Psicothema

Psicothema 2013, Vol. 25, No. 4, 520-528
doi: 10.7334/psicothema2013.58

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2013 Psicothema
www.psicothema.com

Multilevel bootstrap analysis with assumptions violated

Guillermo Vallejo Seco[1], Manuel Ato García[2], María Paula Fernández García[1] and Pablo Esteban Livacic Rojas[3]
[1] Universidad de Oviedo, [2] Universidad de Murcia and [3] Universidad Santiago de Chile

Abstract

Background: Likelihood-based methods can work poorly when the residuals are not normally distributed and the variances across clusters are heterogeneous. **Method:** The performance of two estimation methods, the non-parametric residual bootstrap (RB) and the restricted maximum likelihood (REML) for fitting multilevel models are compared through simulation studies in terms of bias, coverage, and precision. **Results:** We find that (a) both methods produce unbiased estimates of the fixed parameters, but biased estimates of the random parameters, although the REML was more prone to give biased estimates for the variance components; (b) the RB method yields substantial reductions in the difference between nominal and actual confidence interval coverage, compared with the REML method; and (c) for the square root of the mean squared error (RMSE) of the fixed effects, the RB method performed slightly better than the REML method. For the variance components, however, the RB method did not offer a systematic improvement over the REML method in terms of RMSE. **Conclusions:** It can be stated that the RB method is, in general, superior to the REML method with violated assumptions.

Keywords: Multilevel model, heterogeneous variances, nonparametric bootstrap, maximum likelihood.

Resumen

Análisis bootstrap multinivel con supuestos incumplidos. **Antecedentes:** los métodos basados en la verosimilitud pueden trabajar con dificultad cuando los errores no se distribuyen normalmente y las varianzas a través de los grupos son heterogéneas. **Método:** el desempeño de dos métodos de estimación, el bootstrap residual (BR) no paramétrico y el de la máxima verosimilitud restringida (MVR), para ajustar modelos multinivel es comparado mediante estudios de simulación en términos de sesgo, cobertura y precisión. **Resultados:** encontramos que: (a) ambos métodos proporcionan estimaciones no sesgadas de los efectos fijos, pero sesgadas de los efectos aleatorios, aunque el método MVR es más propenso a generar estimaciones sesgadas para los componentes de la varianza; (b) el método BR depara diferencias más pequeñas entre las tasas de cobertura real y nominal de los intervalos de confianza que el método MVR; y (c) los valores de la raíz del error cuadrático medio (RECM) para los efectos fijos son algo más pequeños bajo el método BR que bajo el método REML. Sin embargo, en lo referido a los componentes de la varianza, el método de BR no ofrece una mejora sistemática sobre el método MVR en términos de RECM. **Conclusiones:** en general, se puede afirmar que el método BR resulta superior al método MVR con supuestos incumplidos.

Palabras clave: modelo multinivel, varianzas heterogéneas, bootstrap no paramétrico, máxima verosimilitud.

Multilevel data are prevalent in social and behavioral sciences research. Examples of naturally occurring hierarchies include observations nested within persons, participants nested within therapists, children nested within families, students nested within classrooms, and patients nested within health centers (see Apodaca, Magill, Longabaugh, Jackson, & Monti, 2013; Imel, Hubbard, Rutter, & Simon, 2013; Núñez, Rosário, Vallejo, & González-Pienda, 2013; Núñez, Vallejo, Rosário, Tuero-Herrero, & Valle, in press). The list of areas in which data can be organized into different levels or clusters is endless. Outcomes measured on the same person, therapist, family, classroom, or health center are almost certain to be correlated, and this needs to be taken into account in planning the analyses. In each of these cases, researchers can use multilevel models, special cases of mixed-effects regression models (Raudenbush & Bryk, 2002), because they incorporate the random effects into the model to accommodate the possible intra-cluster or intra-individual correlation. Details of the technique can be found in several texts, including Hox (2010), Raudenbush and Bryk (2002) and Snijders and Bosker (2012).

The dominant approaches to estimating the fixed effect and random effects model in multilevel analysis are based on the principle of maximum likelihood (ML) estimation. Other available tools (e.g., bootstrapping and Bayesian methods) are used less frequently. When distributional assumptions are made about the error forms at each level in the data, both ML and restricted ML (REML) estimation methods provide parameter estimates that are relatively straightforward. However, it is well known that these asymptotic methods can work poorly when the number of clusters is small and/or the residuals are not normally distributed. Numerical studies have shown that the fixed parameter estimates are unbiased, whereas their standard errors tend to be negatively biased as the number of clusters decreases. On the other hand, the variance components and their associated standard errors may be strongly biased downward, especially if the number of groups is too small (see, e.g., Van der Leeden, Meijer, & Busing, 2008).

Received: March 4, 2013 • Accepted: June 19, 2013

Corresponding author: Guillermo Vallejo Seco
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
e-mail: gvallejo@uniovi.es

Also implicit in the assumptions about error is the homogeneity of variance within clusters and across clusters. If the variances are heterogeneous, but vary randomly, it does not appear that the fixed effects or standard errors are biased (Kasim & Raudenbush, 1998), but if the variances depend in some way on the explanatory variables, it can severely affect the validity of inferences about random parameters and fixed effects (Dedrick et al., 2009; Raudenbush & Bryk, 2002).

To improve the accuracy of inferences, correct standard errors are required for data that violate distributional assumptions. As pointed out by Maas and Hox (2004), a well-known correction method for producing robust standard errors is the so-called Huber/White sandwich estimator, which is widely available in software for estimating multilevel models, including MLwiN and HLM. This approach, also available through the PROC MIXED option (see Sterba, 2009 for details), may not be the best choice for alleviating the bias when the sample size is not large enough (Diggle, Heagerty, Liang, & Zeger, 2002). Another commonly used method is to apply a non-linear transformation to the dependent variable. Unfortunately, it is not generally obvious which is the best choice of transformation and the interpretation of results on the selected scale may be unclear. Hodges (1998) discusses Box-Cox transformations for multilevel models. A somewhat different approach that may be viable for making inferences when the data fail to fulfill the assumptions of either normality or homogeneity it is based on deriving the empirical sampling distribution of the statistic of interest by randomly resampling with replacement from the sample available. Although the bootstrap methodology appears to be a viable alternative for improving the accuracy of inferences about parameter values (Carpenter, Goldstein, & Rasbash, 2003; Shieh & Fouladi, 2002), applications of bootstrapping are rare within the multilevel arena.

