## Appendix 2

**Example output in R for the model fit by maximum likelihood with varying-intercepts and varying-slopes**

```
summary(fit.ML.Int.Slope)

Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: y ~ sex + (1 + sex | school)

     AIC      BIC   logLik deviance df.resid
   230.1    245.8   -109.0    218.1       95

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2209 -0.6419 -0.0393  0.5583  3.1720

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 school   (Intercept) 1.64623  1.2831
          sex         0.01516  0.1231   0.14
```

Model Fit Indices in Table 6

Random Components in Table 5

Correlation in Table 5

```
 Residual             0.40408  0.6357
Number of obs: 101, groups:  school, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)   6.1719     0.5807  10.628
sex          -0.5710     0.1399  -4.082

Correlation of Fixed Effects:
    (Intr)
sex -0.043
```

Fixed Components in Table 5

**Example output in R for the model fit by REML with varying-intercepts and varying-slopes.**

```
summary(fit.REML.Int.Slope)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ sex + (1 + sex | school)

REML criterion at convergence: 219.2
```

Model Fit Indices in Table 6

```
Scaled residuals:
     Min      1Q      Median     3Q       Max
-2.14867 -0.57558 -0.08174  0.62449  3.13306

Random effects:
 Groups   Name          Variance    Std.Dev. Corr
 school   (Intercept)   2.06470     1.4369
          sex           0.03977     0.1994    0.05
 Residual               0.40410     0.6357
Number of obs: 101, groups:  school, 5
```

Random Components in Table 5

Correlation in Table 5

```
Fixed effects:
                  Estimate    Std. Error  t value
(Intercept)       6.1700      0.6488      9.510
sex               -0.5750     0.1566      -3.672

Correlation of Fixed Effects:
    (Intr)
sex -0.050
```

Fixed Components in Table 5

**Example output in R using Bayesian estimation for the varying-intercepts, varying-slopes model. The a[*i*] is the estimated intercept for group *i* and the b[*i*] is the estimated slope for group *i*. Mu.a and mu.b are the fixed components for the intercepts and slopes, respectively. Sigma.a and sigma.b are the standard deviations of the random components for intercepts and slopes, respectively. Sigma.y is the standard deviation of the first-level residual.**

**Parameter estimates for varying-intercepts, varying-slopes model**

Inference for Bugs model at "vivc.txt", fit using jags,
 3 chains, each with 21000 iterations (first 1000 discarded)
 n.sims = 60000 iterations saved

| | mu.vect | sd.vect | 2.50% | 25% | 50% | 75% | 97.50% | Rhat | n.eff |
|---|---|---|---|---|---|---|---|---|---|
| a[1] | 7.804 | 0.184 | 7.438 | 7.681 | 7.805 | 7.927 | 8.164 | 1.001 | 22000 |
| a[2] | 4.37 | 0.185 | 4 | 4.247 | 4.371 | 4.494 | 4.727 | 1.001 | 12000 |
| a[3] | 7.07 | 0.175 | 6.731 | 6.952 | 7.067 | 7.185 | 7.42 | 1.001 | 9200 |
| a[4] | 6.584 | 0.199 | 6.177 | 6.453 | 6.59 | 6.72 | 6.957 | 1.001 | 8500 |
| a[5] | 5.018 | 0.187 | 4.658 | 4.891 | 5.016 | 5.14 | 5.394 | 1.001 | 15000 |
| b[1] | -0.532 | 0.225 | -0.974 | -0.678 | -0.536 | -0.391 | -0.072 | 1.001 | 12000 |
| b[2] | -0.513 | 0.23 | -0.956 | -0.664 | -0.521 | -0.372 | -0.032 | 1.001 | 6300 |
| b[3] | -0.732 | 0.263 | -1.318 | -0.889 | -0.7 | -0.548 | -0.29 | 1.001 | 5000 |
| b[4] | -0.391 | 0.241 | -0.805 | -0.561 | -0.415 | -0.24 | 0.132 | 1.001 | 9600 |
| b[5] | -0.717 | 0.24 | -1.24 | -0.866 | -0.698 | -0.549 | -0.3 | 1.001 | 5100 |
| mu.a | 6.167 | 1.383 | 3.499 | 5.537 | 6.168 | 6.8 | 8.837 | 1.005 | 16000 |
| mu.b | -0.576 | 0.276 | -1.128 | -0.709 | -0.572 | -0.442 | -0.056 | 1.001 | 15000 |
| rho | 0.001 | 0.537 | -0.916 | -0.443 | -0.005 | 0.449 | 0.92 | 1.001 | 60000 |
| sigma.a | 2.536 | 1.873 | 0.986 | 1.523 | 2.035 | 2.896 | 7.077 | 1.003 | 1800 |
| sigma.b | 0.390 | 0.379 | 0.018 | 0.146 | 0.29 | 0.507 | 1.404 | 1.002 | 1800 |
| sigma.y | 0.646 | 0.048 | 0.56 | 0.612 | 0.643 | 0.676 | 0.747 | 1.001 | 31000 |
| deviance | 196.73 | 4.625 | 189.2 | 193.48 | 196.181 | 199.394 | 207.339 | 1.001 | 60000 |

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 10.7 and DIC = 207.4
DIC is an estimate of expected predictive error (lower deviance is better).

Model Fit Indices in Table 6
Random Components in Table 5
Correlation in Table 5
Fixed Components in Table 5

## Citation:

Boedeker, Peter (2017). Hierarchical Linear Modeling with Maximum Likelihood, Restricted Maximum Likelihood, and Fully Bayesian Estimation. *Practical Assessment, Research & Evaluation*, 22(2). Available online: http://pareonline.net/getvn.asp?v=22&n=2

## Corresponding Author

Peter Boedeker
University of North Texas

email: peter.boedeker@unt.edu