This paper investigates two issues. The first is to clarify the performance of the standard two-level analysis as implemented through the PROC MIXED (SAS Institute Inc., 2011) module in terms of bias, coverage, and precision when the normality and variance homogeneity assumptions are violated. As mentioned above, if the variances change as a function of some predictor, this will result in incorrect estimates of the sampling variability and it can lead to quite distorted statistical inferences. Nevertheless, little is currently known about the direction and severity of such effects. The second question is to check to what extent the residual bootstrap approach can correct bias in parameter estimates and improve the accuracy of inferences about parameter values. When the residuals are not normally distributed, bootstrapping has been presented (Carpenter, Goldstein, & Rasbash, 2003; Goldstein, 2011; Wang, Carpenter, & Kepler, 2006; Wang, Shi, & Fisher, 2011) as a potential strategy for dealing with the bias in the variance estimates and standard errors that results from using ML or REML estimation.

Notation and definition of the statistical procedures

Let Y_{ij} denote the ith observation ($i = 1,\dots, n_j$) in the jth group (j= 1, …, m) and $n = \Sigma^m_j n_j$ the total number of subjects enrolled in the study. The simplest multilevel model is a random intercept model:

$$Y_{ij} = \beta_{0j} + e_{ij} \tag{1}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{2}$$

where $u_{0j} \sim N(0, \tau_{00})$, $e_{ij} \sim N(0, \sigma^2_e)$, and $cov(e_{ij}, u_{0j}) = 0$. Substituting Equation 2 into Equation 1, we have the one-way random effects ANOVA model

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{3}$$

Covariate information can be introduced at both the individual and group level to create a more general multilevel model, which can be expressed in matrix notation as

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{e}_j \tag{4}$$

$$\boldsymbol{\beta}_j = \mathbf{Z}_j\boldsymbol{\gamma} + \boldsymbol{u}_j \tag{5}$$

where $\boldsymbol{y}_j$ is an n_j -vector of outcomes, $\boldsymbol{X}_j$ is an $n_j \times p$ matrix of explanatory variables at the individual level, $\boldsymbol{\beta}_j$ is a p-vector of individual-level random parameters, $\boldsymbol{e}_j$ is the error term on the individual level distributed normally with mean vector of $\mathbf{0}$ and covariance matrix Σ_j, which is often assumed to be $\boldsymbol{\sigma}^2\boldsymbol{I}$, $\boldsymbol{Z}_j$ is a $p \times q$ matrix of explanatory variables at the group level, $\boldsymbol{\gamma}$ is a q-vector of group-level fixed effects, and the error term on the group level $\boldsymbol{u}_j$ has dispersion matrix $\boldsymbol{\Gamma}$, which expresses the between group variability and covariance of the lowest level regression coefficients. Substituting Equation 5 into Equation 4 yields

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{\gamma} + \boldsymbol{Z}_j\boldsymbol{u}_j + \boldsymbol{e}_j \tag{6}$$

which is a special case of the linear mixed-effects model of Laird and Ware (1982) with $\boldsymbol{X}^*_j = \boldsymbol{X}_j\boldsymbol{Z}_j$.

Although several methods may be employed to estimate the parameters of multilevel models, including simple two-step ordinary and weighted least squares methods, likelihood based-methods, robust methods, and Bayesian methods (de Leeuw & Meijer, 2008; Raudenbush & Bryk, 2002; Goldstein, 2003), in this paper we will use a bootstrap resampling method to estimate the standard errors of the parameters. In addition to examining bootstrapping behavior, REML estimation, as implemented in PROC MIXED, was also conducted for the purpose of comparison.

Residual bootstrap (RB)

Several bootstrap strategies exist for analyzing hierarchically nested data (see van der Leeden, Meijer, & Busing, 2008, and the references therein). In general, these strategies may be divided into three basic categories: (a) parametric bootstrap (which generates new data by keeping the explanatory variables fixed and resampling with replacement the Level-1 and Level-2 residuals from a normal distribution); (b) residual bootstrap (which generates new data by keeping the explanatory variables fixed and resampling with replacement the Level-1 and Level-2 residuals from the observed raw residuals); and (c) cases bootstrap (which generates new data by resampling with replacement from the original sample r-dimensional observation vectors (i.e., $\boldsymbol{Z}$'s, $\boldsymbol{W}$'s and $\boldsymbol{Y}$'s). The estimates of the RB were the most accurate, making it the preferred estimation method of choice (Carpenter et al., 2003). For this reason we focused on the unweighted RB approach.

The version of the RB method used in our simulation is similar to that of Wang et al. (2006), except that the resampling procedure was carried out for every level of the treatment variable. Specifically, the following steps were used.

1. Obtain parameter estimates for the model in Equation 6 from the data by REML, and calculate residuals at each level by the method of shrinkage (i.e., using empirical Bayes estimator of $\boldsymbol{u}$ and generalized least squares estimator of $\boldsymbol{\gamma}$). Both the individual and group residuals are centered to avoid biased estimates caused by the nonzero mean of the residuals. Then, these centered residuals are rescaled to generate new residuals such that their covariance matrix is equal to the model-estimated residual covariance matrix (see formulas in Wang et al., 2006).
2. Draw random samples with replacement from the two sets of rescaled and centered Level-1 and Level-2 residuals, separately. Given the existence of heterogeneous dispersion matrices and the lack of equilibrium presented by the different groups making up the design, the resampling procedure is carried out for every level of the treatment variable.
3. Generate the bootstrap samples $\boldsymbol{y}_j^*$ from $\boldsymbol{y}_j^* = \boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{\gamma}+\boldsymbol{Z}_j\boldsymbol{u}_j^*+\boldsymbol{e}_j^*$, where the coefficient vector $\hat{\boldsymbol{\gamma}}$ is estimated from the multilevel model using the original data and the vectors $\hat{\boldsymbol{u}}_j^*$ and $\hat{\boldsymbol{e}}_j^*$ are the rescaled and centered error terms on the group and individual level, respectively.
4. Compute estimates for all parameters of the two-level model for each artificial (synthetic) data set.
5. Repeat steps 2-4 B times to obtain B sets of bootstrap parameter estimates for inference. As a general guideline, 1000 bootstrap samples are usually considered to be sufficient. The mean and standard deviation of the empirical distribution of the bootstrap estimates for a particular parameter are the bootstrap-estimated parameter and its standard error, respectively.

For more detailed information, see Carpenter et al. (2003), Wang et al. (2006), and Wang et al. (2011).

Method

The ex-post facto design that forms a basis for simulation study is taken from Núñez et al. (in press). This study focused on the relationship between contextual variables and students' academic achievement. To contribute to explaining the stated objective, the students' biology achievement is the outcome variable, predicted by a set of explanatory variables measured at the student level (Level-1) and at the class level (Level-2). Variables at Level-1 are learning approaches (X_1), study time (X_2), prior domain knowledge (X_3), homework completion (X_4), students' gender (X_5), class absence (X_6), and parents' educational level (X_7). In addition to the teaching approaches (Z_1) *per se*, other explanatory variables included in Level-2 were teachers' experience (Z_2), class size (Z_3), and teachers' gender (Z_4).

True data-generating model

In the data-generating process, only the first two explanatory variables at Level-1 and the first two explanatory variables at Level-2 were included. The model (represented as levels) used to simulate the data becomes, at Level-1:

$$Y_{ij} = b_{0j} + b_{1j}X_{1ij} + b_2X_{2ij} + e_{ij} \tag{7}$$

and at Level-2:

$$\begin{aligned} b_{0j} &= \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j} \\ b_{1j} &= \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + u_{1j} \end{aligned} \tag{8}$$

Consistent with common practice in multilevel modeling, we assume that the pupil-level residuals, e_{ij} have a normal distribution with mean zero and variance σ_e^2. We also assume that the group-level residuals, u_{0j} and u_{1j}, have a bivariate normal distribution with zero means, variances τ_{00} and τ_{11}, respectively, and covariance τ_{01}. The Level-1 regression coefficients with subscript j (i.e., b_{0j} and b_{1j}) are random coefficients, which varied across the classes, and were treated as dependent variables in the Level-2 equations; those without subscript j are fixed coefficients.

Study variables

Five variables are manipulated in order to examine the performance by type of method:

1. *Intraclass correlation* (ICC). The amount of variability attributable to clusters was set at values of .1 and .3. These conditions reflect the range of values that have been found in most multilevel studies (Mass & Hox, 2004).
2. *Number of groups* (NG). Since the multilevel analysis is affected by the sample size at the group level, the performance of the test statistics was investigated using two different sizes: *NG* = 50 and *NG* = 100. After half the time, the number of groups in the treatment condition was 20 and 40, and in the control condition the number was 30 and 60; the opposite occurred for the other half. For accurate estimates 100 or more groups would be advisable, however, 50 groups is a frequently occurring number in behavioral and educational research.
3. *Group size* (GS). The number of subjects per cluster was small and moderate. Specifically, *GS* = 15 and *GS* = 30. The size of the groups is based on the literature and on practice (Mass & Hox, 2004; Núñez et al., in press).
4. *Type of pairing* (TP). Previous studies have shown that unequal group sizes, when paired with unequal variances, can affect Type I error control for tests that compare measures independent across groups (see, e.g., Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011a). Therefore, positive and negative pairings of group sizes and variance were investigated. A positive pairing implies that the treatment condition having the smallest number of clusters is associated with the smallest variance, whereas the opposite occurs for a negative pairing. The unequal treatment conditions variances were in the ratio of 1:5.
5. *Distribution shape* (DS). To investigate the influence of non-normality of the error terms at all levels on the robustness of the procedures, we generated data from normal and non-normal distributions. Specifically, besides the multivariate normal distribution with univariate skewness (γ_1) and kurtosis (γ_2) equal to zero, the data were obtained from an asymmetric light–tailed distribution with shape parameters equivalent to those of an exponential distribution (i.e. $\gamma_1 = 2$; $\gamma_2 = 6$). This distribution and its associated values of skew and kurtosis is representative of those encountered in applied psychological research by Micceri (1989).

Under each scenario, 1000 data sets were generated using the power method proposed by Fleishman (1978), and the properties of the estimators were compared for bootstrap and REML procedures. In performing the residual bootstrap, B = 1000 bootstrap samples were taken. For all simulated conditions, the regression parameters (i.e., γ') were set to the value of one. The Level-1 variance component (i.e., σ_e^2) was also set to the value of one. The variance components of the intercept and slope (i.e., τ_{00} and τ_{11}) were assumed to be the same (i.e., .11, and .43 per input ICC .1 and .3), while the residual covariance between the intercept and the slope was constrained to zero. The fixed values for the observations on the *X*' and *Z*' were determined by drawing from a normal distribution with a mean of zero and a variance of one. Later, we dichotomized some variables by an arbitrary threshold (i.e., the mean of all observed data). Data manipulations were performed in SAS/IML and SAS MACRO languages.

Evaluation criteria

To determine the accuracy of the estimation methods being compared, we examine their performance in terms of bias, coverage, and precision.

1. *Bias*. For the generic parameter θ, bias is defined here as the difference between its average bootstrap estimate and its true value, and is given by $(\bar{\hat{\theta}}^{B} - \theta)$, where $\bar{\hat{\theta}}^{B} = \sum_{i=1}^{S} \hat{\theta}_i^{B} /$ S and $\hat{\theta}_i^{B}$ is the bootstrap estimate of interest within each of the $i = 1,..., S$ simulations. The smaller the bias, the closer the estimate is to the true value, and so the more accurate the estimate.
2. *Coverage*. The coverage is the percentage of times that true parameter value is covered in the confidence to the nominal rate. Percentile-based confidence intervals are given by $\hat{\theta}_i^{B} \pm Z_{1\text{-}\alpha/2}\ SE(\hat{\theta}_i^{B})$, where $SE(\hat{\theta}_i^{B})$ denotes the standard error (i.e., the standard deviation of the bootstrap replicates) of the estimate of interest within each simulation and $Z_{1\text{-}\alpha/2}$ the 1-α/2 quantile of the standard normal distribution. For the 95% confidence level employed here, the interval used for defining the robustness of the tests was 92.5 and 97.5. If a procedure is working well, the actual coverage should be close to the nominal (i.e., Type I error rates are properly controlled). For each parameter a non-coverage indicator variable was set up that was equal to zero if its true value was in the confidence interval, and equal to one if its true value was outside the confidence interval.
3. *Precision*. Bias and variance are often combined into a single measure of overall variability of the model called mean squared error (MSE), which is computed as: $(\theta^{B} - \theta)^2 + V(\theta^{B})$, where $V(\theta^{B}) = \sum_{i=1}^{S}\left(\hat{\theta}^{B} - \bar{\hat{\theta}}^{B}\right)^2 / S$. It is common to report the square root of the MSE (RMSE), which is on the same scale as the parameter and makes it easier to interpret (Burton, Altman, Royston, & Holder 2006). The smaller RMSE values indicate more accurate estimations.

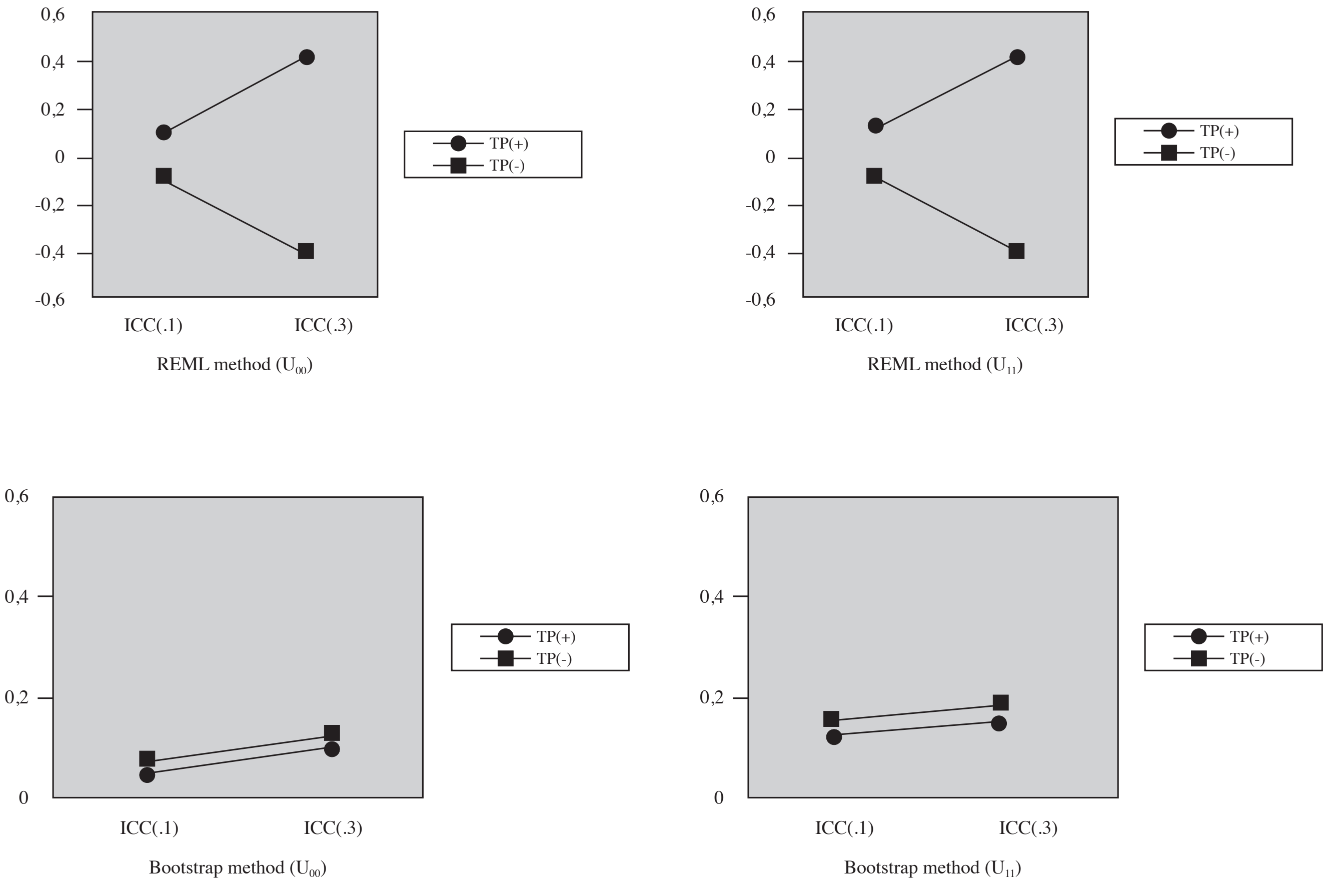

Figure 1. *Interaction plots of intraclass correlation (ICC) and type of pairing (TP) at REML and bootstrap for bias of the second level variances (i.e., u_{00} and u_{11})*

Results

Assessment of bias

Bias in fixed effects estimates. Bias occurs when there is a systematic difference between the estimated mean and the true population mean. The REML method gives results similar to the RB method. Although the bias was slightly larger when the explanatory variables were qualitative (i.e., X_1 and Z_1) rather than quantitative (i.e., X_2 and Z_2), the bias of the estimates was quite small for the two methods. In fact, the largest bias (less than 6%) was observed when the data were skewed in the condition with the smallest sample sizes in combination with the highest ICC. In order to preserve space, the results are not tabled but are available from the authors upon request.

Bias in random effects estimates. For the REML method, the bias ranged in magnitude from -.442 to .438 for u_{00} (M = -.00, SD = .30), -.428 to .438 for u_{11} (M = .00, SD = .30), and .059 to .119 for e_{00} (M = .08, SD = .02). In contrast to the REML method, the bias for the RB method ranged from .020 to .141 for u_{00} (M = .08, SD = .13), .049 to .284 for u_{11} (M = .16, SD = .23), and -.133 to .045 for e_{00} (M = .06, SD = .05). The effect of bias was markedly different for the two methods of estimating the variances, and was primarily affected by the TP and by the ICC. As seen from the Figure 1, both methods provide biased estimates of the second level variances (i.e., u_{00} and u_{11}). For the REML method, these variances were moderately overestimated (underestimated) when the TP was positive (negative) and the ICC value was high, but were slightly overestimated (underestimated) when the TP was positive (negative) and the ICC value was low. In turn, the RB variances were always slightly overestimated.

Assessment of coverage

Coverage rates for the fixed effects. The accuracy of the confidence intervals of the REML and bootstrap methods was evaluated in terms of the coverage rates. For γ_{00}, γ_{20} (REML and RB based), γ_{02}, γ_{12} (REML based), and γ_{11} (RB based) the coverage of the 95% intervals turns out to be nearly nominal significance level. The coverage of the remaining parameter estimates was outside the sampling variation of the true parameter values. In particular, the coverage for the REML method ranged from 96.5 to 97.9 (M = 97.1, SD = .48) for γ_{10}, 88.4 to 97.9 (M = 93.1, SD = 4.20) for γ_{01}, and 88.9 to 97.8 (M = 93.5, SD = 4.06) for γ_{11}. In turn, the coverage for the RB method ranged from 98.1 to 99.2 (M = 98.6, SD = .36) for γ_{10}, 96.4 to 98.7 (M = 97.4, SD = .62) for γ_{02}, and 97.7 to 98.8 (M = 98.2, SD = .35) for γ_{12}. For positive (negative) pairing, the REML method tends to overestimate (underestimate) the standard errors of γ_{01} and γ_{11}. However, RB standard deviations for the same fixed effects were overestimated. The coverage of fixed effects under REML was primarily affected by the TP and the NG, whereas the coverage obtained by the RB method was affected by the ICC and the GS. This is shown in Figure 2.

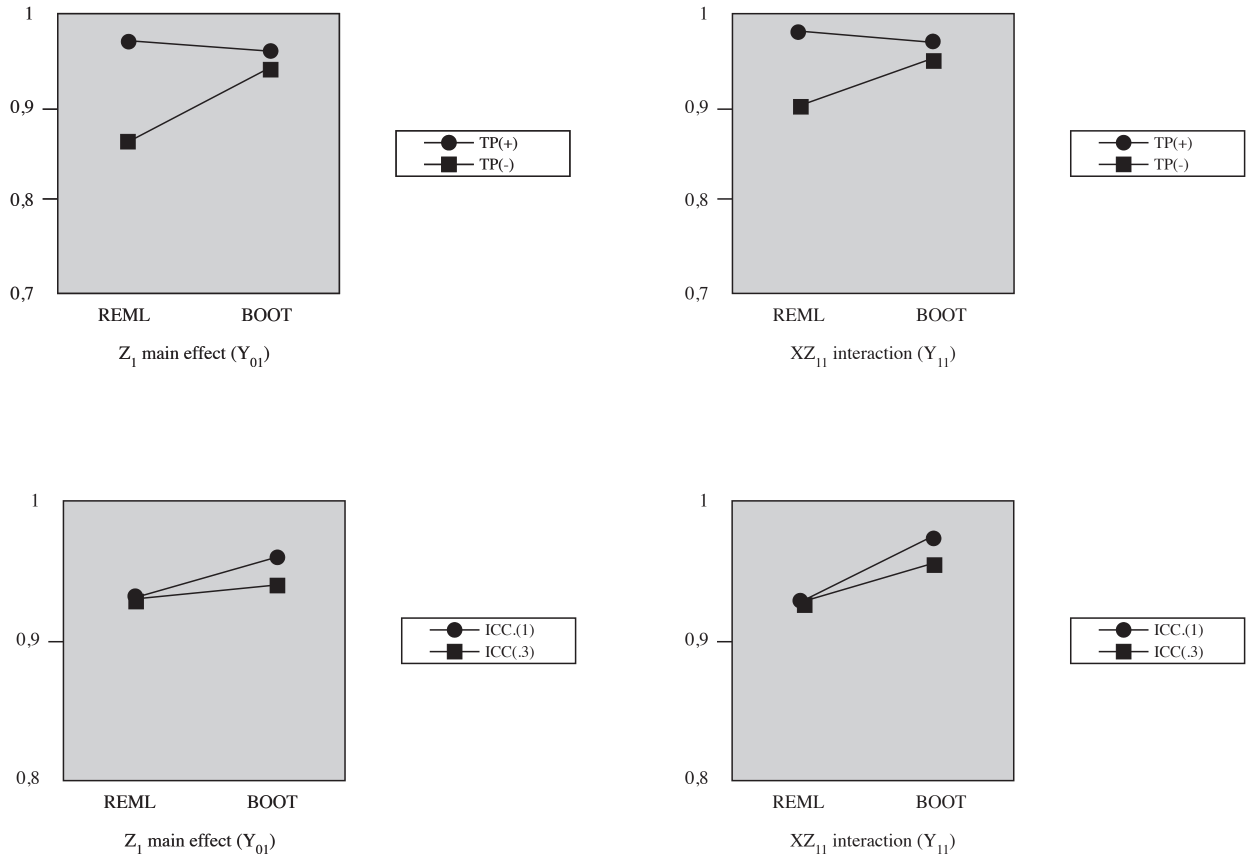

Figure 2. *Interaction plots of type of pairing (TP) and method (i.e., REML and bootstrap), and intraclass correlation (ICC) and method for coverage rates of Z_1 (i.e. γ_{01}) principal effect and the XZ_{11} (i.e. γ_{11}) secondary effect*

Coverage rates for the random effects. As is shown in Table 1, the coverage of the variances under REML ranged from 46.5 to 91.9 (M = 71.5, SD = 13.18) for u_{00}, 53.7 to 91.7 (M = 76.4, SD = 10.68) for u_{11}, and 58.3 to 91.3 (M = 74.4, SD = 14.35) for e_{00}. The coverage of the RB ranged from 78.1 to 99.8 (M = 93.4, SD = 6.96) for u_{00}, 90.4 to 99.9 (M = 98.1, SD = 2.58) for u_{11}, and 74.7 to 94.3 (M = 85.7, SD = 7.72) for e_{00}.

For the group level variances under REML, the coverage was primarily affected by the DS and by the TP, whereas the coverage obtained by the RB was primarily affected by the GS. For the lowest level variance, the coverage of both methods was primarily affected by the DS and by the TP. As expected, however, the difference in the coverage rates between normal and exponential distributions was larger under REML than under the RB.

Assessment of precision

RMSE for the fixed effects. The level of precision associated with the REML and RB estimators was evaluated using an integrated measure of bias and variance. The estimators in RMSE vary widely in respect to the measurement scale used for the explanatory variables (i.e., metric or nonmetric). From Figure 3, it can be seen that there are marked differences between the results obtained by the two methods when the Level-1 regression coefficients are treated as dependent variables in the Level-2 equations (see Eq. 8), but they give quite similar results for the remaining fixed effects.

RMSE for the random effects. The RMSE for the REML method ranged from .19 to 1.50 (M = .62, SD = .39) for u_{00}, .23 to 1.53 (M = .74, SD = .41) for u_{11}, and .18 to .83 (M = .42, SD = .18) for e_{00}.

Table 1
Mean coverage rates (%) of the parameter estimates for the variance components under the REML and bootstrap methods

	Normal				**Exponential**			
	50/15	**50/30**	**100/15**	**100/30**	**50/15**	**50/30**	**100/15**	**100/30**
REML: ICC =.1 & TP(+)								
u_{00}	**90.80**	**90.74**	**89.18**	**88.20**	**79.14**	**76.46**	**76.52**	**72.10**
u_{11}	**90.50**	**90.46**	**89.54**	**89.54**	**82.08**	**82.82**	**84.10**	**81.72**
e_{00}	**91.29**	**90.87**	**90.48**	**89.66**	**62.54**	**60.88**	**64.18**	**60.28**
REML: ICC =.3 & TP(+)								
u_{00}	**91.94**	**90.88**	**87.92**	**85.54**	**71.12**	**69.32**	**68.16**	**65.36**
u_{11}	**91.74**	**90.70**	**88.86**	**87.60**	**79.16**	**74.22**	**76.34**	**71.30**
e_{00}	**90.42**	**90.42**	**91.08**	**89.32**	**64.34**	**61.30**	**64.00**	**61.10**
REML: ICC =.1 & TP(–)								
u_{00}	**75.46**	**72.16**	**72.24**	**69.18**	**64.20**	**56.66**	**61.84**	**53.38**
u_{11}	**77.28**	**76.40**	**79.00**	**73.56**	**72.68**	**66.74**	**71.52**	**65.40**
e_{00}	**88.34**	**89.10**	**89.11**	**86.38**	**58.90**	**57.38**	**58.88**	**58.50**
REML: ICC =.3 & TP(–)								
u_{00}	**69.75**	**68.35**	**65.45**	**64.23**	**54.72**	**50.89**	**49.57**	**46.48**
u_{11}	**72.76**	**70.39**	**70.76**	**67.36**	**61.73**	**56.48**	**58.67**	**53.65**
e_{00}	**86.83**	**84.25**	**84.55**	**82.07**	**59.60**	**59.07**	**59.59**	**58.33**
Bootstrap: ICC =.1 & TP(+)								
u_{00}	**98.64**	**99.76**	**99.04**	**99.02**	**97.92**	94.96	**99.21**	95.32
u_{11}	**99.64**	**99.97**	**99.88**	**99.89**	**99.75**	**99.24**	**99.90**	**99.77**
e_{00}	93.65	**90.78**	94.25	**90.74**	**83.16**	**83.03**	**83.12**	**82.64**
Bootstrap: ICC =.3 & TP(+)								
u_{00}	97.04	95.44	**97.88**	96.55	**86.96**	**81.44**	**86.39**	**79.56**
u_{11}	**99.12**	**97.52**	**99.71**	**98.76**	97.44	**91.72**	**98.52**	94.44
e_{00}	93.24	94.08	92.79	92.99	**83.56**	**86.04**	**85.27**	**82.84**
Bootstrap: ICC =.1 & TP(–)								
u_{00}	**99.45**	97.01	**99.78**	**98.29**	**97.52**	92.95	**99.03**	92.60
u_{11}	**99.47**	**99.23**	**99.90**	**99.57**	**99.60**	**99.73**	**99.92**	**99.81**
e_{00}	**84.72**	**89.55**	**82.41**	**89.68**	**76.35**	**79.42**	**74.69**	**79.81**
Bootstrap: ICC =.3 & TP(–)								
u_{00}	95.76	93.76	**97.84**	96.24	**84.56**	**78.36**	**83.49**	**78.06**
u_{11}	**98.33**	96.00	**99.17**	97.00	96.44	**91.99**	97.05	**90.45**
e_{00}	**90.24**	**89.81**	**87.67**	**87.89**	**80.02**	**76.92**	**79.84**	**79.36**

Note: Bold values correspond to rates outside the interval 92.50 – 97.50, for the 5% level of significance

The RMSE for the RB method ranged from .20 to 1.53 (*M* = .59, *SD* = .38) for u_{00}, .24 to 1.80 (*M* = .80, *SD* = .40) for u_{11}, and .17 to .74 (*M* = .39, *SD* = .16) for e_{00}. From Figure 4, it can be seen that the REML method has largest (smallest) RMSE when the TP was positive (negative). The reverse was true for the RB method. In Figure 4, it can also be seen that the ICC × method interaction displayed a similar pattern, but the difference was smaller. This suggests that the RB method does not offer an systematic improvement over the REML method in terms of RMSE.

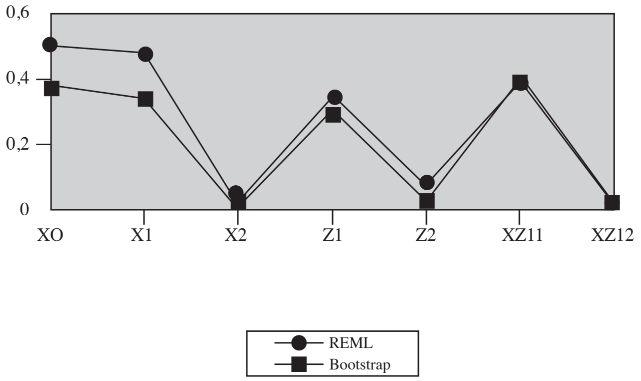

Figure 3. *Plot of the mean RMSE rates for the fixed effects under the REML and bootstrap methods*

On the other hand, for e_{00}, the difference in the RMSE between normal and exponential distributions was larger under REML than under the RB.

Discussion

In this study, we have examined the performance of RB and REML methods of fitting multilevel models when the normality and variance homogeneity assumptions were violated. Until now, the performance of the RB method had been restricted to examination of robustness under departure from normality (Carpenter et al., 2003). Our main findings are summed up in the following points.

Firstly, with respect to the bias, the estimated (REML and RB) fixed parameter values were generally close to the population values in the presence of heterogeneous and non-normal data. In fact, in 88% of the examined conditions the bias was less than 2% and in the remaining conditions it never exceeded 6%. However, both methods provide biased estimates of the second level variances (i.e., u_{00} and u_{11}). For the REML method, these variance components were slightly overestimated (underestimated) when the type of pairing (i.e., relationship between the number of groups in each treatment condition and unequal variances) was positive (negative) and the ICC value was low (i.e., ICC = .1), but they were moderately overestimated (underestimated) when the type of pairing was positive (negative) and the ICC value was high (i.e., ICC = .3). These results are in general agreement with those

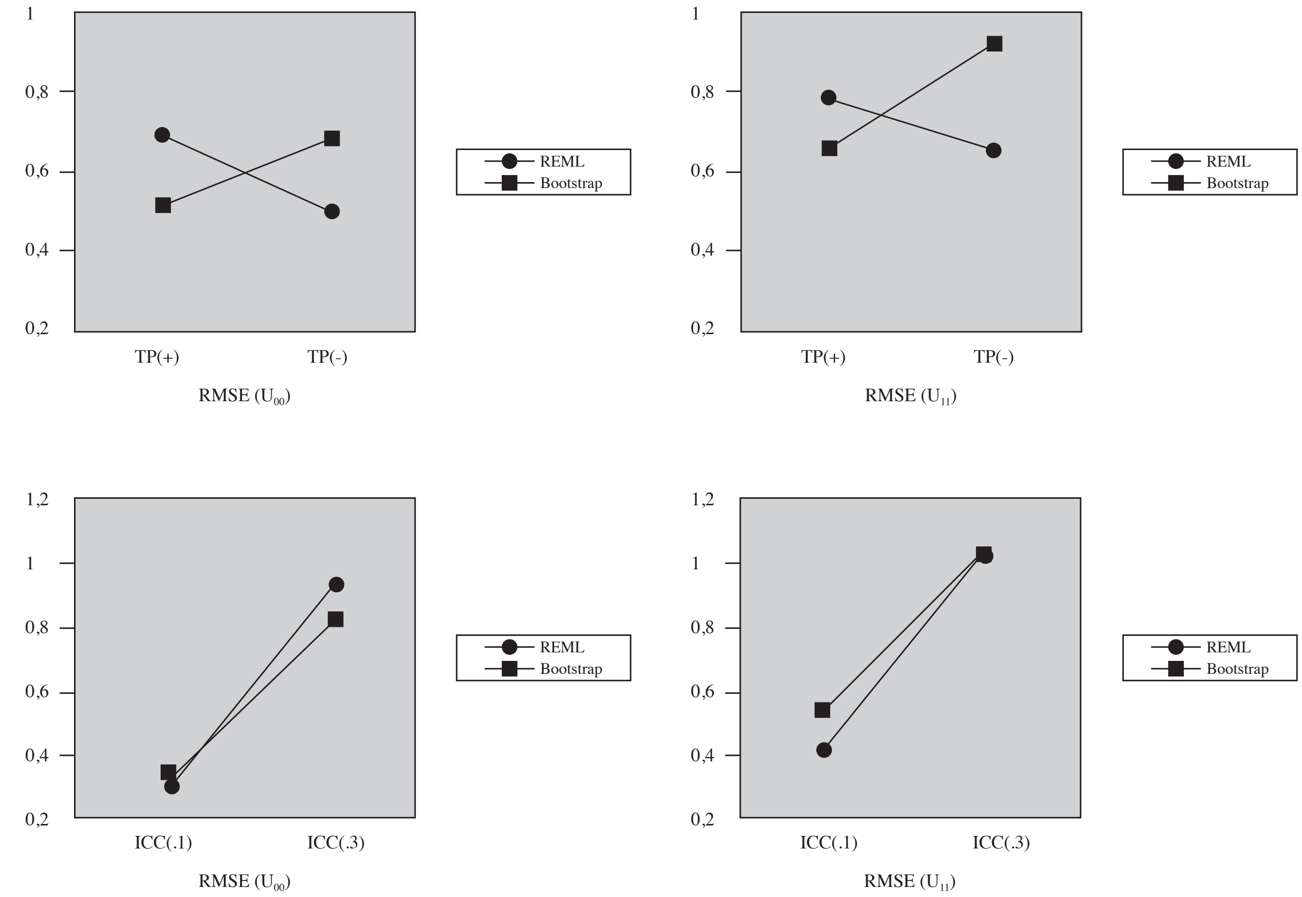

Figure 4. *Interaction plots of type of pairing (TP) and method (i.e., REML and bootstrap), and intraclass correlation (ICC) and method for RMSE rates of the second level variances (i.e., u_{00} and u_{11})*

reported by Vallejo and Ato (2012) in the multivariate context using uni-level analysis methods. In turn, the RB variances were always slightly overestimated. Unfortunately, our data demonstrate that this method is not generally accurate when the number of subjects per cluster is small.

Secondly, with respect to the coverage of the fixed effects, the REML estimation method was significantly affected by the number of groups and the type of pairing. For positive (negative) pairing, the standard errors are biased upward (downward). This results in confidence intervals whose coverage rates are above (below) the nominal 1-α level, and downwardly (upwardly) biased test statistics whose Type I error rates tend to well below (above) the nominal alpha level. In turn, the coverage rates for the RB were significantly affected by the group size and the ICC, but under this method the standard errors tend to be biased upward. On the other hand, the coverage of the variance components obtained via REML estimation had standard errors that were significantly more biased than the corresponding estimates from the RB method. In particular, REML yields standard error estimates that are severely biased downwards, whereas the RB method yields standard error estimates that are moderately biased either upward or downward.

Thirdly, for the RMSE of the fixed effects, an indicator of the accuracy of the estimates, the results revealed that the REML method performed slightly worse than the RB method, particularly when the assumption of normally distributed residuals did not hold. For the variance terms, however, the RB method did not offer a systematic improvement over the REML method in terms of RMSE. Again as expected, the difference in the RMSE between normal and exponential distributions was larger under REML than under RB. Thus, the results of this study suggest that the choice of procedure rests, in part, on a priori information about the shape of the population distribution. Techniques for evaluating the tenability of the model assumptions can be found in Snijders and Berkhof (2008).

Consequently, the non-parametric RB approach provides a robust alternative to the likelihood-based methods which could be used in preference to the ML/REML methods. This is especially true when it is likely that data depart from normality and the variances across clusters are heterogeneous. However, for a very small group size, the bootstrap resampling methods should be used with caution. For this reason, in future research, it would be informative to examine the performance of the linear model, using techniques that allow distributions of error terms other than the normal, and relax the requirement of constant variability (e.g., generalized linear models). Another question that remains open here is how the slope-intercept correlation affects performance of methods. The correlation between the two random effects is set at 0 in the simulation, which is a very special case of real data.

To conclude, we note that researchers face two problems when dealing with the bootstrap method. The first problem is the lack of an automated option for performing bootstrapping with the popular software packages, which requires programming some macros. Unfortunately, as Roberts and Fan (2004) noted, this can be a daunting task for investigators who do not have the skills, knowledge, or interest required to carry out it. The second problem is the computational load needed to obtain accurate results. For example, model fit by REML takes 4 seconds on a Workstation dual processor 3.0 GHz versus 15 minutes using 1000 bootstrap samples. If we add that it is common practice in multilevel modeling to compare the adequacy of different models rather than simply evaluating the fit of a single model in isolation, the problem becomes even more severe. Assuming heterogeneity of variance across groups, usually due to an interaction of treatments with some unspecified subject characteristics, we recommended using a hybrid modeling strategy, in which likelihood-based selection criteria are used in the model exploration phase, remembering that the variance components are generally underestimated) and bootstrap methods are used to report the final inferential results. Model specification may include semi-automatic search procedures, such as information criteria, and procedures that are more subjective, such as collapsing categorical predictors based on the observed relationship with the outcome (Vallejo, Arnau, Bono, Fernández, & Tuero-Herrero, 2010; Vallejo, Fernández, Livacic-Rojas, & Tuero-Herrero, 2011b; Vallejo, Tuero-Herrero, Núñez, & Rosário, in press).

Authors' Note

We thank Jichuan Wang for their SAS program to conduct nonparametric residual bootstrap multilevel modeling and two referees for constructive comments. This work was supported by a grant from the Spanish Ministry of Science and Innovation (Ref: PSI-2011-23395) awarded to the authors.

References

Apodaca, T.R., Magill, M., Longabaugh, R., Jackson, K.M., & Monti, P.M. (2013). Effect of a significant other on client change talk in motivational interviewing. *Journal of Consulting and Clinical Psychology*, *81*, 35-46.

Burton, A., Altman, D.G., Royston, P., & Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*, 4279-4292.

Carpenter, J.M., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society, Series C, 52*, 431-443.

de Leeuw, J., & Meijer, E. (2008). Introduction to multilevel analysis. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 1-75). New York: Springer-Verlag.

Dedrick, R.F., Ferron, J.M., Hess, M.R., Hogarty, K.Y., Kromry, J.D., Lang, T.R. & Lee, R.S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79,* 69-102

Diggle, P.J., Heagerty, P.J., Liang, K.Y., & Zeger, S.L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43,* 521-532.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Goldstein, H. (2011). Bootstrapping in multilevel models. In J.J. Hox & J.K. Roberts (Eds.), *Handbook advanced multilevel analysis* (pp. 163-171). New York: Routledge.

Hodges, J.S. (1998). Some algebra and geometry for hierarchical linear models, applied to diagnostics. *Journal of the Royal Statistical Society, Series B, 60*, 497-536.

Hox, J.J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York: Routledge.

Imel, Z.E., Hubbard, R.A., Rutter, C.M., & Simon, G. (2013). Patient-rated alliance as a measure of therapist performance in two clinical settings. *Journal of Consulting and Clinical Psychology*, *81*, 154-165.

Kasim, R., & Raudenbush, S. (1998). Application of Gibbs sampling to nested variance components models with heterogenous within group variance. *Journal of Educational and Behavioral Statistics, 23*, 93-116.

Laird, N.M., & Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

Maas, C.J.M., & Hox, J.J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127-137.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 92*, 778-785.

Núñez, J.C., Rosário, P., Vallejo, G., & González-Pienda, J.A. (2013). A longitudinal assessment of the effectiveness of a school-based mentoring program in middle school. *Contemporary Educational Psychology, 38,* 11-21.

Núñez, J.C., Vallejo, G., Rosário, P., Tuero-Herrero, E., & Valle, A. (in press). Variables from the students, the teachers and the school context predicting academic achievement: A multilevel perspective. *Journal of Psychodidactics*.

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

Roberts, J.K., & Fan, X. (2004). Bootstrapping within the multilevel/hierarchical linear modeling framework: A primer for use with SAS and SPLUS. *Multiple Linear Regressions Viewpoints*, *30*, 23-34.

Shieh, Y.Y., & Fouladi, R.T. (2002). The application of bootstrap methodology to multilevel mixed effects linear models under conditions of error term nonnormality. In *ASA Proceedings of the Joint Statistical Meetings*, pp. 3191-3196 . American Statistical Association, Alexandria, VA.

Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Snijders, T.A.B., & Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 141-175). New York: Springer-Verlag.

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,* 711-740

van der Leeden, R., Meijer, E., & Busing, F.M.T.A. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401-434). New York: Springer-Verlag.

Vallejo, G., Arnau, J., Bono, R., Fernández, M.P., & Tuero-Herrero, E. (2010). Nested model selection for longitudinal data using information criteria and the conditional adjustment strategy. *Psicothema*, *22*, 323-333.

Vallejo, G., & Ato, M. (2012). Robust tests for multivariate factorial designs under heteroscedasticity. *Behavior Research Methods, 44*, 471-489.

Vallejo, G., Fernández, M.P., Livacic-Rojas, P.E., & Tuero-Herrero, E. (2011a). Comparison of modern methods for analyzing unbalanced repeated measures data with missing values. *Multivariate Behavioral Research*, *46*, 900-937.

Vallejo, G., Fernández, M.P., Livacic-Rojas, P.E., & Tuero-Herrero, E. (2011b). Selecting the best unbalanced repeated measures model. *Behavior Research Methods, 43*, 18-36.

Vallejo, G., Tuero-Herrero, E., Núñez, J.C., & Rosário, P. (in press). Performance evaluation of recent information criteria for selecting multilevel models in behavioral and social sciences. *International Journal of Clinical and Health Psychology.*

Wang, J., Carpenter, J.R., & Kepler, M.A. (2006). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine*, *82*, 130-143.

Wang, J., Xie, H., & Fisher, J. (2011). *Multilevel models: Applications using SAS*. Berlin: De Gruyter & Higher Education Press.